Model validation for aggregate inferences in out-of-sample prediction[*]

Lauren Kennedy[†], Aki Vehtari[‡], Andrew Gelman[§]

15 Feb 2024

Abstract

Generalization to new samples is a fundamental rationale for statistical modeling. For this purpose, model validation is particularly important, but recent work in survey inference has suggested that simple aggregation of individual prediction scores does not give a good measure of the score for population aggregate estimates. In this manuscript we explain why this occurs, propose two scoring metrics designed specifically for this problem, and demonstrate their use in three different ways. We show that these scoring metrics correctly order models when compared to the true score, although they do underestimate the magnitude of the score. We demonstrate with a problem in survey research, where multilevel regression and poststratification (MRP) has been used extensively to adjust convenience and low-response surveys to make population and subpopulation estimates.

Keywords: sample surveys, model validation, Bayesian statistics

[*] We gratefully acknowledge support from National Institutes of Health grants 5R01AG067149-02. We acknowledge the Office of Naval Research for partial support of this work. We acknowledge the support by the Research Council of Finland Flagship programme: Finnish Center for Artificial Intelligence, and Research Council of Finland project (340721). We thank Daniel Simpson and Jonah Gabry for early discussions on validating models for MRP (particularly leave-one-cell-out approaches). We thank Swen Kuh for generously sharing simulation code for this work. We thank Alex Cooper for early work as a research assistant investigating scoring rules. Code for the simulations presented in this work can be found at `https://github.com/lauken13/MRP_validation_ALT`

[†] School of Computer and Mathematical Sciences, University of Adelaide

[‡] Department of Computer Science, Aalto University

[§] Department of Statistics and Department of Political Science, Columbia University

1 Introduction

A central challenge of statistics is to generalizing from observed data to new settings. To do this, sample surveys increasingly rely on model-based estimates to overcome low response rates and non-probability samples. These model-based estimates use a prediction framework, but the purpose is not to predict each unobserved individual accurately. Instead the purpose is to estimate accurate population summaries from individual predictions. Although these two aims might seem similar, recent work suggests that the best model for the individual predictions might not be the best model for population summaries (Kuh et al. 2023). This leads to challenges in assessing both model adequacy and model selection (Gelfand et al. 1992).

In this manuscript we focus on a specific model-based technique, multilevel regression and poststratification (MRP; Gelman & Little 1997, Park et al. 2006), which has been widely used in political science (e.g., Lax & Phillips 2009b,a, Ghitza & Gelman 2013, Wang et al. 2015) and increasingly deployed in social (e.g., Alsalti et al. 2023, Kolczynska et al. 2024) and health sciences (e.g., Machalek et al. 2022, Downes et al. 2018). As with all methods of survey adjustment, the method relies on assumptions of model suitability.

MRP has three stages (see, e.g., Lopez-Martin et al. 2022). First, a model (originally a Bayesian multilevel model, but increasingly other regularising models, e.g., Ornstein 2020, Bisbee 2019, Gao et al. 2021) is fit to the observed sample data conditional on discrete predictors (typically geographic and demographic) whose joint distribution in the population is known or can be estimated. Second, a poststratification table is constructed to represent this population joint distribution, ideally using counts derived from census data or high quality official statistics. The model from the sample is then used to predict the expected value for each cell. Third, the population totals in each cell are used to weight the cell estimates to obtain an estimate for the population (or subpopulation). The three steps can be remembered as model-predict-aggregate.

The population estimate relies on the model prediction, and so it is essential to consider how to fairly validate the model. Given the emphasis on prediction in this method, previous work has investigated the sum of leave-one-out (LOO) expected-log-predictive-density (ELPD) (Kuh et al. 2023). This is a commonly used metric for model goodness in the Bayesian literature, and is easily implemented through the LOO package (Vehtari et al. 2021). Kuh et al. (2023) also consider a weighted alternative to this metric (initially proposed by Lumley & Scott 2015) to account for non-random sampling.

Kuh et al. (2023) use results from Little & Vartivarian (2005) to demonstrate that neither standard sum of LOO-ELPD nor the weighted alternative correctly distinguish between a model that correctly produces an approximately unbiased estimate of the mean and one that does not.

	Individual			Mean	
	$\hat{y}_i$	$(\hat{y}_i - y_i)^2$	$\frac{1}{2}\sum(\hat{y}_i - y_i)^2$	$\frac{1}{2}\sum \hat{y}_i$	$(\frac{1}{2}\sum \hat{y}_i - \frac{1}{2}\sum y_i)^2$
Model 1	$\{0,1\}$	$\{0,1\}$	0.5	0.5	0.25
Model 2	$\{-2,2\}$	$\{-2,2\}$	4	0	0

Table 1: *Demonstration of difference between mean of individual squared error and squared error of mean in a simple example of* $n = 2$. *True outcome* $y_i = 0$.

1.1 Individual scores and aggregate scores

We argue that the underlying difficulty described in Kuh et al. (2023) is that the best model for predicting individuals may not correspond to the best model for an aggregate estimate. To understand why, consider a simple example of a sample of size 2, where these individuals have a true outcome value 0. We will say $y_{\text{true}} = \{0, 0\}$. The mean of y_{true} is $E(y_{\text{true}}) = \frac{0+0}{2} = 0$. Now we have two potential models to estimate y_{true}. Each produces an estimate (column 2, Table 1), from which we can calculate the squared error and the mean squared error over both individuals. We can also use these estimates to create an estimate of the mean, and from this calculate the squared error of the mean. As we can see, while model 1 is preferred by the mean squared individual error, model 2 is far preferred at estimating the mean.

This simple example demonstrates what we believe to be one of the underlying challenges of validating multilevel regression and poststratification (MRP) models. An MRP based estimator is by definition an estimator that uses cellwise estimates to compose a population or subpopulation estimates. Previous investigations used sum of individual goodness to estimate the goodness of the MRP estimate, which was not effective (Kuh et al. 2023).

1.2 Notation and key terms

Consider a finite population P with N individuals, where interest lies in a binary variable Y[1]. Assume a set of k predictors whose values are known in the population and are represented by a $N \times k$ matrix, $\mathbf{X}$. Assume the predictors are discrete, with the k^{th} variable, $X^{(k)}$, having $l^{(k)}$ levels. This means the distribution of X in the population can be represented by the number of items within each of the $J = \prod_k l^{(k)}$ possible configurations of predictors.

The population mean of variable of interest can be written as,

$$\mathrm{E_P}(Y) = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{\sum_{j=1}^{J} N_j \Pr(Y_{i\in j} = 1)}{N}, \tag{1}$$

where $\Pr(Y_{i\in j} = 1) = \sum_{i\in j} Y_i / N_j$.

[1] Whilst some authors have used a continuous outcome variable (Liu et al. 2023), the prevalent applications currently focus on estimating probabilities.

Next take a sample of size n. The probability of inclusion in the sample for every j^{th} cell is assumed to be the same and denoted π_j.

Following traditional sampling notation, we use lower case for the sample and upper case for the population, and where a matrix is needed, it will be presented in boldface or evident from context. Our first step is to estimate the probability of the outcome in the j^{th} cell in the population using the sample, denoted $\Pr(Y_{i\in j}|y,\theta)$ as shorthand for $\Pr(Y_{i\in j}=1|y,\mathbf{x},\theta)$. We then use this to estimate the population mean, $E_P(Y|y,\theta)$. We do this by modifying (1) into

$$\mathrm{E_P}(Y|y,\theta) = \frac{\sum_{j=1}^{J} N_j \Pr(Y_{i\in j}|y,\theta)}{\sum_{j=1}^{J} N_j}. \tag{2}$$

Our goal is to identify the best model M and estimate the parameters θ, denoted $\hat{\theta}$, using the sample. We can then use this to estimate $\Pr(Y_{i\in j}|y,\hat{\theta})$, which can then be used in (2).

$$\mathrm{E_P}(Y|y,\hat{\theta}) = \frac{\sum_{j=1}^{J} N_j \Pr(Y_{i\in j}|y,\hat{\theta})}{\sum_{j=1}^{J} N_j}. \tag{3}$$

In this paper we discuss aggregation at a range of different levels. For clarity we describe all terms with a simple example. Consider a survey of voting intention. The smallest unit is the *individual*, the person who is being surveyed. These individuals can be aggregated in different ways. For example, we could aggregate all individuals with a specific set of demographic qualities (e.g., aged 18–25, female, college educated). We call this a cell if these three demographics are the full set of X variables. We could also aggregate at a particular *level* (e.g., 18-25 year) of a variable (age), which we call a *subpopulation* (also known as a small area). In the limit the subpopulation is the full *population* (e.g., all US voters).

1.3 Contributions

In this manuscript we

- Propose two alternative scoring metrics for estimating the probability of a binary outcome in the population (Section 2)

- Demonstrate the use of this method in a leave-one-cell-out cross validation scheme, and demonstrate that unlike previous work, this method retains the correct model ordering (Section 3)

- Apply this method to subpopulation estimation (Section 4)

- Extend the leave-one-cell-out cross validation to a leave-one-cell-out reference validation score where instead of using the sample as a proxy for the cellwise population truth we use a reference model estimate (Section 5)
- The leave-one-cell-out approach requires all cells to be observed in the sample. We also propose an approximate leave-one-cell-out reference approach for use when not all cells are observed in the sample (Section 5).

Together we intend for these contributions to provide a convincing argument that the proposed scoring metric is more appropriate for scoring models for MRP.

2 Aggregate scoring from cellwise components

In this section we demonstrate that by decomposing a score of the MRP population mean into its cellwise components, we can produce an estimate of the population mean score using cellwise goodness. We complete this for a squared error score (Error^2) for the population mean and the continuous ranked probability score (CRPS).

2.1 Squared Error

First consider the square of the error for the population mean,

$$\text{Error}_{\text{MRP}} = E_P(Y|\theta, y) - E_P(Y).$$

Using (1) and (3), we can instead write:

$$\text{Error}_{\text{MRP}} = \frac{\sum_{j=1}^{J} N_j \text{Pr}(Y_{i\in j}|y, \hat{\theta})}{\sum_{j=1}^{J} N_j} - \frac{\sum_{j=1}^{J} N_j \text{Pr}(Y_{i\in j})}{\sum_{j=1}^{J} N_j}$$

which can be simplified to

$$\text{Error}_{\text{MRP}} = \frac{\sum_{j=1}^{J} N_j (\text{Pr}(Y_{i\in j}|y, \hat{\theta}) \; - \; \text{Pr}(Y_{i\in j}))}{\sum_{j=1}^{J} N_j}. \tag{4}$$

This sum itself needs to be squared to get the squared error for the population mean

$$\text{Error}^2_{\text{MRP}} = \left(\frac{\sum_{j=1}^{J} N_j (\text{Pr}(Y_{i\in j}|y, \hat{\theta}) \; - \; \text{Pr}(Y_{i\in j}))}{\sum_{j=1}^{J} N_j} \right)^2,$$

which is why simply summing the squared error of the individuals or cells (as is often done when scoring models; see Gelman et al. 2014) would not provide an acceptable score for the population mean. For clarity, the mean of the squared error of the cells, adjusted for sample representation would be

$$\frac{\sum_{j=1}^{J} N_j(\Pr(Y_{i\in j}|y,\hat{\theta}) - \Pr(Y_{i\in j}))^2}{\sum_{j=1}^{J} N_j}.$$

For notation simplicity we write the error instead of squared error for formulae throughout this manuscript, but figures show the squared error.

2.2 Continuous ranked probability scores

The continuous ranked probability score (CRPS) is an extension of squared error that incorporates the full probability distribution. This relationship suggests that the CRPS for the population level MRP estimate be similarly decomposed. CRPS is described as

$$\mathrm{CRPS}(F, y) = -\int_{-\infty}^{\infty} (F(y) - \mathbb{1}(y \geq x))^2 \, dy,$$

where F is the cumulative distirbution function of the predictive distribution and x is the true value. One benefit of this score is that it can be implemented through approximate draws of this distribution (Equation (21) of Gneiting & Raftery 2007). To harmonise notation, we denote ϕ as the posterior estimate for the population expectation and ϕ' as a permutation of the posterior draws of ϕ,

$$\mathrm{CRPS}(\phi, E_P(Y)) = \frac{1}{2}E(|\phi - \phi'|) - E(|\phi - E_P(Y)|).$$

In our setting we focus on cases where we use posterior draws to estimate expectation. Denoting the b^{th} posterior draw as ϕ^b, we can write this as

$$\mathrm{CRPS}(\phi, E_P(Y)) = \frac{1}{2}\frac{1}{B}\sum_{b}^{B}|\phi^b - \phi'^b| - \frac{1}{B}\sum_{b}^{B}|\phi^b - E_P(Y)|.$$

The ϕ represents our MRP estimate, and $E_P(Y)$ the population mean truth, so we can rewrite as

$$\begin{aligned}&\mathrm{CRPS}_{\mathrm{MRP}}(Y, y, \hat{\theta}) = \\ &\frac{1}{B}\sum_{b}^{B}\left(\frac{1}{2}\left|\frac{\sum_{j=1}^{J}(N_j\Pr(Y_{i\in j}|y,\hat{\theta}^b))}{\sum_{j=1}^{J} N_j} - \frac{\sum_{j=1}^{J}(N_j\Pr(Y_{i\in j}|y,\hat{\theta}'^b))}{\sum_{j=1}^{J} N_j}\right| - \right.\\ &\left.\left|\frac{\sum_{j=1}^{J}(N_j\Pr(Y_{i\in j}|y,\hat{\theta}^b))}{\sum_{j=1}^{J} N_j} - \frac{\sum_{j=1}^{J}(N_j\Pr(Y_{i\in j}))}{\sum_{j=1}^{J} N_j}\right|\right).\end{aligned}$$

Within each summation, this can be reordered to produce a weighted sum of cellwise components:

$$\text{CRPS}_{\text{MRP}}(Y, y, \hat{\theta}) = \frac{1}{B}\sum_{b}^{B}\left(\frac{1}{2}\left|\frac{\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|y, \hat{\theta}^b) - \Pr(Y_{i\in j}|y, \hat{\theta}'^b)\Big)}{\sum_{j=1}^{J} N_j}\right| - \left|\frac{\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|y, \hat{\theta}^b) - \Pr(Y_{i\in j})\Big)}{\sum_{j=1}^{J} N_j}\right|\right). \quad (5)$$

As with squared error, the absolute value is outside of the summation over cells, rather than inside as we would expect for a usual sum of the cellwise scores:

$$\text{CRPS}(Y, y, \hat{\theta}) = \frac{1}{N}\sum_{j=1}^{J} N_j\left(\frac{1}{2}\frac{1}{B}\sum_{b}^{B}\Big|\Pr(Y_{i\in j}|y, \hat{\theta}^b) - \Pr(Y_{i\in j}|y, \hat{\theta}'^b)\Big| - \frac{1}{B}\sum_{b}^{B}\Big|\Pr(Y_{i\in j}|y, \hat{\theta}^b) - \Pr(Y_{i\in j})\Big|\right). \quad (6)$$

3 Scoring rule applied with cross validation

We aim to score our model without the population truth through the observed values. We first consider a scenario where at least one individual in every cell is observed. We continue this assumption in Section 4 where we consider subpopulation estimates. In Section 5.2 we relax this assumption with a reference validation approach.

3.1 Simulation Design

For simplicity, we borrow the simulation scenario in Example 1 of Kuh et al. (2023), but add an additional constraint that all cells in the population must be observed at least once in the sample. Kuh et al. (2023) had a less restrictive constraint that at least one observation in every level needed to be observed. To create our simulated data, first four independent normal$(0, 2)$ variables are sampled. The probability density of the outcome ($\Pr(Y_{i\in j=1})$) and the probability of being included in the sample ($\pi_{i\in j}$) were created as follows:

$$\text{Probability of outcome: } \Pr(Y_{i\in j} = 1) = \text{logit}^{-1}(0.1X_1 + 1X_2 + 0.1X_3 + 1X_4), \quad (7)$$

$$\text{Inclusion probability: } \pi_{i\in j} = \text{logit}^{-1}(0.1X_1 + 0.1X_2 + 1X_3 + 1X_4). \quad (8)$$

Here, X_2 is a precision variable (related strongly to the outcome but weakly to the probability of inclusion in the sample), and X_4 is a bias variable (related strongly to both). According to Little & Vartivarian (2005), models with X_4 and not X_2 should be strongly preferred over models with X_2 and not X_4.

We create binary observations Y_i for every individual in the population using $\Pr(Y_{i\in j} = 1)$. We also discretize the X_k variable into 5 groups of equal range.

We simulate with a population size of $20,000$, twice the size of Kuh et al. (2023). We sample from the population by first ensuring every cell has one observation and then sampled the remaining obsevations with probability $\pi_{i\in j}$. The sample size is held constant at $1,000$. As the outcome is binary, we choose to fit a binomial likelihood of counts of $y_{i\in j} = 1$ per cell j. For simplicity we fit a subset of the models used by Kuh et al. (2023), focusing on the following models:

- Full model: $\Pr(y_{i\in j} = 1|n_j) = \text{logit}^{-1}(\beta_0 + \alpha_j^{(X_1)} + \alpha_j^{(X_2)} + \alpha_j^{(X_3)} + \alpha_j^{(X_4)})$,
- Precision variable model: $\Pr(y_{i\in j} = 1|n_j) = \text{logit}^{-1}(\beta_0 + \alpha_j^{(X_1)} + \alpha_j^{(X_2)} + \alpha_j^{(X_3)})$,
- Bias variable model: $\Pr(y_{i\in j} = 1|n_j) = \text{logit}^{-1}(\beta_0 + \alpha_j^{(X_1)} + \alpha_j^{(X_3)} + \alpha_j^{(X_4)})$,
- Nuisance variable model: $\Pr(y_{i\in j} = 1|n_j) = \text{logit}^{-1}(\beta_0 + \alpha_j^{(X_1)} + \alpha_j^{(X_3)})$,

where $\alpha_j^{(k)} \sim \text{normal}(0, \sigma^{(k)})$, $\sigma^{(k)} \sim t(3, 0, 2.5)$ and $\beta_0 \sim t(3, 0, 2.5)$, which are the standard priors in the brms R package (Bürkner 2017, 2018).

For each sample iteration and model we perform the following steps:

1. Create a population.
2. Create a sample from this population so that each cell has at least one individual observed, and then according to $\pi_{i\in j}$.
3. Fit all four models on this sample using brms with default settings.
4. For each model we calculate
   (a) The population mean estimate using MRP and score it compared to the truth using CRPS and squared error,
   (b) An estimate for each cell and compare to the true $\Pr(Y_{i\in j} = 1)$ to calculate the cellwise CRPS and squared error as per (4) and (5).

The cellwise score using the population cellwise truth is equivalent, up to a small amount of noise, to the population score; see supplementary materials for visualisation of results). However, in practice the population truth (cellwise or at the aggregate level) is not available.

One option in this case is to approximate the population truth with the sample observation, $\Pr(y_{i\in j}|n_j, y_{i\in j})$. This leads to

$$\widehat{\text{Error}}_{\text{MRP}} = \frac{\sum_{j=1}^{J} N_j(\Pr(Y_{i\in j}|y,\hat{\theta}) - \Pr(y_{i\in j}|n_j, y_j))}{\sum_{j=1}^{J} N_j}, \tag{9}$$

and

$$\begin{aligned}\widehat{\text{CRPS}}_{\text{MRP}}(y,n,\hat{\theta},\mathbf{x}) = \frac{1}{2}\frac{1}{B}\sum_{b}^{B}\left(\left|\frac{\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|y,\hat{\theta}^{b}) - \Pr(Y_{i\in j}|y,\hat{\theta}'^{b})\Big)}{\sum_{j=1}^{J} N_j}\right|\right) - \\ \frac{1}{B}\sum_{b}^{B}\left(\left|\frac{\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|y,\hat{\theta}^{b}) - \Pr(y_{i\in j}=1|n_j,y_j)\Big)}{\sum_{j=1}^{J} N_j}\right|\right).\end{aligned} \tag{10}$$

When there is only one observation per cell, the cellwise probability estimate is either 0 or 1 accordingly. Previous research suggests that this will underestimate the magnitude of the score, which we also find (see supplementary materials for a visual demonstration). However, the ordering of models is preserved—namely, the separation between models that contain the bias variable as a predictor and models that contain only the precision variable as a predictor. The standard sum of individual scores does not maintain this ordering (e.g., see Kuh et al. 2023 for sum of ELPD, and supplementary materials for mean of CRPS and squared error scores), which suggests the scoring rule could be useful for ordering models but not estimating the expected magnitude of error.

3.2 Brute-force cross validation

To resolve the underestimation of the magnitude of scores we consider a cross validation approach. There is a wide and established literature on cross validation for estimating predictions. We considered three potential schemes: K-fold cross validation, leave-one-cell-out (LOCO) and leave-one-level-out (LOLO).

K-fold cross validation partitions the sample into K non-overlapping folds. The k^{th} fold is removed to fit the model and score calculated using this fold. To use this approach in with MRP, we would need to adjust the score estimated in this fold to population. To do this we would need at least K observations present in every poststratification matrix cell (one for each fold) which isn't plausible given the size and complexity of survey data and the set of potential adjustment variables.

Leave-one-cell-out cross validation uses a similar idea, but instead partitions the data so each cell is it's own fold. The j^{th} cell is removed, the model fit and the score estimated using the observations in the removed cell $\Pr(y_j = 1|n_j)$. In this scenario, only one observation is required per cell (to estimate the probability), and the estimate of population error can be made with our proposed score decomposition. This is the method that we will focus on in this manusucript.

Despite this, leave-one-level-out (LOLO) cross validation could be useful specifically in subpopulation estimation (see Section 4). In this instance the data are partitioned by the level of a adjustment variable and a similar process to the other validation techniques applied. In general though, it is not clear which variable to focus on when multiple variables are in the model when assessing a whole population estimate.

To use LOCO, we propose replacing our estimate $\Pr(Y_{i\in j} = 1|y, \hat{\theta})$ with $\Pr(Y_{i\in j} = 1|y_{i\notin j}, \hat{\theta})$, the probability of $y = 1$ in that cell given that it was not used when the model was fit to get

$$\text{Error}^{(\text{LOCO}-\text{CV})} = \frac{\sum_{j=1}^{J} N_j(\Pr(Y_{i\in j}|y_{i\notin j}, \hat{\theta}) - \Pr(y_{i\in j}|n_j, y_j))}{\sum_{j=1}^{J} N_j}, \tag{11}$$

and

$$\begin{aligned}\text{CRPS}^{(\text{LOCO}-\text{CV})}(y, n, \hat{\theta}) = \frac{1}{2}\frac{1}{B}\sum_{b}^{B}\left(\left|\frac{\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|y_{i\notin j}, \hat{\theta}^{b}) - \Pr(Y_{i\in j}|y_{i\notin j}, \hat{\theta}'^{b})\Big)}{\sum_{j=1}^{J} N_j}\right|\right) - \\ \frac{1}{B}\sum_{b}^{B}\left(\left|\frac{\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|y_{i\notin j}, \hat{\theta}^{b}) - \Pr(y_{i\in j}|n_j, y_j)\Big)}{\sum_{j=1}^{J} N_j}\right|\right).\end{aligned} \tag{12}$$

From here on in this manuscript we will only be discussing an aggregate MRP score and so no longer note this for notation readability.

When the model is refit excluding each cell (leading to $J + 1$ model fits), we call this a brute-force implementation of leave-one-cell-out (LOCO). It is not efficient in terms of time and computational resources, but it does allow us to directly investigate whether the underestimation impact of cross validation on the magnitude of score estimate. Figure 1 shows that the score estimated with a brute-force LOCO approach is indeed slightly larger than that estimated using the full sample to model and estimate the score. However, this is still a considerable underestimate for the true score, as shown by Figure 2.

Despite the underestimate of the magnitude of the score, we believe that this approach is still useful to practitioners who wish to evaluate the relative efficacy of different models when estimating the population mean. Unlike previous attempts, ordering is maintained. We advise that the magnitude of the score be interpreted with caution.

3.3 Approximate cross validation

The magnitude of the score is not the only challenge for a brute-force leave-one-cell-out approach. A practical issue is the sheer amount of computational required for a brute-force approach. For example, in the first iteration of the previously described simulation study,

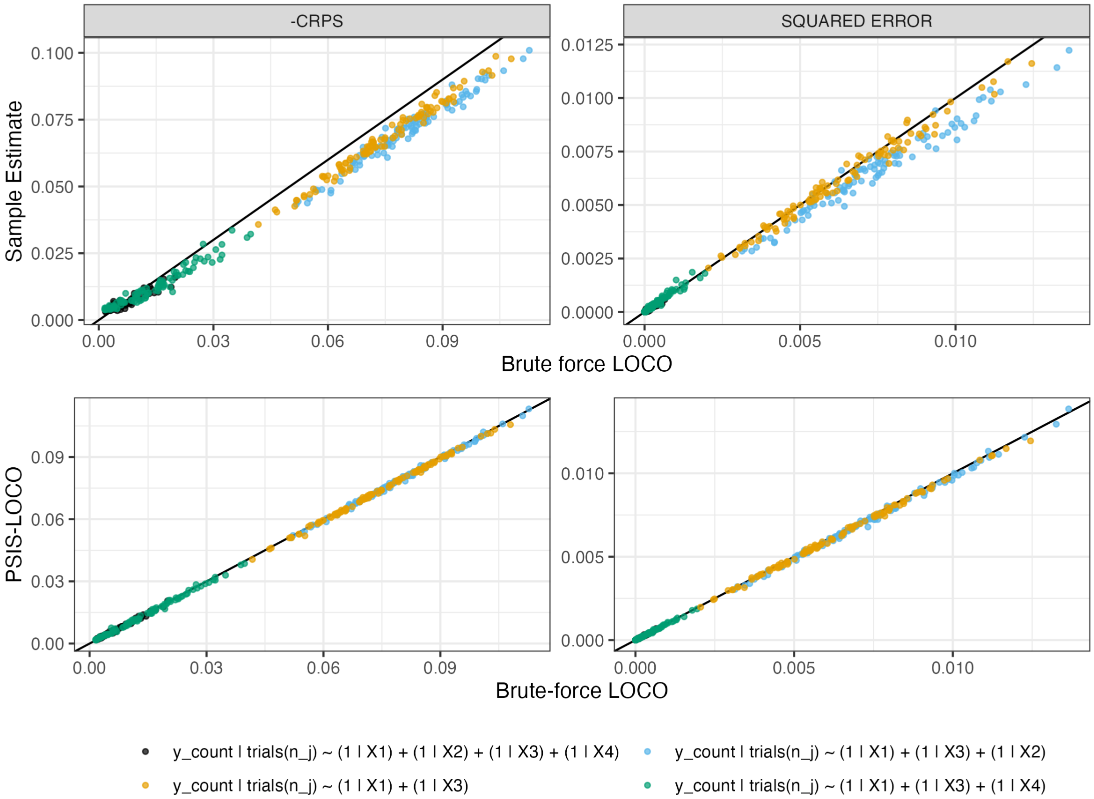

Figure 1: *Comparison of the brute-force leave-one-cell-out approximation (LOCO, x-axis) with a non cross validation scheme (y-axis, top row) and approximate cross validation scheme (y-axis, bottom row). Colour of point represents model, points within this represent different simulation iterations. The black line represents an identity line. We plot (-)CRPS(left panels) and squared error (right panels).*

the posterior inference had to be rerun 294 times (once for each cell plus one for full data), and this example is a relatively small toy problem in terms of the size of the poststratification matrix. Many MRP examples involve poststratifying to an area level variable like postcode, which might have 100 levels (Machalek et al. 2022). In this section, we apply an approximate LOO method based on Pareto smoothed importance sampling (PSIS; Vehtari et al. 2017, 2024).

PSIS-LOO (equation (10) of Vehtari et al. 2017) estimates the sum expected log individual predictive density by

$$\widehat{\text{elpd}}_{\text{PSIS-LOO}} = \sum_{i=1}^{n} \log\left(\frac{\sum_{b=1}^{B} w_i^b \Pr(y_i|\theta^b)}{\sum_{b=1}^{B} w_i^b}\right),$$

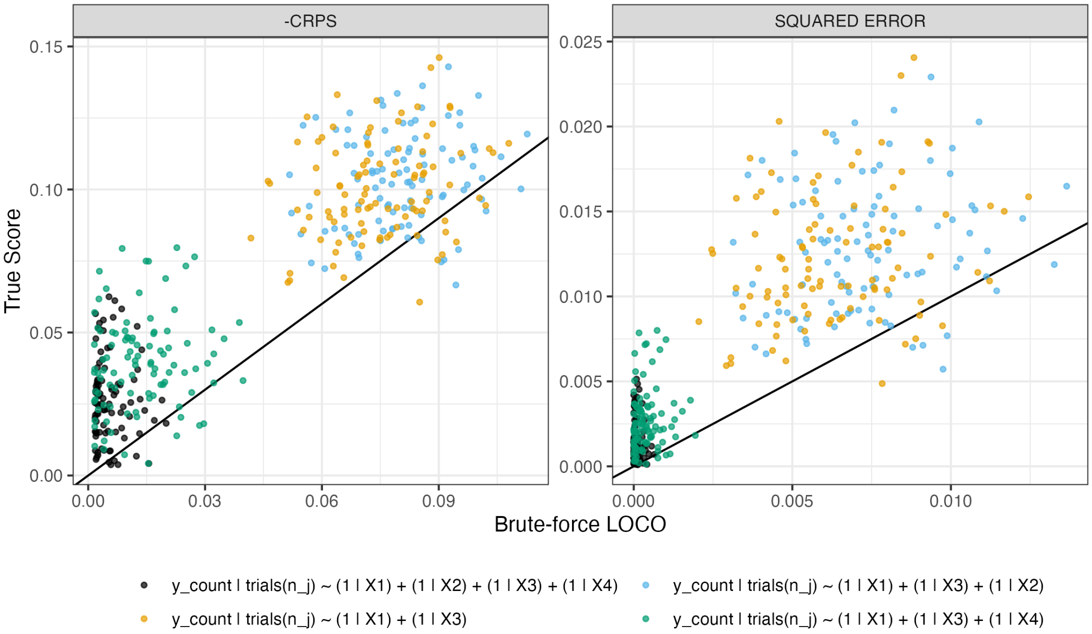

Figure 2: *Comparison of a direct score of the MRP estimate (using known population truth, y-axis) against an indirect score of the MRP estimate (using a brute-force leave-one-cell-out approach, x-axis). Colour of point represents the different models fitted, points within this different simulation iterations. The black line represents an identity line. We plot squared error and the negative of CRPS, so higher can be interpreted as a worse model. These results demonstrate that using LOCO still underestimates the model score.*

where w_i^b is the Pareto smoothed importance weight for the b^{th} posterior draw (denoted s by Vehtari et al. 2017) and the i^{th} individual. We modify this for our purposes by using other scoring metrics. We begin with substituting the error in each cell as before, with for shorthand we denote $\epsilon_j^b = \Pr(Y_{i\in j}|y_{i\notin j}, \hat{\theta}^b) - \Pr(Y_{i\in j}|n_j, y_j)$. For each cell j we can compute Pareto smoothed weights w_j^b for each posterior sample b to estimate ϵ_j^b as outlined by Vehtari et al. (2017). Using these, we can approximate ϵ_j^b and substitute into (11) as

$$\widehat{\text{Error}}^{(\text{LOCO}-\text{CV})} = \frac{1}{N}\sum_{j=1}^{J} N_j \frac{\sum_{b=1}^{B} w_j^b \epsilon_j^b}{\sum_{b=1}^{B} w_j^b} \tag{13}$$

and thus estimate the squared error of the MRP estimate using the full model fit.

It is slightly more complicated to achieve this with the CRPS, but we begin by denoting the $g(y_{i\notin j}, \hat{\theta}^b, \hat{\theta}'^b) = \Pr(Y_{i\in j}|y_{i\notin j}, \hat{\theta}^b) - \Pr(Y_{i\in j}|y_{i\notin j}, \hat{\theta}'^b)$ and $h(y_{i\notin j}, y_j, n_j, \hat{\theta}^b) =$

$\Pr(Y_{i\in j}|y_{i\notin j},\hat{\theta}^b) - \Pr(y_{i\in j}|n_j,y_j)$. This means we can write (12) more concisely as

$$\mathrm{CRPS}^{(\mathrm{LOCO-CV})}(y,n,\hat{\theta},\hat{\theta}') = \frac{1}{2}\frac{1}{B}\sum_{b=1}^{B}\left(\left|\frac{\sum_{j=1}^{J} N_j g(y_{i\notin j},\hat{\theta}^b,\hat{\theta}'^b)}{\sum_{j=1}^{J} N_j}\right|\right) - \frac{1}{B}\sum_{b=1}^{B}\left(\left|\frac{\sum_{j=1}^{J} N_j h(y_{i\notin j},y_j,n_j,\hat{\theta}^b)}{\sum_{j=1}^{J} N_j}\right|\right).$$

While it is possible to simply take an importance-weighted sum for the squared error, CRPS doesn't have an amenable form. To apply the PSIS weights, which are the b^{th} weight for the j^{th} cell, we do importance resampling of the posterior draws for each cell using a stratified resampling algorithm (Kitagawa 1996) using the B PSIS weights for each cell. This procedure is implemented in the posterior R package (Bürkner et al. 2023). This produces B resampled posterior draws $\hat{\theta}_{RW}$, which we permute to calculate PSIS reweighted $\hat{\theta}'_{RW}$, and thus $g'_j(\hat{\theta}^b_{RW},\hat{\theta}'^b_{RW}) = \Pr(Y_{i\in j}|\hat{\theta}^b_{RW}) - \Pr(Y_{i\in j}|\hat{\theta}'^b_{RW})$ and $h'_j(y_j,n_j,\hat{\theta}^b_{RW}) = \Pr(Y_{i\in j}|\hat{\theta}^b_{RW}) - \Pr(Y_{i\in j}|n_j,y_j)$, to get

$$\widehat{\mathrm{CRPS}}^{(\mathrm{LOCO-CV})}(y,n,\hat{\theta}_{RW},\hat{\theta}'_{RW}) = \frac{1}{BN}\left(\frac{1}{2}\sum_{b=1}^{B}\left(\left|\sum_{j=1}^{J} N_j g_j(\hat{\theta}^b_{RW},\hat{\theta}'^b_{RW})\right| - \left|\sum_{j=1}^{J} h_j(y_j,n_j,\hat{\theta}^b_{RW})\right|\right)\right). \quad (14)$$

The second row of Figure 1 shows that the approximate PSIS-LOCO scores are approximately equivalent to the brute-force LOCO scores.

Through these simulations, we demonstrate the efficacy of our proposed cellwise scoring in a leave-one-cell-out approach. Whilst the model ordering is well preserved (which is desirable as other approaches do not preserve this), the scores as estimated with the sample are more favourable to the quality of the estimate than they should be. Using a leave-one-cell-out approach only marginally improves on this, but can be well approximated with a fast approximation such as PSIS-LOCO. Although it is frustrating to underestimate to this degree, we feel the preservation of model ordering still makes this a useful tool, albeit one where the scale must be interpreted carefully.

When we followed the traditional approach of comparing the mean of cell scores to the true mean of cell scores, the leave-one-cell-out approach appeared to slightly over estimate the magnitude of score (see supplementary), suggesting that the observed underestimation is a feature of our particular score decomposition.

4 Differentiating between population and subpopulation goodness

One of the benefits of MRP is that the same model can be used for both population and subpopulation estimates. To demonstrate how, we follow Kuh et al. (2023) to use set notation to describe the set of cells that form the subpopulation as $\mathbb{S}$. The total number of cells in this set is S and any particular cell as s. In the case of the full population, $\mathbb{S}$ is all cells in the population, $S = J$ and $s = j$. Thus we can express a population weighted estimate for any subpopulation (small area) in the population as

$$\mathrm{E}_{\mathbb{S}}(Y|y,\hat{\theta}) = \frac{\sum_{s=1}^{S} N_s \mathrm{Pr}(Y_{i\in s}|y,\hat{\theta})}{\sum_{s=1}^{S} N_s}. \tag{15}$$

One benefit of the leave-one-cell-out estimation technique described in the previous section is that it is also constructed within each cell. This means that we can apply both the CRPS score and squared-error score described previously by focusing on the set of cells that describe a particular subpopulation. In the previous section we demonstrated the accordance of the PSIS-LOCO approximation to brute-force LOCO, and so in this section we use only the PSIS-LOCO for computational efficacy.

In this study we add two more models to our comparison set to add great distinction for the X_1 and X_3 subpopulation estimates:

- X_1 only : $\mathrm{Pr}(y_{i\in j} = 1|n_j) = \mathrm{logit}^{-1}(\beta_0 + \alpha_{i\in j}^{(X_1)})$,
- X_3 only : $\mathrm{Pr}(y_{i\in j} = 1|n_j) = \mathrm{logit}^{-1}(\beta_0 + \alpha_{i\in j}^{(X_3)})$,

We can modify (13) to estimate the subpopulation score with

$$\widehat{\mathrm{Error}}_S^{(\mathrm{LOCO-CV})} = \frac{1}{\sum_{s=1}^{S} N_s} \sum_{s=1}^{S} N_s \frac{\sum_{b=1}^{B} w_s^b \epsilon_s^b}{\sum_{b=1}^{B} w_s^b}, \tag{16}$$

by simply limiting to the relevant subpopulation rather than the full population. Similarly we can estimate the CRPS for the subpopulaton by modifying (14) to similarly focus on the relevant subpopulation:

$$\widehat{\mathrm{CRPS}}_S^{(\mathrm{LOCO-CV})}(y, n, \hat{\theta}_{RW}, \hat{\theta}'_{RW}) = \frac{1}{BN} \sum_{b=1}^{B} \left(\frac{1}{2} \left| \sum_{s=1}^{S} N_s g_s(\hat{\theta}_{RW}^b, \hat{\theta}_{RW}^{\prime b}) \right| - \left| \sum_{s=1}^{S} N_s h_s(y_s, n_s, \hat{\theta}_{RW}^b) \right| \right), \tag{17}$$

using the same importance resampling technique as described previously.

It is common for multiple subpopulations to be estimated, most often all levels of a variable (e.g., geographic regions). This means there is interest in scoring the goodness of estimates for all levels of a particular variable rather than one specific level. To describe this we use l to denote a particular level of the k^{th} variable, with the total number of levels in this variable L. Where we it is necessary to identify a specific variable we denote with a superscript (k) but otherwise assume this is clear from the context for simplicity. For a particular score, it is possible to plot and consider Score_l, for every l, but it is also possible to consider the average of scores for every level:

$$\text{Score}^{(k)} = \frac{\sum_{l=1}^{L^{(k)}} \text{Score}_l}{L^{(k)}}, \tag{18}$$

which in the case of the score being squared error, corresponds to the mean squared error over levels of k.

Figure 3 shows a comparison between the mean over levels of the estimated PSIS-LOCO score for each variable and the true mean over levels. Focusing on the two variables that are strongly predictive of the outcome (X_2 and X_4; final two columns) we can cleanly distinguish between models with a low mean true score and models with a high mean true score. This was a demonstrated challenge in Kuh et al. (2023) that these new metrics overcome. Of particular interest for the practitioner is that the models with the lowest true score differ for the different variables, which again reinforces the concept proposed by Kuh et al. (2023) that the best model differs based on the overall aim.

Turning our attention to the variables that are less predictive of the outcome (X_1 and X_3; first two columns), we see less clear delineation between good and bad models. However, we still retain the respective ordering despite the noisiness of estimates. The respective score of each level can also be plotted individually; see supplementary materials.

5 Reference validation score

To use our scoring decomposition we require at least one observation in each cell. However, in many real-world cases not all cells will be observed in the sample. To overcome this, we follow the approach of Vehtari & Ojanen (2012) and substitute the observed cellwise probability with an estimated probability using a reference model denoted M_*. To investigate this, we first consider a full reference validation score where we make this substitution for every cell (Section 5.1), which allows us to compare to cross validation. Then we describe a new simulation scenario where not all cells are observed. We first demonstrate PSIS LOCO reference validation, and then propose a LOCO combined validation approach in which we combine cross validation and reference validation. As a consequence of this, we describe a way of validating the chosen reference model.

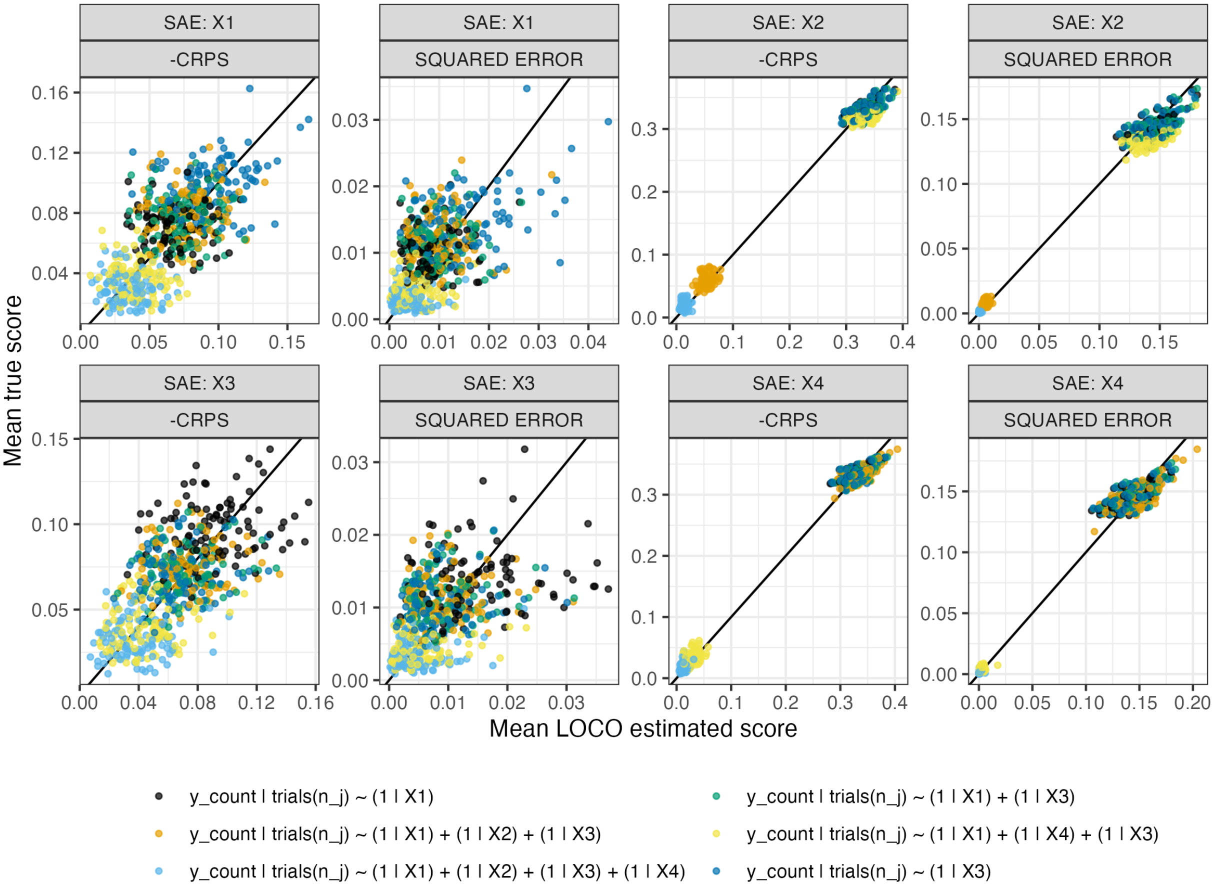

Figure 3: *Comparison of the mean of true scores across levels (y-axis) against the mean of estimated scores (x-axis, alternating vertical facets) for each level in each variable (horizontal and vertical facets). Colour of the point represents the different models fitted, whilst different points within this represent different simulation iterations. The black line represents an identity line. These results can be separated into two classes, those estimating levels of X2 and X4 (strongly predictive of outcome) and those estimating levels of X1 and X3 (weakly predictive of outcome).*

5.1 All cells observed

The reference validation technique follows the approach of Vehtari & Ojanen (2012). We denote the reference model as M_*, and compare to a candidate model, M_c, with the overall aim to evaluate to rank the C candidate models and identify candidate models that perform similarly to the reference model. The reference model can be difficult to propose (as discussed by Vehtari & Ojanen 2012), but in an MRP context there is often need to identify if a smaller or simpler model would suffice . In this paper the model with X_4 could be considered of similar efficacy to a model with all predictors.

We can apply a reference validation approach by substituting the reference model estimate for the population truth in (4) and (5). In Equation (30) of Vehtari & Ojanen (2012), the expected loss function under the actual belief model M_* is decomposed in the variance for a predicted value given M_* and the squared difference in expectations between the reference model M_* and the proposed model M_c. As mentioned by Vehtari & Ojanen (2012) the model minimising the squared loss is that which the predicted expectation is closest to that with the reference model. Incorporating this into (4) in an MRP context, we would aim to minimise the square of

$$\text{Error}^{(\text{REF})}(M_c, M_*) = E_P(Y|M_c, \theta_c) - E_P(Y|M_*, \theta_*),$$

which can be expanded in a similar form as in previous sections to

$$\text{Error}^{(\text{REF})}(M_c, M_*) = \frac{\sum_{j=1}^{J} N_j(\Pr(Y_{i\in j}|\theta_c) - \Pr(Y_{i\in j}|\theta_*))}{\sum_{j=1}^{J} N_j}, \tag{19}$$

with $\Pr(Y_{i\in j}|\theta_c)$ shorthand for $\Pr(Y_{i\in j} = 1|M_c, \theta_c, \mathbf{x}, y)$.

Similarly we use Lemma (17) of Székely & Rizzo (2005) to measure the difference between the predictive distribution of M_* and M_c. This lemma is also used by Gneiting & Raftery (2007) when discussing CRPS:

$$\text{CRPS} = \int_{-\infty}^{\infty} (G(t) - F(t))^2 dt = \frac{1}{2}E|X - X'| + \frac{1}{2}E|Y - Y'| - E|X - Y|,$$

where F and G are the cumulative distribution functions for the reference and candidate models, respectively, and so we use $\Pr(Y|\theta_*)$ as Y and $\Pr(Y|\theta_c)$ as X. Plugging these in and simplifying as in previous sections, we get

$$\begin{aligned}\text{CRPS}^{(\text{REF})}(M_c, M_*) = \frac{1}{BN}\Bigg(&\frac{1}{2}\sum_{b=1}^{B}\left|\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|\theta_c^b) - \Pr(Y_{i\in j}|\theta_c'^b\Big)\right| + \\ &\frac{1}{2}\sum_{b=1}^{B}\left|\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|\theta_*^b) - \Pr(Y_{i\in j}|\theta_*'^b\Big)\right| - \\ &\sum_{b=1}^{B}\left|\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|\theta_c^b) - \Pr(Y_{i\in j}|\theta_*'^b\Big)\right|\Bigg).\end{aligned}$$

We can implement the LOCO estimate of $\Pr(Y_{i\in j}|\theta_c)$, $\Pr(y_{i\in j}|\theta_c, y_{\notin j})$ to avoid overfitting and more accurately capture the predictive power for unseen observations. For squared error, we use the square of

$$\text{Error}^{(\text{LOCO}-\text{REF})}(M_c, M_*) = \frac{\sum_{j=1}^{J} N_j(\Pr\left(Y_{i\in j}|M_c, y_{i\notin j}\right) - \Pr\left(Y_{i\in j}|M_*, y_{i\notin j}\right))}{N}, \tag{20}$$

and for CRPS we have

$$\begin{aligned}\text{CRPS}^{(\text{LOCO}-\text{REF})}(M_c, M_*) = \frac{1}{BN}\sum_{b=1}^{B}\Bigg(\frac{1}{2}\Bigg|\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|\theta_c^b, y_{i\notin j}) - \Pr(Y_{i\in j}|\theta_c'^b, y_{i\notin j}\Big)\Bigg| + \\ \frac{1}{2}\Bigg|\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|\theta_*^b, y_{i\notin j}) - \Pr(Y_{i\in j}|\theta_*'^b, y_{i\notin j}\Big)\Bigg| - \\ \Bigg|\sum_{j=1}^{J} N_j\Big(\Pr(Y_{i\in j}|\theta_c^b, y_{i\notin j}) - \Pr(Y_{i\in j}|\theta_*'^b, y_{i\notin j}\Big)\Bigg|\Bigg).\end{aligned}$$

The PSIS approximation of the LOCO reference model estimate for each cell can be substituted. For the error score, we simply use an importance-weighted estimate for each cell:

$$\begin{aligned}\widehat{\text{Error}}^{(\text{LOCO}-\text{REF})}(M_c, M_*) = \\ \frac{1}{N}\sum_{j=1}^{J} N_j\Bigg(\frac{\sum_{b=1}^{B} w_{c,j}^b \Pr(Y_{i\in j}|\hat{\theta}_c^b)}{\sum_{b=1}^{B} w_{c,j}^b} - \frac{\sum_{b=1}^{B} w_{*,j}^b \Pr(Y_{i\in j}|\hat{\theta}_*^b)}{\sum_{b=1}^{B} w_{*,j}^b}\Bigg).\end{aligned} \tag{21}$$

For the LOCO reference validation CRPS score we can use the importance-resampled posteriors for reference model M_*, $\hat{\theta}_{*,RW}$ and candidate model M_c, $\hat{\theta}_{c,RW}$. We then modify as follows: $g'_{j,\text{REF}}(\hat{\theta}^b_{c,RW}, \hat{\theta}'^b_{c,RW}) = \Pr(Y_{i\in j}|\hat{\theta}_{c,RW}) - \Pr(Y_{i\in j}|\hat{\theta}'^b_{c,RW})$ and $h'_{j,\text{REF}}(\hat{\theta}^b_{c,RW}, \hat{\theta}^b_{*,RW}) = \Pr(Y_{i\in j}|\hat{\theta}^b_{c,RW}) - \Pr(Y_{i\in j}|\hat{\theta}^b_{*,RW})$.

$$\begin{aligned}\widehat{\text{CRPS}}^{(\text{LOCO}-\text{REF})}(M_c, M_*) = \frac{1}{BN}\sum_{b=1}^{B}\Bigg(\frac{1}{2}\Bigg|\sum_{j=1}^{J} N_j g'_{j,\text{REF}}(\hat{\theta}^b_{c,RW}, \hat{\theta}'^b_{c,RW})\Bigg| + \\ \frac{1}{2}\Bigg|\sum_{j=1}^{J} N_j g'_{j,\text{REF}}(\hat{\theta}^b_{*,RW}, \hat{\theta}'^b_{*,RW})\Bigg| - \Bigg|\sum_{j=1}^{J} N_j h'_{j,\text{REF}}(\hat{\theta}^b_{c,RW}, \hat{\theta}^b_{*,RW})\Bigg|\Bigg).\end{aligned} \tag{22}$$

To demonstrate the efficacy of the PSIS LOCO reference validation approach, we let the full model with all covariates be M_* and the candidate models be the model without

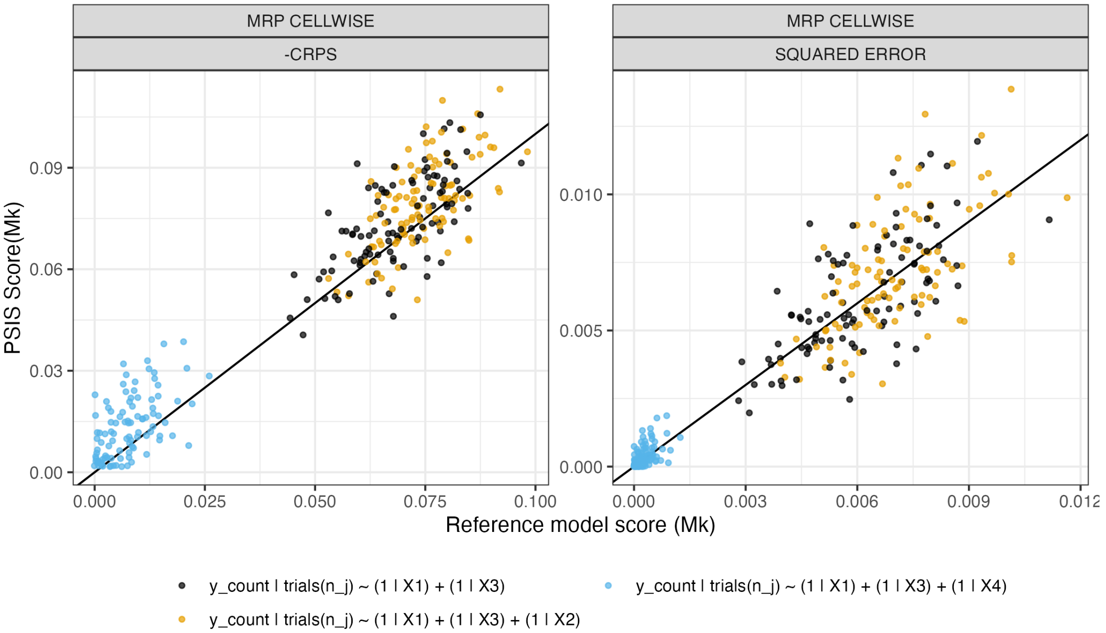

Figure 4: *Comparison of the candidate model LOCO cross validation score (when estimated using Pareto smoothed importance sampling, y-axis) against the LOCO reference validation approach (x-axis). Colour of point represents the different candidate models, whilst different points within this represent different simulation iterations. The black line represents an identity line. The strong relationship indicates that when the reference model is well chosen, this closely approximates the PSIS-LOCO score.*

the precision variable, the model without the bias variable, and the model without either. For each we use a PSIS approximation of the predictive error of cell j and apply this to our previous example with all cells observed. Figure 4 demonstrates that the PSIS LOCO reference validation approach and the PSIS LOCO cross validation approach described in Section 3 produce similar estimates when the appropriate reference model is chosen.

5.2 Not all cells observed

Previously, we demonstrated the equivalence of a reference validation approach to a cross validation approach when all cells are observed. Here, we relax the constraint that all cells are observed. We contribute three components. Firstly we detail how to use the reference validation approach in this case. Secondly, we propose a technique to validate the reference model. Thirdly, we propose a combined validation method using LOCO cross validation and LOCO reference validation.

5.2.1 Reference validation score

We begin by partitioning the population into two sets. The first set, $\mathbb{V}$, is the set of cells where at least one observation in each is observed in the sample. The total number of cells in this set is V and any particular cell in this set is v. The second set, U, is the set of cells that do not have any observations in the sample. The total number of cells in this set is U and any particular cell in this set is u. For clarity note that $J = V + U$.

When cells are observed, we could use a cross validation or reference validation approach to assess predictive power. When cells are not observed, we can only use a reference validation approach. To minimise the reliance on correctly specifying the reference model, we could combine cross validation for observed cells and reference validation for unobserved cells. For simplicity, we write the Pareto smoothed LOCO approximation as

$$\Pr(Y_{i\in j}|\hat{\theta}_c^b, w_{c,j}) = \frac{\sum_{b=1}^{B} w_{c,j}^b \Pr(Y_{i\in j}|\hat{\theta}_c^b)}{\sum_{b=1}^{B} w_{c,j}^b}.$$

and our modified equation as

$$\text{Error}^{(\text{REF})}(M_c, M_*) = \frac{1}{N}\Bigg(\sum_{v=1}^{V} N_v \bigg(\Pr(Y_{i\in v}|\hat{\theta}_c^b, w_{c,v}) - \Pr(Y_{i\in v}|\hat{\theta}_*^b, w_{*,v})\bigg) + \sum_{u=1}^{V} N_u \bigg(\Pr(Y_{i\in u}|\hat{\theta}_c) - \Pr(Y_{i\in u}|\hat{\theta}_*)\bigg)\Bigg). \quad (23)$$

Each component of CRPS can similarly be decomposed into an approximate LOCO reference validation for the cells with observations and approximate LOCO reference validation for the cells with no sample observations. We denote $g_j^{(\text{REF})}(.)$ and $h_j^{(\text{REF})}(.)$ as the corresponding $g_j(.)$ and $h_j(.)$ for the reference model approach.

$$\widehat{\text{CRPS}}^{(\text{REF})}(M_c, M_*) = \frac{1}{BN}\sum_{b=1}^{B}\Bigg(\frac{1}{2}\bigg|\sum_{v=1}^{V} N_v g_v^{(\text{REF})}(\hat{\theta}_c^b, \hat{\theta}_c'^b) + \sum_{u=1}^{U} N_u g_u^{(\text{REF})}(\hat{\theta}_c^b, \hat{\theta}_c'^b)\bigg| + {} + \frac{1}{2}\bigg|\sum_{v=1}^{V} N_v g_v^{(\text{REF})}(\hat{\theta}_*^b, \hat{\theta}_*'^b) + \sum_{u=1}^{U} N_u g_u^{(\text{REF})}(\hat{\theta}_*^b, \hat{\theta}_*'^b)\bigg| - \bigg|\sum_{v=1}^{V} N_v h_v^{(\text{REF})}(\hat{\theta}_c^b, \hat{\theta}_*^b) + \sum_{u=1}^{U} N_u h_u^{(\text{REF})}(\hat{\theta}_c^b, \hat{\theta}_*^b)\bigg|\Bigg). \quad (24)$$

5.2.2 Combined validation score

The challenge with the method described in Section 5.2.1 is that it doesn't use the information that we have observed in the sample to test model goodness. In this section we utilise this information in two ways. Firstly we propose a technique to provide partial validation for the reference model. Secondly, we propose a modification of (23) and (24) to use LOCO reference validation when the cells are not observed and LOCO cross validation when they are.

We begin by focusing on the portion of the population that is observed. We calculate the PSIS LOCO reference validation score and the PSIS LOCO cross validation score by treating the observed cells as a subpopulation; see Section 4. By comparing the resultant model ordering we can identify whether the two approaches suggest similar ordering for the subpopulation where the sample has an observation in each cell.

To demonstrate this, we extend the simulations in Section 2 to relax the assumption that every cell needs to be observed in the sample. Instead we simply constrain that every level of a variable needs to be observed in the sample when taking a sample from the population. With this constraint relaxed, our simulated samples contained 29% of the cells in the population (minimum 25% and maximum 33%).

Figure 5 plots the PSIS LOCO cross validation scores scores against the reference model scores. The first column uses the full model as the reference model, whilst the second column uses the precision model as the reference model. When the reference model is chosen appropriately (the full model) the ordering of scores align, but when an inappropriate reference model is chosen (the precision model), they do not.

However, this only evaluates part of the population and should not be used to score the models. Instead we propose an approach that combines the reference validation score and the cross validation score. We use the reference validation approach when we don't have an observation in a cell, and the cross validation approach when we do. For squared error we modify (23) to

$$\widehat{\text{Error}}^{(\text{CB})}(M_c, M_*) = \frac{1}{N}\left(\sum_{v=1}^{V} N_v \frac{\sum_{b=1}^{B} w_v^b \epsilon_v^b}{\sum_{b=1}^{B} w_v^b} + \sum_{u=1}^{U} N_u \left(\Pr(Y_{i\in u}|\hat{\theta}_c) - \Pr(Y_{i\in u}|\hat{\theta}_*)\right)\right),$$

where w_{obs}^b and ϵ_{obs}^b are as defined in Section 3.3. Similarly we take (24) and modify to use

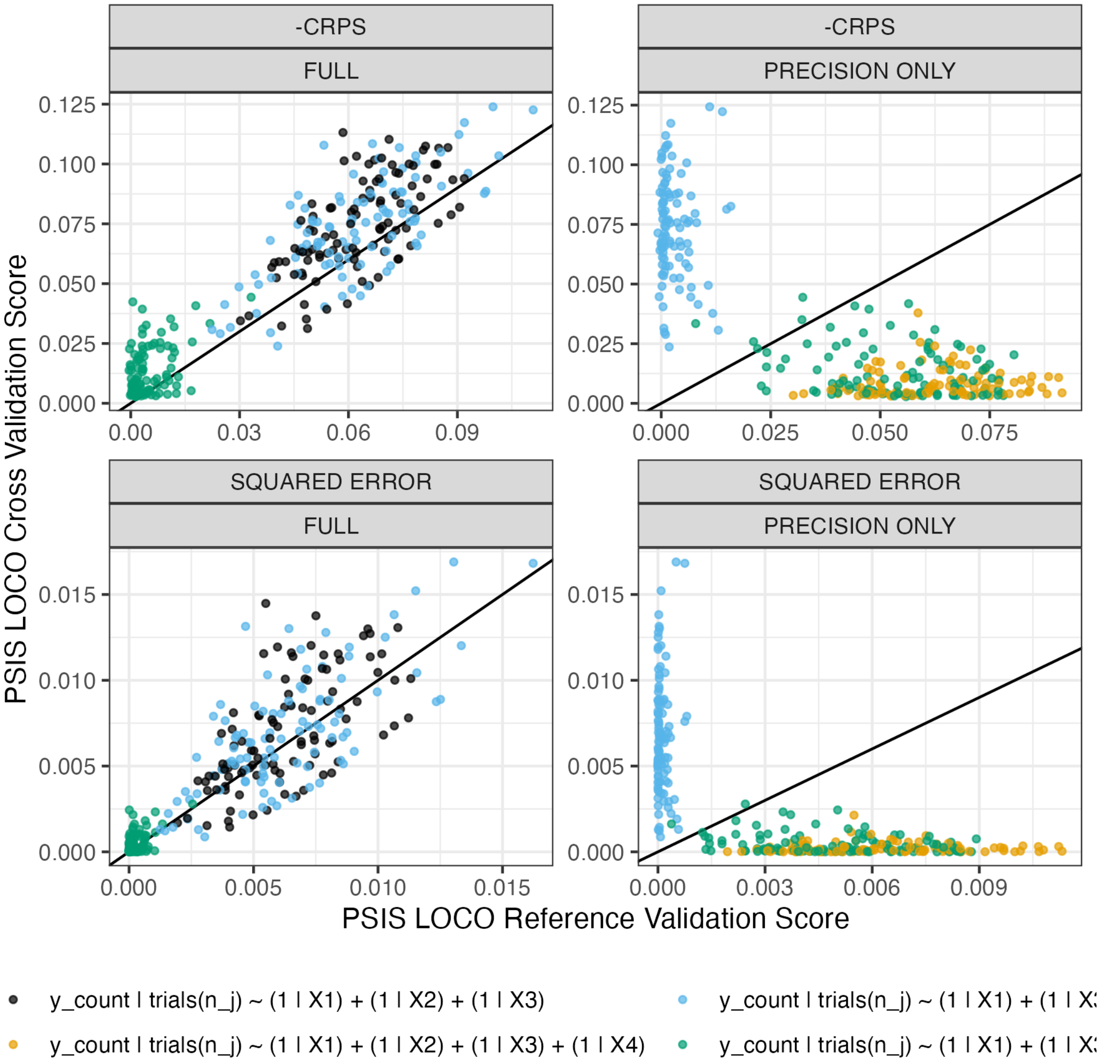

Figure 5: *Comparison of the PSIS LOCO cross validation score (x-axis) against the PSIS LOCO reference validation score (y-axis). Colour of point represents the different candidate models, whilst different points within this represent different simulation iterations. The top row represents the CRPS scores when the reference model is the full model (left panel) and the precision model (right panel). The bottom represents the same for the squared error.*

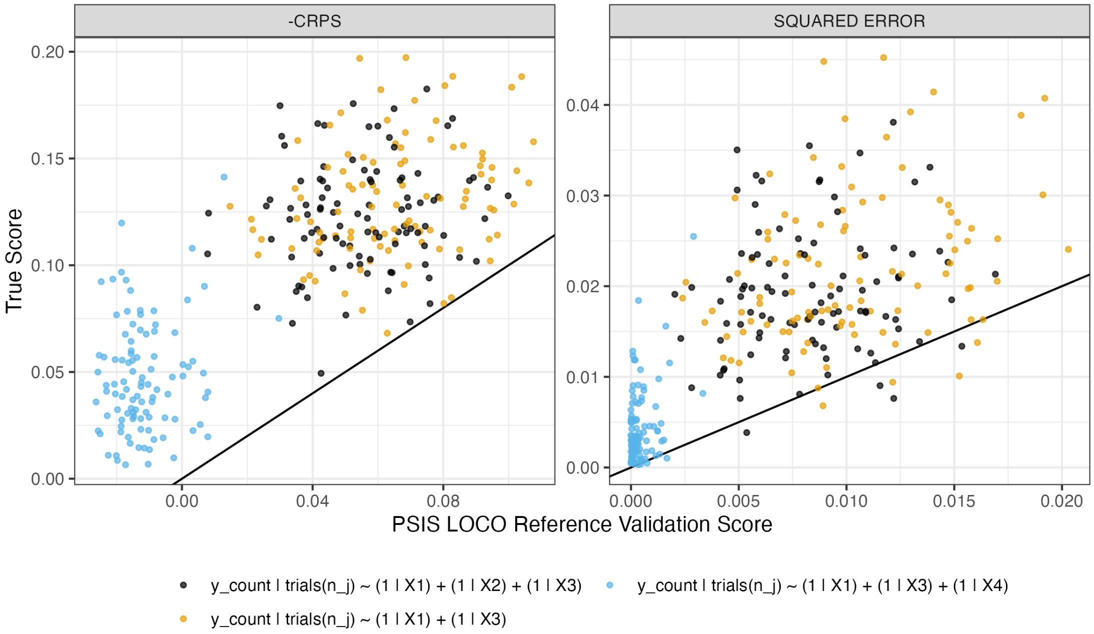

Figure 6: *Comparison of the true score for the estimate (y-axis) against the combined validation approach (x-axis). Colour of point represents the different candidate models, whilst different points within this represent different simulation iterations. Points below the black line represent underestimation of the magnitude for CRPS and above the identity line represent underestimation of the magnitude for squared error.*

$g(.)$ and $h(.)$ as defined in Section 3.3 where the cell is observed:

$$\begin{aligned}
&\widehat{\mathrm{CRPS}}^{(\mathrm{CB})}(M_c, M_*) = \\
&\qquad \frac{1}{BN}\sum_{b=1}^{B}\left(\frac{1}{2}\left|\sum_{v=1}^{V} N_v g_v(\hat{\theta}^b_{c,RW}, \hat{\theta}'^b_{c,RW}) + \sum_{u=1}^{U} N_u g_u^{(\mathrm{REF})}(\hat{\theta}^b_{c,RW}, \hat{\theta}'^b_{c,RW})\right| + \right. \\
&\left. + \ \frac{1}{2}\left|\sum_{v=1}^{V} N_v g_v(\hat{\theta}^b_{*,RW}, \hat{\theta}'^b_{*,RW}) + \sum_{u=1}^{U} N_u g_u^{(\mathrm{REF})}(\hat{\theta}^b_{*}, \hat{\theta}'^b_{*})\right| - \left|\sum_{u=1}^{U} N_u h_u^{(\mathrm{REF})}(\hat{\theta}^b_{c}, \hat{\theta}^b_{*})\right|\right). \qquad (25)
\end{aligned}$$

There is one less term on the final line as there is no distribution for these estimates in the observed cell case.

In Figure 6 we demonstrate the efficacy of this approach against the true model score. Similar to previous results, the magnitude of the score is underestimated for both scores but ordering is maintained.

6 Conclusion

We have consider the challenge of assessing model adequacy and selection in the case where models are being used to predict and then aggregate. Our specific focus is on multilevel regression and poststratification (MRP), but this challenge is relatively common across a range of fields. We propose adaptions to the squared error and continuous ranked probability score (CRPS) and demonstrate that this adaption correctly recovers model ordering at the population and subpopulations, suggesting a successful tool for selecting models. We also provide an approximate leave-one-cell-out approach for fast estimation. We propose a reference validation approach and a combined validation approach for use when not all cells are observed in the sample. Together we feel this provides encouraging progress towards MRP model validation and consider this a considerable improvement on previous model validation attempts (which do not retain correct ordering).

This work provides evidence that failure to correctly recover model ordering is based in the sum of individual scores. However, our method does consistently underestimate the magnitude of the loss, which should be further investigated. Until a solution has been identified, the magnitude of loss should not be directly used as an approximation of true loss, suggesting that caution should be used when evaluating model adequacy. This work used the simulation design proposed by Kuh et al. (2023). Future work could consider different sampling conditions and model comparisons (including comparisons of different priors) to ensure generalisability as this scenario is severe and may not be illustrative of real world applications.

References

Alsalti, T., Hussey, I., Elson, M. & Arslan, R. C. (2023), ‘Using multilevel regression and poststratification to efficiently derive accurate norms’.
URL: *https://osf.io/preprints/psyarxiv/fcm3n*

Bisbee, J. (2019), ‘BARP: Improving Mister P using Bayesian additive regression trees’, *American Political Science Review* **113**(4), 1060–1065.

Bürkner, P.-C. (2017), ‘brms: An R package for Bayesian multilevel models using Stan’, *Journal of Statistical Software* **80**(1), 1–28.

Bürkner, P.-C. (2018), ‘Advanced Bayesian multilevel modeling with the R package brms’, *R Journal* **10**(1), 395–411.

Bürkner, P.-C., Gabry, J., Kay, M. & Vehtari, A. (2023), ‘posterior: Tools for working with

posterior distributions'. R package version 1.4.1.
URL: *https://mc-stan.org/posterior/*

Downes, M., Gurrin, L. C., English, D. R., Pirkis, J., Currier, D., Spittal, M. J. & Carlin, J. B. (2018), 'Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples', *American Journal of Epidemiology* **187**(8), 1780–1790.

Gao, Y., Kennedy, L., Simpson, D. & Gelman, A. (2021), 'Improving multilevel regression and poststratification with structured priors', *Bayesian Analysis* **16**(3), 719–744.

Gelfand, A. E., Dey, D. K. & Chang, H. (1992), 'Model determination using predictive distributions with implementation via sampling-based methods', *Bayesian statistics* **4**, 147–167.

Gelman, A., Hwang, J. & Vehtari, A. (2014), 'Understanding predictive information criteria for Bayesian models', *Statistics and Computing* **24**(6), 997–1016.

Gelman, A. & Little, T. C. (1997), 'Poststratification into many categories using hierarchical logistic regression', *Survey Methodology* **23**, 2127–2136.

Ghitza, Y. & Gelman, A. (2013), 'Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups', *American Journal of Political Science* **57**(3), 762–776.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Kitagawa, G. (1996), 'Monte Carlo filter and smoother for non-Gaussian nonlinear state space models', *Journal of Computational and Graphical Statistics* **5**(1), 1–25.

Kolczynska, M., Bürkner, P.-C., Kennedy, L. & Vehtari, A. (2024), 'Trust in state institutions in Europe, 1989–2019', *Survey Research Method* . In press. Preprint `https://osf.io/preprints/socarxiv/3v5g7`.

Kuh, S., Kennedy, L., Chen, Q. & Gelman, A. (2023), 'Using leave-one-out cross-validation (LOO) in a multilevel regression and poststratification (MRP) workflow: A cautionary tale', *Statistics in medicine* .
URL: *https://doi.org/10.1002/sim.9964*

Lax, J. R. & Phillips, J. H. (2009a), 'Gay rights in the states: Public opinion and policy responsiveness', *American Political Science Review* **103**(3), 367–386.

Lax, J. R. & Phillips, J. H. (2009b), 'How should we estimate public opinion in the states?', *American Journal of Political Science* **53**(1), 107–121.

Little, R. J. & Vartivarian, S. (2005), 'Does weighting for nonresponse increase the variance of survey means?', *Survey Methodology* **31**(2), 161–168.

Liu, Y., Gelman, A. & Chen, Q. (2023), 'Inference from non-random samples using Bayesian machine learning', *Journal of Survey Statistics and Methodology* **11**(2), 433–455.

Lopez-Martin, J., Phillips, J. H. & Gelman, A. (2022), 'Multilevel regression and poststratification case studies'.
URL: *https://bookdown.org/jl5522/MRP-case-studies/*

Lumley, T. & Scott, A. (2015), 'AIC and BIC for modeling with complex survey data', *Journal of Survey Statistics and Methodology* **3**(1), 1–18.

Machalek, D. A., Vette, K. M., Downes, M., Carlin, J. B., Nicholson, S., Hirani, R., Irving, D. O., Gosbell, I. B., Gidding, H. F., Shilling, H. et al. (2022), 'Serological testing of blood donors to characterise the impact of COVID-19 in Melbourne, Australia, 2020', *PLoS One* **17**(7), e0265858.

Ornstein, J. T. (2020), 'Stacked regression and poststratification', *Political Analysis* **28**(2), 293–301.

Park, D. K., Gelman, A. & Bafumi, J. (2006), State-level opinions from national surveys: Poststratification using multilevel logistic regression, *in* J. E. Cohen, ed., 'Public Opinion in State Politics', Stanford University Press, pp. 209–228.

Székely, G. J. & Rizzo, M. L. (2005), 'A new test for multivariate normality', *Journal of Multivariate Analysis* **93**(1), 58–80.

Vehtari, A., Gelman, A. & Gabry, J. (2017), 'Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC', *Statistics and Computing* **27**, 1413–1432.

Vehtari, A., Gelman, A., Gabry, J. & Yao, Y. (2021), 'loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models', *R package* .

Vehtari, A. & Ojanen, J. (2012), 'A survey of Bayesian predictive methods for model assessment, selection and comparison', *Statistics Surveys* **6**, 142–228.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y. & Gabry, J. (2024), 'Pareto smoothed importance sampling', *Journal of Machine Learning Research* **25**. In press.

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2015), 'Forecasting elections with non-representative polls', *International Journal of Forecasting* **31**(3), 980–991.