A problem with the use of cross-validation for selecting among multilevel models

Wei Wang[1] and Andrew Gelman[1,2]

[1]Department of Statistics, Columbia University, New York, NY, USA
[2]Department of Political Science, Columbia University, New York, NY, USA

October 24, 2013

Abstract

As a simple and compelling approach for estimating out-of-sample predictive error, cross-validation naturally lends itself to the task of model comparison. However, we feel that the legitimacy of cross-validation methods in model comparison is often being taken for granted. In this work, we want to clarify what cross-validation methods are measuring when they are used for model comparison. Using a hierarchical model fit to large survey data with a battery of questions, we show that even though cross-validation might give good estimates of out-of-sample performance, it is not always a sensitive instrument for model comparison.

1 Introduction

1.1 Cross-validation for Hierarchical Models

Cross-validation is a widely-used method for estimating out-of-sample predictive error of statistical models. By fitting the model on the training data set and then evaluating it on the testing set, the over-optimism of using data twice is avoided. Naturally, it is also a popular tool for model comparison, especially when the main focus is on predictive power. It is customary for researchers to resort to cross-validation when there are no clear alternatives. Furthermore, attempts have been made to use cross-validated objective functions for statistical inference (Craven and Wahba, 1978; Seeger, 2008), thus integrating out-of-sample predictive error estimation and model selection into one step.

However, for multilevel data (as well as other structures such as arise with time series, spatial, or network data), several challenges arise in the use of cross-validation for estimating out-of-sample predictive error and model selection. The first challenge is the lack of clear

protocol of the cross-validation procedure: to truly test the model, the holdout set cannot be a simple random sample of the data but instead needs to have some multilevel structure itself, so that entire groups as well as individual observations are held out. The second challenge is that, in multilevel models, the observed loss function for data-level cross-validation can be so close to flat that the cross-validation estimates of predictive errors under candidate models can be swamped by random fluctuations.

We demonstrate these issues via a set of multilevel models fit to a large cross-tabulated national survey. An innovative aspect of our analysis is that we evaluate separately on 71 different survey responses, taking each in turn as the outcome in a comparison of regression models. This allows us to construct a relatively large corpus of data out of a single survey.

Multilevel models are effective in survey research, as partial pooling can yield accurate state-level estimates from national polls (Gelman and Hill, 2007). Multilevel models have been successfully applied both to representative and non-representative surveys to obtain accurate small-area estimation and prediction (Fay and Herriot, 1979; Lax and Phillips, 2009; Ghitza and Gelman, 2013; Wang et al., 2013), and the practical application of such methods is currently being actively discussed in social science research (Buttice and Highton, 2013; Lax and Phillips, 2013). In the present paper, we conduct model selection procedures based on k-fold cross-validation and find that under this framework, the improvement of the multilevel models over classical models is surprisingly small. Furthermore, we demonstrate that this lack of notable improvement is related to the sample size and structure of the particular data set by repeating the analysis on simulated data sets that vary in terms of these two factors.

Our findings suggests that researchers should be cautious in using and interpreting cross-validation-based model selection procedures in presence of multilevel structure.

1.2 Quick Calculation of Expected Predictive Errors in Binary-data Regressions

What sorts of improvements in terms of expected predictive loss can we should expect to find from improved models applied to public opinion questions? We can perform a back-of-envelope calculation. Suppose that we only have one cell, with true proportion 0.4, and the good model gives a posterior estimate of log proportion at $\log(0.41)$, with poorer models giving estimates of $\log(0.44)$ (overestimating) and $\log(0.38)$ (underestimating). Then using log loss, the predictive loss under the good model is $-[0.4\ \log(0.41)+0.6\ \log(0.59)] = 0.67322$, and under the others is $-[0.4\ \log(0.44) + 0.6\ \log(0.56)] = 0.67386$ and $-[0.4\ \log(0.38) + 0.6\ \log(0.62)] = 0.67628$.

In this example, the improvement of the predictive loss is between 0.0006 and 0.003 per observation. The lower bound is given by $-[0.4\ \log(0.4) + 0.6\ \log(0.6)] = 0.67301$, so the potential gain from moving to the best possible model in this case is only 0.0002.

These differences in expected predictive error are tiny, implying that they would hardly be noticed in a cross-validation calculation unless the number of observations in the cell were huge (in which case, no doubt the analysis would be more finely grained and there would not

be so many data points per cell). At the same time, a chance in prediction from 0.38 to 0.41, or from 0.41 to 0.44, can be meaningful in a political context. For example, Mitt Romney in 2012 won 38% of the two-party vote in Massachusetts, 41% in New Jersey, and 44% in Oregon; these differences are not huge but they are politically relevant, and we would like a model to identify such differences if it is possible from the data.

The above calculations are idealized but they gives a sense of how real differences can correspond to extremely small changes in predictive loss for binary data.

2 Model Assessment and Selection via Cross-Validation

2.1 Predictive Loss

We start with a loss function $l(\tilde{y}, a)$ corresponding to the inferential action a_M based on a model M, in face of future observations $\tilde{y}$. The available data, typically consisted of predictors y and outcomes x, are labeled as D. The corresponding predictive loss is then,

$$PL(p^t, M, D) = E_{p^t} l(\tilde{y}, a_M) = \int l(\tilde{y}, a_M) p^t(\tilde{y}) d\tilde{y}, \tag{1}$$

where $p^t(\cdot)$ is the true distribution from which the future observations $\tilde{y}$ are generated.

The predictive loss is affected by the form of the inferential action a_M, the loss function l, and the data D. For example, a_M could be the mean of the posterior predictive distribution and l the mean square error loss. However, it is often convenient and theoretically desirable to use the whole posterior predictive distribution as the inferential action and a logarithmic loss function. In addition, using the whole posterior predictive distribution has a Bayesian justification, as it reflects the full inferential uncertainty conditional on the model (Vehtari and Ojanen, 2012). Substituting the choice of a_M and l into (1) yields,

$$PL(p^t, M, D) = E_{p^t} \log p(\tilde{y}|D, M) = -\int p^t(\tilde{y}) \log p(\tilde{y}|D, M) d\tilde{y} \tag{2}$$

This quantity will be the cornerstone of the model selection framework. The fundamental difficulty in estimating this quantity is that the true distribution $p^t(\cdot)$ is unknown.

Another important quantity arises when we approximate the true distribution with the empirical distribution, which gives the training loss,

$$TL(M, D) = -\int \log p(y|D, M) d\hat{F}(y) = -\frac{1}{N} \sum_{y \in D} \log p(y|D, M). \tag{3}$$

The training loss uses the same data for both estimation and evaluation and so in general underestimates prediction error.

2.2 Predictive Error

With (2), the model selection task is straightforward. Among the candidate models, the best model under this framework is the one that minimizes the predictive loss:

$$-\min_{M} \int p^t(\tilde{y}) \log p(\tilde{y}|D, M) d\tilde{y}, \tag{4}$$

which has a lower bound, $-\int p^t(\tilde{y}) \log p^t(\tilde{y}) d\tilde{y}$, which is the entropy of the true distribution. It is often more informative to look at the excess of the predictive loss over this lower bound, as shown in (5). We call this quantity the predictive error. Conceptually, predictive error indicates how far the posterior predictive distribution is from the oracle. In fact, this quantify is the Kullback-Leibler divergence between the posterior predictive distribution of the candidate model and the true generative model. As its form suggests, predictive error is the difference between log posterior predictive density and log true predictive density, averaged over the true predictive distribution,

$$PE(p^t, M, D) = PL(p^t, M, D) - LB(p^t) \tag{5}$$

$$= -\int p^t(\tilde{y}) \log p(\tilde{y}|D, M) d\tilde{y} + \int p^t(\tilde{y}) \log p^t(\tilde{y}) d\tilde{y}. \tag{6}$$

So to estimate the predictive error, we need to estimate the two terms in (5).

2.3 k-fold Cross-Validation for Estimating Predictive Loss

In the predictive framework, the central obstacle of estimating the predictive loss (2) is that the future observations are not available. Simple sample reuse methods of estimating predictive loss introduce downward bias and is prone to over-fitting. One thread of research attempts to estimate and correct this bias and thus gives rise to various information criteria, whose validity hinges on a number of assumptions and simplifications. Another thread of research is to use hold-out data for testing, thus making training and testing data independent. This leads to a variety of resampling procedures, including leave-one-out cross-validation, k-fold cross-validation, Monte Carlo cross-validation, and bootstrapping. In practice, k-fold cross-validation is popular due to its computational convenience and stability (Kale et al., 2011). Formally, the k-fold cross-validation of the predictive loss is given by

$$\widehat{PL}^{\mathrm{CV}}(M, D) = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathrm{test}_k} \log p(y_i|D^k, M) = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i|D^{(\setminus i)}, M), \tag{7}$$

in which D^k represents the k^{th} training set and $D^{(\setminus i)}$ denotes the training set that doesn't include the i^{th} observation. Because the k-fold cross-validation does not use all the data, the predictive error estimates are biased, but in the cases where there are relatively few predictors, this bias is not essential (Burman, 1989).

The practical impediment of using cross-validation is the computational burden: with k-fold cross-validation, we need to fit the model k times. However, it is often possible to use perform the k steps in parallel.

2.4 Surrogate for Lower Bound of Predictive Loss

Now the problem remains as to how to estimate the second term in (5), namely the lower bound of predictive loss. In this paper, we will use the in-sample training loss $TL(M_{\mathrm{s}}, D)$ of the saturated model M_s as the surrogate for the lower bound. So the estimated predictive error is given by

$$\widehat{PE}(M, D) = \widehat{PL}^{\mathrm{CV}}(M, D) - TL(M_{\mathrm{s}}, D) \tag{8}$$

$$= -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i | D^{(\setminus i)}, M) + \frac{1}{N}\sum_{y \in D} \log p(y | D, M_{\mathrm{s}}) \tag{9}$$

2.5 Cross-Validation of Structured Data

Standard cross-validation procedure assumes that data are independent and with no distributional differences between the training and testing sets. For structured data, it is not always clear how to apply the cross-validation procedure. Burman et al. (1994) discusses a modification of ordinary cross-validation procedure for stationary time series. In this paper, we focus on the cross-tabulated structure, which is the central characteristic of survey data. In an unbalanced cross-tabulated data set, simple random sampling might result in under-sampling of small cells. Thus, we adopt a stratified sampling approach to guarantee that each cell is partitioned into a training part and a testing part. Another possibility is to do a cluster sampling and train the model on some cells and test the fitted model on other cells. This approach is related to transfer learning (Pan and Yang, 2010). In the analysis of survey data, the focus is mostly on the existing cells rather than on hypothetical new cells, and so we only discuss cross-validation procedures using stratified sampling on structured data.

3 Comparing Multilevel Models for Binary Survey Outcomes

The 2006 Cooperative Congressional Election Survey, the demonstrative data set in this paper, is a national stratified sample of size 30,000. It includes a wide variety of response outcomes, thus providing an ideal setting to evaluate cross-validation. Although various demographic predictors are available in this data set, we keep our model simple by using only two predictors, state and income. Under this setting, multilevel model is the preferred model over no pooling (saturated model) or complete pooling (additive model). On one hand, the saturated model will trigger over-fitting. On the other hand, it is well observed that income and state have strong interactions on electoral choice (Gelman et al., 2009), so

the additive model must be substantively inadequate.

3.1 Complete Pooling, No Pooling, and Partial Pooling Models

Bayesian multilevel modeling is a natural choice for analyzing cross-tabulated data. When the data provide many explanatory variables, and thus a potentially complex cross-tabulated structure, it is difficult to model the interactions among explanatory variables in classical models, since each single cell is getting sparser and the estimates become unstable. By borrowing strength across cells, a multilevel model (or, alternatively, some other structured model such as a Gaussian process) can produce stable estimates even for cells that have few observations and thus can be viewed as a multivariate regression or interpolation procedure..

We develop our model on a simple two-way cross-tabulation of survey data, with state and income as the two explanatory variables, having J_1 and J_2 levels respectively. We assume no additional continuous predictors in our model. Let N be the total sample size of the survey, then the array of cell counts follows a multinomial distribution,

$$\boldsymbol{N} \sim \text{Multinomial}(N, \boldsymbol{p}),$$

where

$$\begin{aligned}\boldsymbol{N} &= (N_{j_1 j_2})_{J_1 \times J_2},\\ \boldsymbol{p} &= (p_{j_1 j_2})_{J_1 \times J_2}.\end{aligned}$$

The population is thus divided into $J_1 \times J_2$ cells. We constrain our discussion to binary outcomes. Then for a respondent in cell (j_1, j_2), the probability that he or she gives a positive response is $\pi_{j_1 j_2}$, which is modeled using logistic regression:

$$\text{logit}(\pi_{j_1 j_2}) = \boldsymbol{Z}\boldsymbol{\beta},$$

in which $\boldsymbol{Z}$ is the design matrix, and $\boldsymbol{\beta}$ includes the main and interaction effects. Since our goal of inference is on cell proportions $\pi_{j_1 j_2}$ rather than cell assignment probabilities $p_{j_1 j_2}$, we will treat $p_{j_1 j_2}$ as fixed throughout.

Under this setup, we consider three models:

- Complete pooling: $\pi_{j_1 j_2} = \text{logit}^{-1}\left(\beta_{j_1}^{\text{state}} + \beta_{j_2}^{\text{inc}}\right)$
- No pooling: $\pi_{j_1 j_2} = \text{logit}^{-1}\left(\beta_{j_1}^{\text{state}} + \beta_{j_2}^{\text{inc}} + \beta_{j_1 j_2}^{\text{state*inc}}\right)$
- Partial pooling: $\pi_{j_1 j_2} = \text{logit}^{-1}\left(\beta_{j_1}^{\text{state}} + \beta_{j_2}^{\text{inc}} + \beta_{j_1 j_2}^{\text{state*inc}}\right)$, with $\beta_{j_1 j_2}^{\text{state*inc}} \overset{i.i.d.}{\sim} \Phi(\cdot)$.

In the remaining sections of this paper, we compare the predictive error of these three models under various real data and simulation settings.

3.2 Computation

Ideally we want to do full Bayesian inference on our model, but for computational reasons we are currently using an approximate marginal posterior mode estimate provided by blme (Dorie, 2011) in R. We specify the prior distribution $\Phi(\cdot)$ of the interaction as a normal distribution. We have developed an R package, mrp (Gelman et al., 2012), to streamline the multilevel model fitting and cross-validation procedure.

3.3 Estimation Procedure

For each outcome, we fit a multilevel logistic regression model, with additive, fully-interacted and multilevel models. 5-fold cross-validation is used to estimate predictive loss (using more folds gives essentially identical results.) The lower bound is estimated by the training loss of the saturated model.

Under the aforementioned setting, the cross-validation loss estimate is,

$$\begin{aligned}
\widehat{PL}^{\mathrm{CV}}(M,D) &= -\frac{1}{N}\sum_{k=1}^{K}\sum_{j\in \mathrm{test}_k}\log p(y_j|D^k,M) \\
&= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i,j}[y_{ij}^{\mathrm{test}_k}\log\hat{\pi}_{ij}^{D^k}+(n_{ij}^{\mathrm{test}_k}-y_{ij}^{\mathrm{test}_k})\log(1-\hat{\pi}_{ij}^{D^k})] \\
&= -\frac{1}{N}\sum_{i,j}\sum_{k=1}^{K}[y_{ij}^{\mathrm{test}_k}\log\hat{\pi}_{ij}^{D^k}+(n_{ij}^{\mathrm{test}_k}-y_{ij}^{\mathrm{test}_k})\log(1-\hat{\pi}_{ij}^{D^k})] \\
&= -\frac{1}{N}\sum_{i,j}\left[y_{ij}\overline{\log\hat{\pi}_{ij}}+(n_{ij}-y_{ij})\overline{\log(1-\hat{\pi}_{ij})}\right] \\
&= -\sum_{i,j}\frac{n_{ij}}{N}\left[\pi_{ij}\overline{\log\hat{\pi}_{ij}}+(1-\pi_{ij})\overline{\log(1-\hat{\pi}_{ij})}\right],
\end{aligned}$$

in which $n_{ij}^{\mathrm{test}_k}$ is the number of respondents in cell (i,j) of the k-th testing set, $y_{ij}^{\mathrm{test}_k}$ is the number of respondents who answered yes in cell (i,j) of the k-th testing set, correspondingly, n_{ij} and y_{ij} are the numbers of total respondents and respondents who answered yes in cell (i,j), $\hat{\pi}_{ij}^{D^k}$ is the estimated π_{ij} using the k-th training data set, and $\overline{\log\hat{\pi}_{ij}}$ is the weighted average log posterior proportion from each fold, $\left(\sum_{k=1}^{K}y_{ij}^{\mathrm{test}_k}\log\hat{\pi}_{ij}^{D^k}\right)/y_{ij}$, and $\overline{\log(1-\hat{\pi}_{ij})}$ has the similar form. We can see that cross-validation estimate is approximately a measure of loss under cell proportion distribution $\{\exp(\overline{\log\hat{\pi}_{ij}}),\exp(\overline{\log(1-\hat{\pi}_{ij})})\}$ (here we say "approximately" because these two probabilities don't in general add up to one). The quick calculation in section 1.2 suggests that we should expect to see improvement that is small in scale.

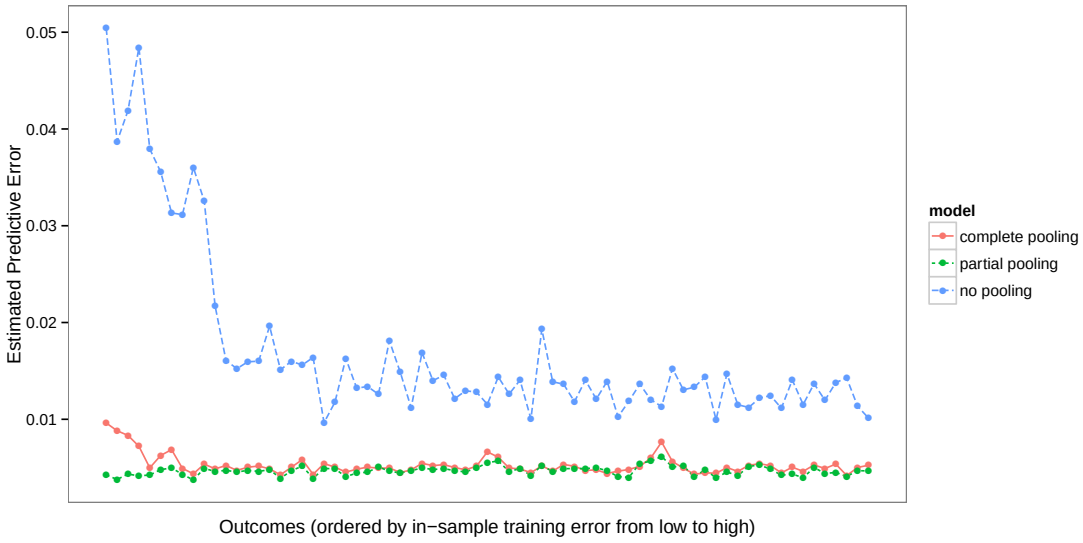

Figure 1: *Measure of fit (estimated predictive error) for all response outcomes in the 2006 Cooperative Congressional Election Survey. Outcomes are ordered by the lower bound (in-sample loss of the saturated model). The no pooling model gives a bad fit. Partial pooling does best but in most cases is almost indistinguishable from complete pooling under the cross-validation criterion.*

4 Results

4.1 Predictive Errors for a Corpus of Outcomes

We begin by estimating the predictive errors of all outcomes in the survey. The results are shown in Figure 1. The x-axis is ordered by the in-sample training loss of the saturated model $TL(M_s, D)$, which we use as a surrogate for a lower bound of predictive loss. For complete pooling and partial pooling, the predictive error stays stable across different outcomes, while the no pooling model has huge predictive error for outcomes with small lower bounds. This finding makes sense since these are the settings where over-fitting is most severe. However, the difference in predictive error between complete pooling and partial pooling seems negligible. Partial pooling is giving essentially the same result as complete pooling, at least according to cross-validation on individual survey responses.

This seems to suggest that partial pooling does not have enough information to estimate cell-to-cell variation, thus giving a overly conservative estimate. Indeed, when we plot the estimates of $\pi_{j_1 j_2}$ for one particular outcome, vote preference for in the congressional election (see the left panel of Figure 2), the estimates from partial pooling are almost identical to those from complete pooling. Even for populous states where, because of their large sample size, the amount of partial pooling should be small, there are no major differences between estimates from partial pooling model and estimates from complete pooling model (see the right panel of Figure 2). This pattern is consistent across different outcomes.

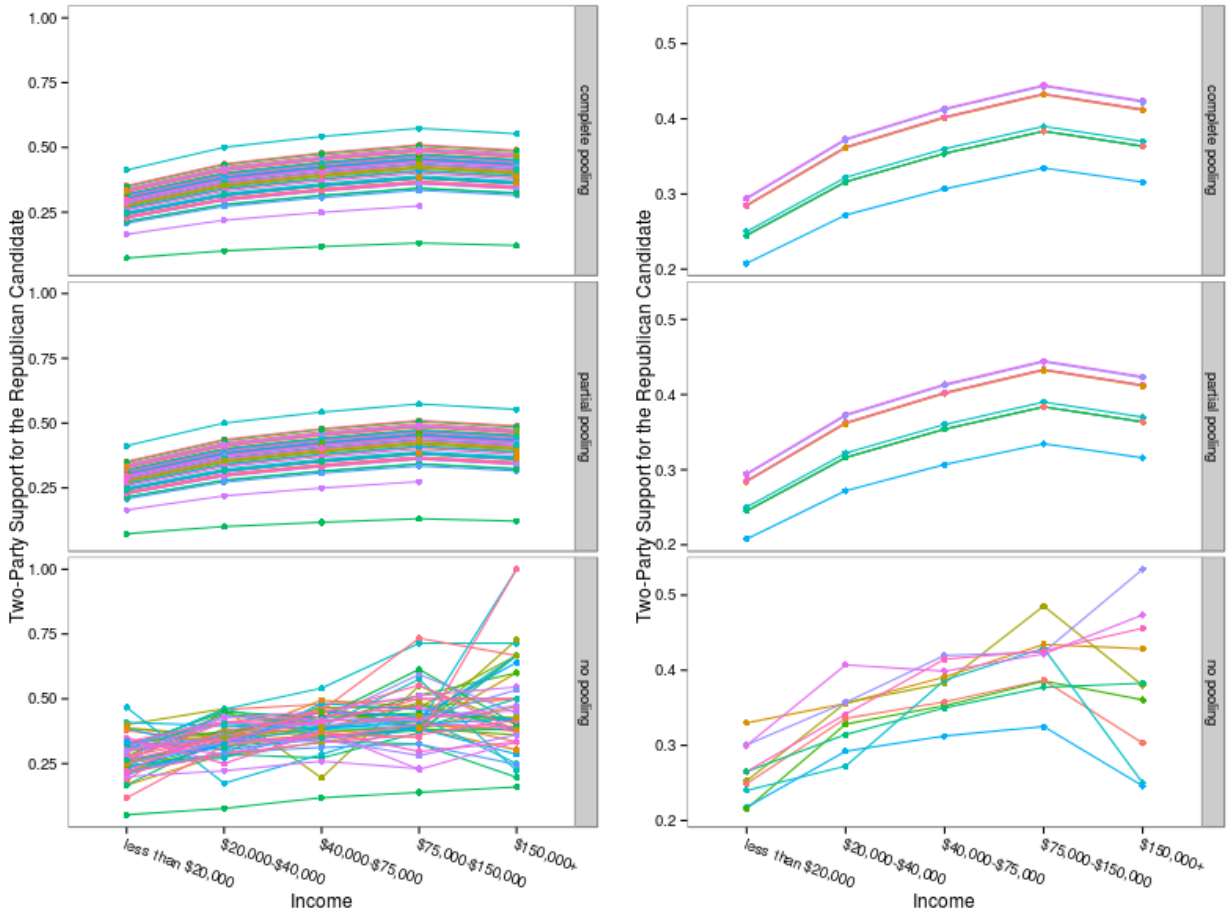

Figure 2: *Left panel: Cell proportion estimates for three models of vote intention for the U.S. House of Representatives. Each line is a state. The partial pooling model pools so much that it is indistinguishable from complete pooling. Right panel: The same estimates for the 10 most populous states. Still, partial pooling estimates are similar to complete pooling estimates.*

Although we believe partial pooling is intrinsically better than complete pooling, it seems that the given data are not sufficient for the partial pooling model to pick up the interaction and unpool the estimates appropriately. It is a result of the particular characteristics of this data set? There are three factors determining the structure of the data that might affect the extent of pooling of the model. First is the sample size. If we increase the sample size to a sufficiently large level, the partial pooling model will be able to partially pool the estimates to an appropriate amount. As sample size grows, the no pooling model will eventually have the same performance as partial pooling, and it might be interesting to see at what point the saturated model becomes acceptable. The second factor affecting the relative performance of the different models is the size of the interactions that are being estimated, and the third factor is the level of imbalance in the hierarchical structure. Survey data classified by demographic and geographic predictors are typically highly unbalanced due to the long tails of sizes typical in taxonomic structures (Mandelbrot, 1955). For example, the 2006 CCES includes 3637 respondents from California but only 131 from Arkansas. This unbalanced structure will affect the amount of pooling performed by a multilevel model.

In the following subsections, we conduct simulations that vary sample size and the structure of the cells to investigate how these factors affect the relative performance of the three models as captured by cross-validation.

4.2 How Sample Size Changes the Dynamics

We artificially augment the data set by combining the data set with itself. New data sets with sample size that are 2, 3 and 4 times as large are generated. This augmentation still maintains the same level of interactions and cell structure as those of the original data. Then we estimate the predictive errors for all outcomes for the three models. Results are plotted in Figure 3. As we expected, as sample size grows, the predictive error of complete pooling model, which is essentially a wrong model, dominates the other two; while the predictive error of no pooling model keeps decreasing. When the sample size is 4 times as large as the original data set, no pooling model has almost the same predictive error as partial pooling model. This makes sense, since the problem of over-fitting eventual goes away if we have sufficiently large sample size and fixed model structure.

These results suggest that for a fixed data structure, partial pooling decisively outperforms no pooling and complete pooling only for a certain window of sample sizes. To have a closer look at the range of the window, we look at one particular outcome, the vote preference in the upcoming election for the U.S. House of Representatives. We augment the sample size and plot the relative performance of the three models in Figure 4. Partial pooling model is noticeably better than complete pooling in this setup when the total sample size exceeds larger than 50,000. Other outcomes have similar patterns.

Figure 3: *Estimated predictive error of all response outcomes for augmented data sets. From top to bottom, the data sets have 2, 3, and 4 times as many data points as the original data set. The outcomes are ordered by the in-sample predictive loss. As sample size grows, complete pooling gradually gets worse and no pooling gets better.*

4.3 Balancedness of the Hierarchical Structure

One possible explanation for the steep learning curve of the partial pooling model is the highly unbalanced structure of the data. Although we have 50 states, the estimate of the covariance of the state random effects might not be reliable since some of the states have small sample sizes. To see how the balancedness of the structure affects the model, we simulate a data set based on partial pooling estimates from the original data set, but make each demographic-geographic cells of roughly the same size. The overall sample size is the same as that of the real data. Relative performance of the three models for all outcomes is plotted in Figure 5. The graph shows that with balanced hierarchical structure, at the same sample size and amount of interaction, partial pooling kicks in much more quickly. Thus partial pooling is consistently better than complete pooling in this scenario. As in the

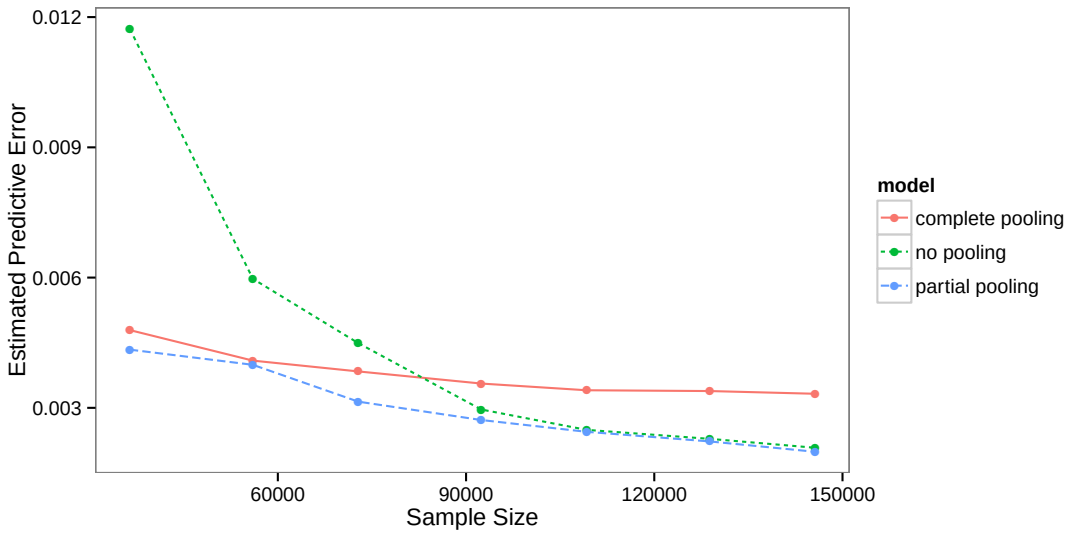

Figure 4: *Predictive error of the three models as sample size grows. The outcome under consideration is partisan vote preference in the upcoming congressional election. By this criterion, partial pooling and complete pooling perform very similarly until sample size exceeds 50,000.*

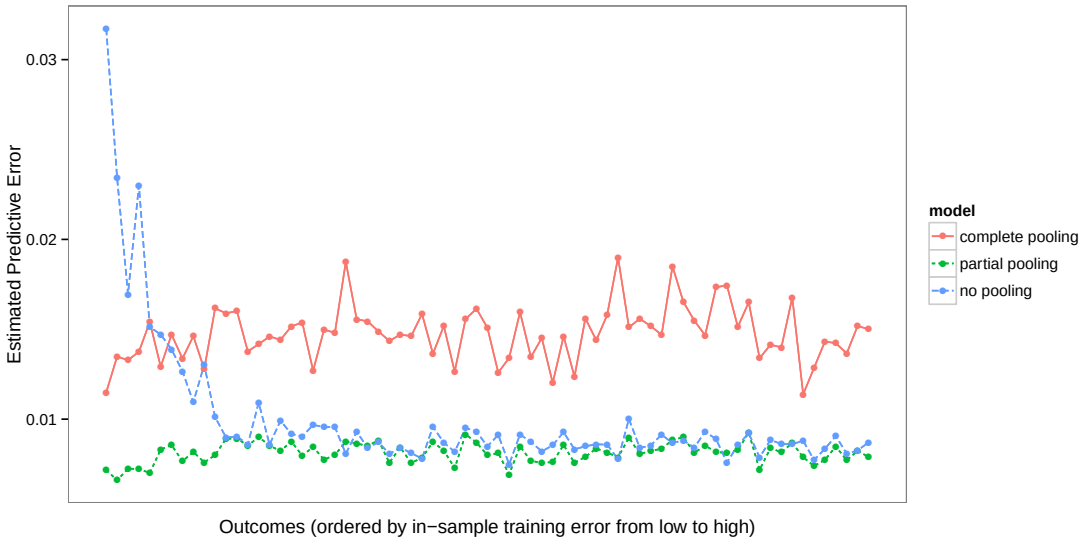

Figure 5: *Measure of fit (predictive error) for all outcomes, ordered by in-sample training loss. The data set is simulated from real data set, and has the same sample size in total as the real data set, but keeping all demographic-geographic cells balanced. In this case, complete pooling model has much higher predictive errors than no pooling and partial pooling. Partial pooling is slightly but consistently better than no pooling. In particular, no pooling model has huge predictive error for outcomes that have smaller in-sample training loss.*

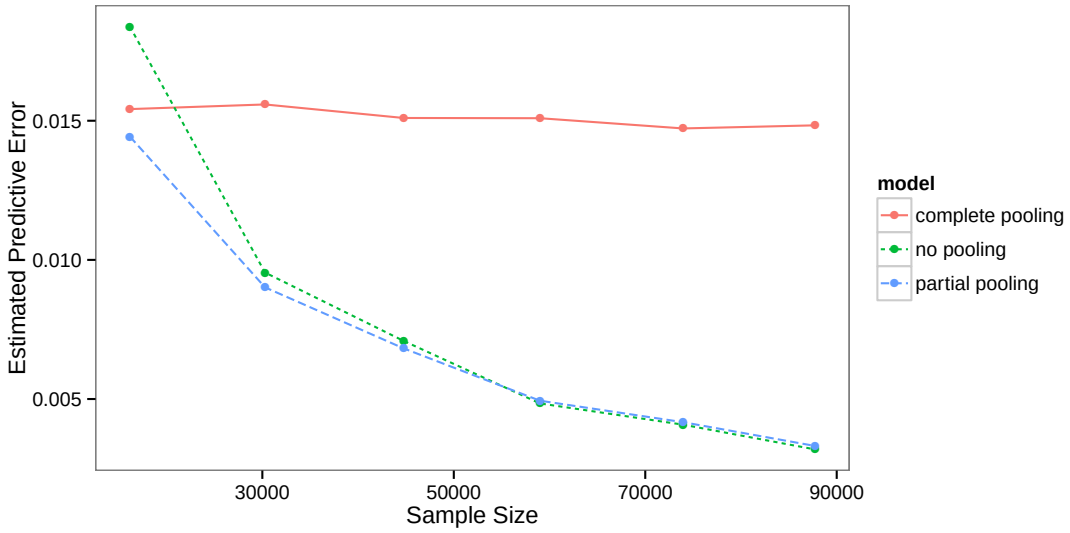

Figure 6: *Predictive error of the three models as sample size grows under the simulated balanced data set. The outcome under consideration is the vote for the Republican candidate in the U.S House of Representatives. Partial pooling has the lowest predictive error when sample size is under 70,000.*

previous analysis, we also look at the relative performance of the three models as sample size grows. The results are plotted in Figure 6.

5 Discussion

Cross-validation is an important tool used to evaluate a wide variety of statistical methods and has been widely used in model comparison when predictive power is of concern. Some theoretical treatments have pointed out situations where cross-validation might have problems. For example, Shao (1993) shows that, under frequentist setting, using leave-one-out cross-validation for linear model variable selection is not consistent. However, the simplicity and transparency of cross-validation gives it a near-universal appeal. In this paper, we investigate the sensitivity of cross-validation as a model comparison instrument in a cross-tabulated multilevel survey data set.

Posed as a model selection problem, three models for this structured data are considered. Two of them are classical models of complete pooling and no pooling, while the other one is a Bayesian multilevel model. The Bayesian multilevel model captures important interactions that are not included in the complete pooling model, while at the same time avoiding the inevitable over-fitting from the no pooling model. However, the improvement of the multilevel model as given by cross-validation is surprisingly tiny, almost negligible to unsuspecting eyes. The problem is that improved fits with binary data yield minuscule improvements in log loss, in moderate sample sizes nearly indistinguishable from noise even if the improved estimates

are substantively important when aggregated (for example, state-level public opinion). Further, real-data-based simulations shows that sample size and structure of the cross-tabulated cells play important roles in the relative margins of different models in cross-validation based model selection. Caution should be exercised in applying cross-validation for model selection with structured data.

References

Burman, P., 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 76, 503–514.

Burman, P., Chow, E., Nolan, D., 1994. A cross-validatory method for dependent data. Biometrika 81, 351–358.

Buttice, M.K., Highton, B., 2013. How does multilevel regression and poststratification perform with conventional national surveys? Political Analysis 21, 449–467.

Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numerische Mathematik 31, 377–403.

Dorie, V., 2011. blme: Bayesian linear mixed-effects models. URL: `http://CRAN.R-project.org/package=blme`.

Fay, R.E., Herriot, R.A., 1979. Estimates of income for small places: an application of james-stein procedures to census data. Journal of the American Statistical Association 74, 269–277.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Gelman, A., Malecki, M., Lee, D., Su, Y.S., Wang, W., 2012. mrp: multilevel regression and poststratification. URL: `http://CRAN.R-project.org/package=mrp`. r package version 0.81-6.

Gelman, A., Park, D.K., Shor, B., Cortina, J., 2009. Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do, second edition. Princeton University Press.

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

Kale, S., Kumar, R., Vassilvitskii, S., 2011. Cross-validation and mean-square stability, in: ICS, pp. 487–495.

Lax, J.R., Phillips, J.H., 2009. How should we estimate public opinion in the states? American Journal of Political Science 53, 107–121.

Lax, J.R., Phillips, J.H., 2013. How should we estimate sub-national opinion using MRP? preliminary findings and recommendations. Presented at Midwest Political Science Association URL: `http://CRAN.R-project.org/package=mrp`.

Mandelbrot, B., 1955. On the language of taxonomy: An outline of a "thermostatistical" theory of systems of categories with willis (natural) structure, in: Cherry, C. (Ed.), Information Theory—Third London Symposium, pp. 135–145.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359.

Seeger, M.W., 2008. Cross-validation optimization for large scale structured classification kernel methods. Journal of Machine Learning Research 9, 1147–1178.

Shao, J., 1993. Linear model selection by cross-validation. Journal of the American statistical Association 88, 486–494.

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

Wang, W., Rothschild, D., Goel, S., Gelman, A., 2013. Forecasting elections with non-representative polls. International Journal of Forecasting .