On Model-Selection and Applications of Multilevel Models in Survey and Causal Inference

Wei Wang

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

On Model-Selection and Applications of Multilevel Models in Survey and Causal Inference

Wei Wang

This thesis includes three parts. The overarching theme is how to analyze multilevel structured datasets, particularly in the areas of survey and causal inference. The first part discusses model selection of hierarchical models, in the context of a national political survey. I found that the commonly used model selection criteria based on predictive accuracy, such as cross validation, don't perform very well in the case of political survey and explore the possible causes. The second part centers around a unique data set on the presidential election collected through an online platform. I show that with adequate modeling, meaningful and highly accurate information could be extracted from this highly-biased data set. The third part builds on a formal causal inference framework for group-structured data, such as meta-analysis and multi-site trials. In particular, I develop a Gaussian Process model under this framework and demonstrate additional insights that can be gained compared with traditional parametric models.

Table of Contents

List of Figures

Acknowledgments

Firstly, I would like to thank my advisors Prof. Michael Sobel and Prof. Andrew Gelman for their continuous support during my PhD study. They have provided tremendous patience, encouragement, and guidance in all the time of research and the writing of the thesis.

Besides, I would like to thank Dr. David Rothschild from the Microsoft Research for providing me an opportunity of internship at Microsoft Research. It was a very intellectually stimulating experience and led to a major part of this dissertation. I would also like to thanks Prof. Lauren Hannah and Prof. John Cunningham for being on my committee and providing insightful comments and encouragements.

Also, I want to express my gratitude to the whole Columbia Department of Statistics, including the faculty, staff, and my fellow graduate students. They made this long journey a happy and meaningful experience.

Last but not the least, I would like to thank my parents for bringing me up and teaching me the importance of integrity, honesty and caring for others.

To my parents and teachers.

Part I

Introduction

Chapter 1

Introduction

Multilevel datasets are ubiquitous in social and behavioral sciences. For example, a national survey intrinsically has respondents belonging to different states, age groups, income brackets, age cohorts etc. In meta-analysis and multi-site randomized experiments, study subjects are grouped with respect to studies/sites, possibly sharing common traits or characteristics. Hierarchical models, by partial pooling group specific coefficients (parametric cases) or functional forms (nonparametric cases), often yield inferential results that are more realistic and less prone to over-fitting.

In this thesis, I discuss three projects related to multilevel datasets in social and behavioral sciences, each constituting one chapter. In chapter 2, which is based on [68], I discuss model-selection issues of hierarchical models. It is widely accepted in machine learning and statistics literature to use out-of-sample prediction as the gold standard of measuring model utilities. However, I demonstrate with an example of a large national survey, that actual predictive accuracy based criteria are often not the most reliable or sensitive tools to compare models. I discuss why out-of-sample prediction fails to select the better models under an extensive set of simulations. This serves as a cautionary tale against the popular one number summary of model comparison, especially when dealing with multilevel data and hierarchical models. In chapter 3, which is based on [69], the motivating example is a political survey data

set collected on an online platform. Being a non-traditional survey, the multilevel structure of the data is highly skewed, i.e., certain groups are heavily over/under-representative compared with the general voter population. How to salvage this dataset and extract meaningful information is the theme of this chapter. In chapter 4, which is an extension of [62], I discuss how to conduct proper causal inference on multilevel data, motivated by examples of meta-analysis and multi-site randomized experiments. A suite of machine learning and non-parametric methods have been entering the toolbox of causal inference, but it is rare to see these modern methods being applied to multilevel data. Based on the extended potential outcome framework developed in a previous paper, I discuss a Gaussian Processes based model for a multi-site randomized experiment in education.

Multilevel datasets provide both opportunities and challenges for statistical analysis of social and behavioral research. I hope the three essays contribute to our understanding and toolbox of tackling multilevel data.

Part II

Cross-validation

Chapter 2

Insensitivity of Predictive Accuracy for Selecting among Multilevel Models

Models selection is an integral part of any data analysis. In an ideal world, iteratively improving and comparing model fits of different specifications should be the routine of all statistical procedures, especially when developments in methodology and computation facilitate evermore sophisticated and complex models. Often, the most important question is not that whether a more complicated model is computational tractable, but why this model is an improvement over the older and simpler ones. Multilevel models (also known as Hierarchical Models) are an example of modern statistical models, which specifically handles data with group structure, for example, a national survey data with geographic and demographic information or an educational intervention applied to different schools and neighborhoods.

The gold standard of model comparison is out-of-sample prediction accuracy, i.e., in the hypothetical case of more observations coming in, which model gives the best prediction of new case of outcomes based on new cases of predictors. Cross-validation is a perhaps the most widely-used method for estimating out-of-sample prediction

error and comparison of statistical models. By fitting the model on the training dataset and then evaluating it on the hold-out testing set, the over-optimism of using data twice is avoided. Furthermore, attempts have been made to use cross-validated objective functions for statistical inference [16, 58], thus integrating out-of-sample prediction error estimation and model selection into one step.

In this chapter, I will discuss several challenges I encounter in using cross-validation predictive accuracy in evaluating and selecting among multilevel models, specifically in binary classification models. The first challenge is the lack of clear protocol for 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. Hierarchical cross-validation can be performed in the context of particular applications [50], but it is not clear how best to subsample structured data for cross-validation in a general way. 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 prediction errors under candidate models can be swamped by random fluctuations.

I focus on the second of these concerns, demonstrating the limitations of prediction error in the context of a set of multilevel models fit to a large cross-tabulated national survey. An innovative aspect of this analysis is that I 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. This chapter is based on a published paper [68], and a collaboration with Andrew Gelman.

2.1 Multilevel Models and Survey Research

There are two types of survey researchers, as identified by the classic book "Survey Errors and Survey Costs" [27], the *describers*, who "use surveys to describe char-

acteristics of a fixed population", and the *modelers*, who "seek to identify causes of phenomena constantly occuring in a society". The latter group developed models to generate less biased estimates, as a result of using more data and handling more inherent structure within the data. Multilevel models, an example of the *modeler* approach, are effective in survey research, as partial pooling can yield accurate state-level estimates from national polls [22]. Multilevel models have been successfully applied both to representative and nonrepresentative surveys to obtain accurate small-area estimation and prediction [20, 24, 40, 69], and the practical application of such methods is currently being actively discussed in social science research [9, 40]. In this chapter, I conduct model selection procedures based on k-fold cross-validation and find that under this framework, the improvement of multilevel models over classical models is surprisingly small when measured on the scale of prediction error. Furthermore, I demonstrate that this lack of notable improvement is related to the sample size and data structure by repeating the analysis on simulated datasets that vary in terms of these two factors.

The results illustrate that under multilevel structure, it could be tricky to use cross-validation in model selection, as the size of the data and how balanced the structure is heavily affect the relative performance of the models.

In the next section, I will present a fully Bayesian model comparison framework, a preparation for the real data analysis.

2.2 Model Assessment and Selection via Cross-Validation

2.2.1 Predictive Loss

I 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 consisting

of predictors x and outcomes y, are labeled as D. The corresponding predictive loss is then,

$$PL(p^t, M, D) = E_{p^t} l(\tilde{y}, a_M) = \int l(\tilde{y}, a_M) p^t(\tilde{y}) d\tilde{y} \tag{2.1}$$

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 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 [67]. Substituting the choice of a_M and l into (2.1) yields,

$$\begin{aligned} PL(p^t, M, D) &= E_{p^t}[-\log p(\tilde{y}|D, M)] \\ &= -\int p^t(\tilde{y}) \log p(\tilde{y}|D, M) d\tilde{y} \end{aligned} \tag{2.2}$$

This quantity is central to predictive model selection. The fundamental difficulty in estimating it is that the true distribution $p^t(\cdot)$ is unknown.

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

$$\begin{aligned} TL(M, D) &= -\int \log p(y|D, M) d\hat{F}(y) \\ &= -\frac{1}{N} \sum_{y \in D} \log p(y|D, M). \end{aligned} \tag{2.3}$$

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

2.2.2 Prediction Error

With (2.2), the model selection task is straightforward. Among the candidate models, the best model under this framework is the one that minimizes the predictive loss:

$$-\min_{M} \int p^t(\tilde{y}) \log p(\tilde{y}|D, M) d\tilde{y}, \tag{2.4}$$

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 (2.5). I label this quantity as the prediction error. Conceptually, the prediction error indicates how far the posterior predictive distribution is from the oracle, and it is the Kullback-Leibler divergence between the posterior predictive distribution of the candidate model and the true generative model. As its form suggests, the prediction error is the difference between log posterior predictive density and log true predictive density, averaged over the true predictive distribution,

$$\begin{aligned} PE(p^t, M, D) &= PL(p^t, M, D) - LB(p^t) \\ &= -\int p^t(\tilde{y}) \log p(\tilde{y}|D, M) d\tilde{y} + \int p^t(\tilde{y}) \log p^t(\tilde{y}) d\tilde{y}. \end{aligned} \tag{2.5}$$

So to estimate the prediction error, I need to estimate the two terms in (2.5).

2.2.3 k-fold Cross-Validation for Estimating Predictive Loss

In the predictive framework, the central obstacle of estimating the predictive loss (2.2) is that the future observations are not available. One thread of research attempts to estimate and correct the bias introduced by reusing the sample 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 [35]. Formally, the k-fold cross-validation of the predictive loss is given by

$$\begin{aligned}\widehat{PL}^{\mathrm{CV}}(M,D) &= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in \mathrm{test}_k}\log p(y_i|D^k,M)\\ &= -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|D^{(\setminus i)},M),\end{aligned} \tag{2.6}$$

where D^k represents the k^{th} training set, test_k represents the k^{th} testing set under the random partition and $D^{(\setminus i)}$ denotes the training set that excludes the i^{th} observation. Because k-fold cross-validation does not use all the data, the prediction error estimates are biased, but in the cases where there are relatively few predictors, this bias is small [7].

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

The problem remains of estimating the second term in (2.5), namely the lower bound of predictive loss. In this chapter, I 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 prediction error is

$$\begin{aligned}&\widehat{PE}(M,D) = \widehat{PL}^{\mathrm{CV}}(M,D) - TL(M_{\mathrm{s}},D)\\ &= -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|D^{(\setminus i)},M) + \frac{1}{N}\sum_{y\in D}\log p(y|D,M_{\mathrm{s}}).\end{aligned} \tag{2.7}$$

2.3 Cross-Validation of Structured Data

Standard cross-validation 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 best to perform this partition. [8] discusses a modification of ordinary cross-validation procedure for stationary time series. In this chapter, I focus on the cross-tabulated structure, which is the characteristic of survey data with discrete responses. In an unbalanced cross-tabulated dataset, simple random sampling might result in undersampling of small cells. Thus, I adopt a stratified sampling approach to guarantee that each cell is partitioned into a training part and a testing part. Another possibility is to perform a cluster sampling and train the model on some cells and test the fitted model on others. This approach is related to transfer learning [47]. In the analysis of survey data, the focus is mostly on the existing cells rather than on hypothetical new cells, and so I only discuss cross-validation using stratified sampling on structured data.

2.4 Comparing Multilevel Models for Binary Survey Outcomes

The 2006 Cooperative Congressional Election Survey, the example dataset in this chapter, is a national stratified sample of size 30,000 that includes a wide variety of response outcomes, thus providing an ideal setting to evaluate cross-validation. Although various demographic predictors are available in this dataset, I keep this model simple by using only two predictors, state and income. Under this setting, the multilevel model is the preferred model over no pooling (saturated model) or complete pooling (additive model). On one hand, the saturated model will trigger overfitting. On the other hand, income and state are known to have strong interactions when predicting electoral choice [23], so the additive model must be substantively

inadequate.

2.4.1 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..

I develop this 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.[1] I assume no continuous predictors in this 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

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

The population is thus divided into $J_1 \times J_2$ cells. I constrain the discussion to binary

[1]For the 2006 Cooperative Congressional Election Survey dataset, there are 50 states ($J_1 = 50$), and 5 income levels ($J_2 = 5$), including less than \$20,000, \$20,000-\$40,000, \$40,000-\$75,000, \$75,000-\$150,000, and \$150,000+.

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 covariate vector and $\boldsymbol{\beta}$ includes the main and interaction effects. Since the goal of inference is on cell proportions $\pi_{j_1 j_2}$ rather than cell assignment probabilities $p_{j_1 j_2}$, I treat $p_{j_1 j_2}$ as fixed throughout.

Under this setup, I consider three models:

- Complete pooling of interactions:

$$\pi_{j_1 j_2} = \text{logit}^{-1}\left(\beta_{j_1}^{\text{state}} + \beta_{j_2}^{\text{inc}}\right)$$

- 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 $\beta_{j_1 j_2}^{\text{state inc}} \overset{i.i.d.}{\sim} \text{N}(0, \sigma^2)$ where the scale parameter σ is estimated from the data (with a separate value for each survey outcome).

Although nonparametric multilevel modeling, both in the Bayesian [32] and the frequentist [57] perspectives, have been under rapid development, I adopt a linear parametric specification for the multilevel model, because linear parametric models are still the standard specification, and software that fit the routine linear parametric models are widely available and easily accessible to practitioners. In the remaining sections of this chapter, I compare the prediction error of these three models under various real data and simulation settings.

Multilevel models in big-data applications can be much more complicated [24]; I use a relatively simple example here to explore the basic ideas.

2.4.2 Computation

Ideally I want to do full Bayesian inference on the model, but for computational reasons I am currently using an approximate marginal posterior mode estimate provided by blme [18] in R, which is an extension of the widely-used lme4 [5] package. The lme4 package approximately integrates out the random effects to obtain an approximate marginal MLE of the scale parameter and the fixed effects. However, modal estimates can end up on the boundary due to sampling variability [14], which in the case makes the partial pooling model reduce to complete pooling. In blme, the scale parameter σ is also given a gamma prior with shape parameter 2.5 and rate parameter 0. The gamma prior is used to regularize the prior of the scale and pull the estimates of the interactions away from zero, a situation that often happens in modal estimation.

2.4.3 Estimation Procedure

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

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

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

in which $n_{ij}^{\text{test}_k}$ is the number of respondents in cell (i,j) of the k-th testing set, $y_{ij}^{\text{test}_k}$ 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 dataset, and $\overline{\log\hat{\pi}_{ij}}$ is the weighted average log posterior proportion from each fold, $\left(\sum_{k=1}^{K} y_{ij}^{\text{test}_k}\log\hat{\pi}_{ij}^{D^k}\right)/y_{ij}$, and $\overline{\log(1-\hat{\pi}_{ij})}$ has the similar form. The cross-validation loss estimate is approximately a measure of loss under cell proportion distribution $(\exp(\overline{\log\hat{\pi}_{ij}}), \exp(\overline{\log(1-\hat{\pi}_{ij})}))$ ("approximately" because these two probabilities do not in general add up to 1). The quick calculation in section 1.2 suggests that I should expect to see only small improvements in cross-validation loss even from substantively important model improvements.

2.5 Results

2.5.1 Prediction Errors for a Corpus of Outcomes

I begin by estimating the prediction errors of all outcomes in the survey. The results are shown in Figure 2.1. The x-axis is ordered by the in-sample training loss of

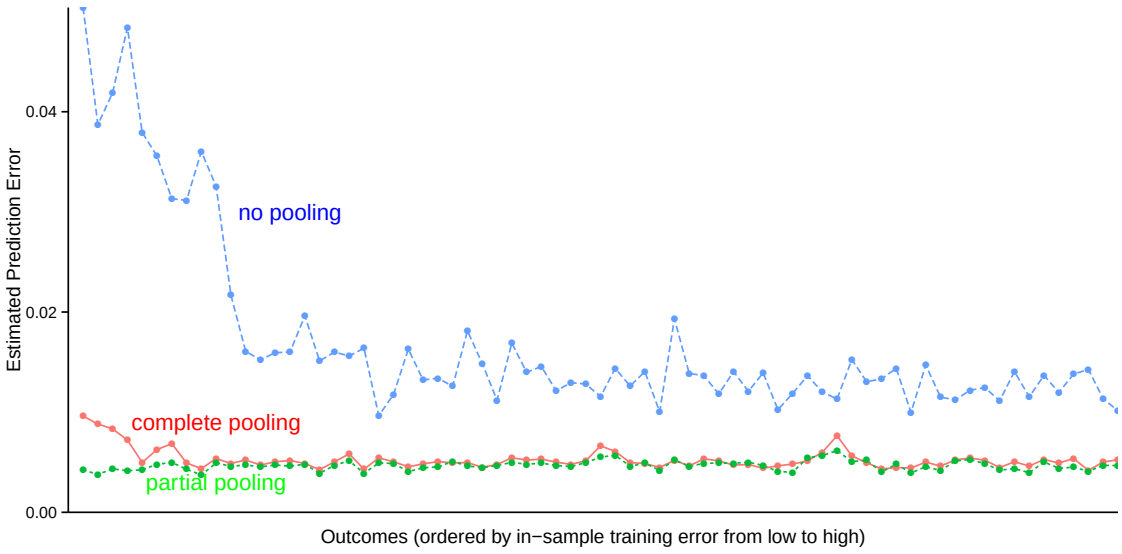

Figure 2.1: *Measure of fit (estimated prediction error) for all response outcomes in the 2006 Cooperative Congressional Election Survey. Outcomes are ordered by the lower bound (in-sample loss of the saturated model). The no pooling model gives a bad fit. Partial pooling does best but in most cases is almost indistinguishable from complete pooling under the cross-validation criterion.*

the saturated model $TL(M_s, D)$, which I use as a surrogate for a lower bound of predictive loss. For complete pooling and partial pooling, the prediction error stays stable across different outcomes, while the no pooling model has huge prediction error for outcomes with small lower bounds. This finding makes sense since these are the settings where overfitting is most severe (saturated models achieve the lowest in-sample training error). However, the difference in prediction 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 an overly conservative estimate. Indeed, when I 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 2.2), 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 2.2). This pattern is consistent across different outcomes.

Although 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 dataset? 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 I 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

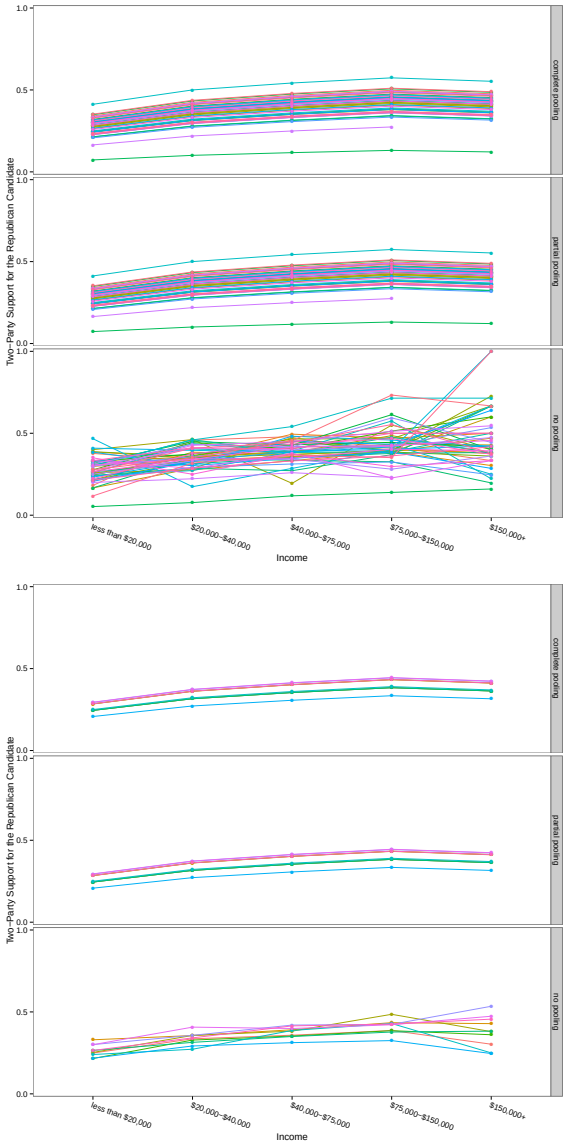

Figure 2.2: *Left panel: Cell proportion estimates for three models of vote intention. Each line is a state. The partial pooling model pools so much that it is indistinguishable from complete pooling. Right panel: The same estimates for the 10 most populous states. Still, partial pooling estimates are similar to complete pooling estimates.*

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 [43]. For example, the 2006 CCES includes 3,637 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, I 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.

2.5.2 How Sample Size Changes the Dynamics

I artificially augment the dataset by combining the dataset with itself. New datasets 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 I estimate the prediction errors for all outcomes for the three models. Results are plotted in Figure 2.3. As I expected, as sample size grows, the prediction error of complete pooling model, which is essentially a wrong model, dominates the other two; while the prediction error of no pooling model keeps decreasing. When the sample size is 4 times as large as the original dataset, no pooling model has almost the same prediction error as partial pooling model. This makes sense, since the problem of overfitting eventual goes away if there are 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, I look at one particular outcome, the vote preference in the upcoming election for the U.S. House of Representatives. I augment the sample size and plot the relative performance of the three models in

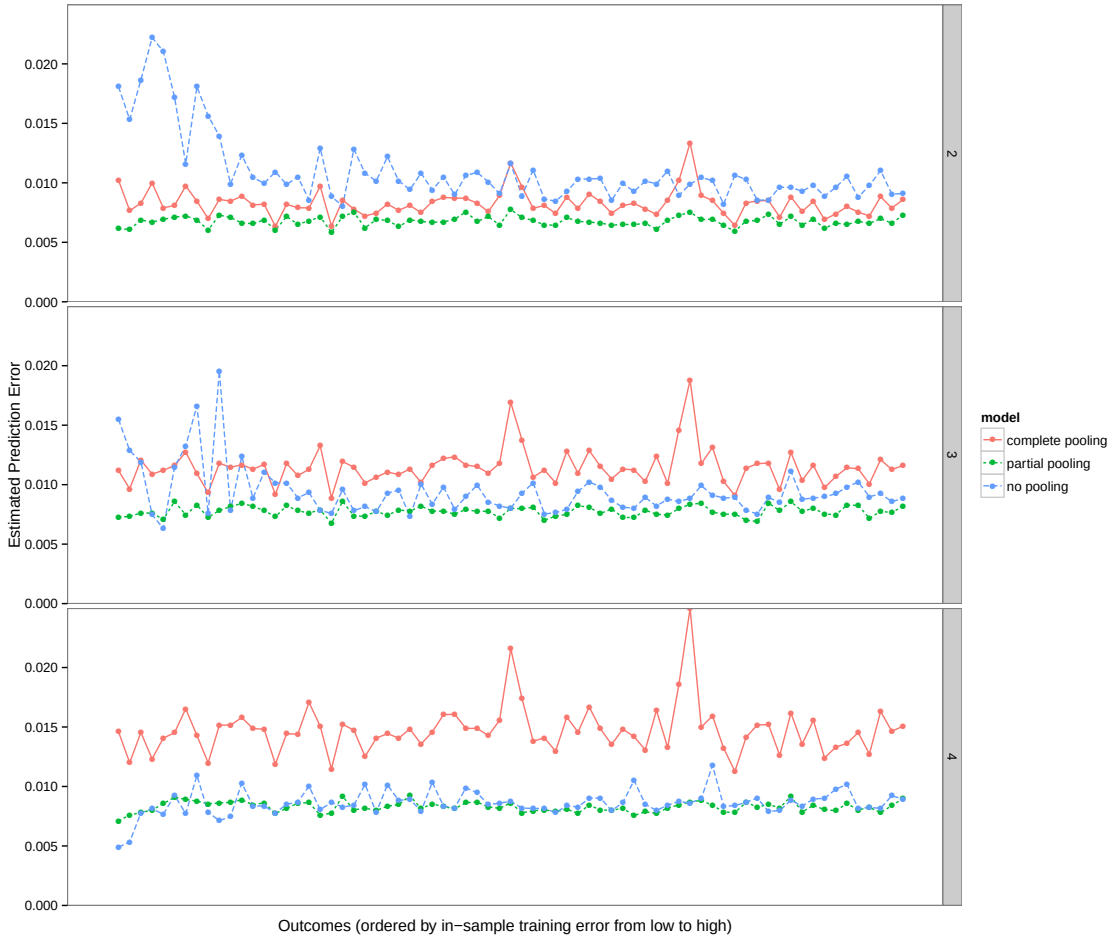

Figure 2.3: *Estimated prediction error of all response outcomes for augmented datasets. From top to bottom, the datasets have 2, 3, and 4 times as many data points as the original dataset. The outcomes are ordered by the in-sample predictive loss. As sample size grows, complete pooling gradually gets worse and no pooling gets better.*

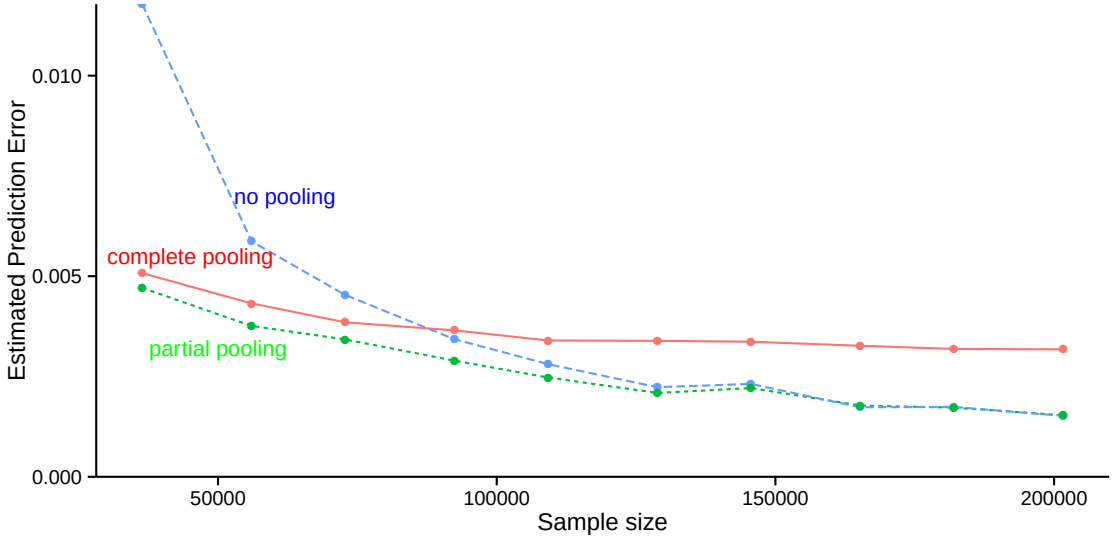

Figure 2.4: *Prediction error of the three models as sample size grows. The outcome under consideration is partisan vote preference in the upcoming congressional election. By this criterion, partial pooling and complete pooling perform similarly until sample size exceeds 50,000.*

Figure 2.4. 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.

2.5.3 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 there are 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, I simulate a dataset based on partial pooling estimates from the original dataset, 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 2.5. 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 previous analysis, I also look at the relative performance of the three models as sample size grows. The results are plotted in Figure 2.6.

2.6 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, [59] shows that, under the 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 chapter, I investigate the sensitivity of cross-validation

Figure 2.5: *Measure of fit (prediction error) for all outcomes, ordered by in-sample training loss. The dataset is simulated from real dataset, and has the same sample size in total as the real dataset, but keeping all demographic-geographic cells balanced. In this case, complete pooling model has much higher prediction errors than no pooling and partial pooling. Partial pooling is slightly but consistently better than no pooling. In particular, no pooling model has huge prediction error for outcomes that have smaller in-sample training loss.*

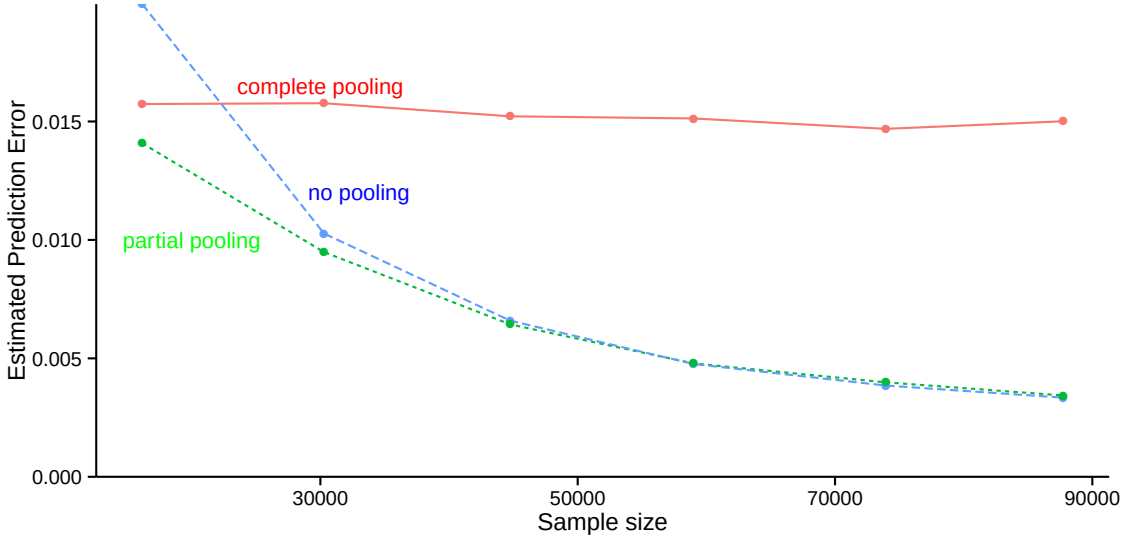

Figure 2.6: *Prediction error of the three models as sample size grows under the simulated balanced dataset. The outcome under consideration is the vote for the Republican candidate in the U.S House of Representatives. Partial pooling has the lowest prediction error when sample size is under 70,000.*

as a model comparison instrument in a cross-tabulated multilevel survey dataset.

I set up the model selection problem, considering three models for these structured data: the classical models of complete pooling and no pooling, and a Bayesian multilevel model. The multilevel model captures important interactions that are not included in the complete pooling model, while at the same time avoiding the inevitable overfitting 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). Simulations based on real data show 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 prediction error for model selection with structured data.

Part III

Biased Survey

Chapter 3

Hierarchical Modeling of High-Frequency Non-Representative Polls

In this chapter, I will discuss an application of hierarchical modeling to non-representative survey sampling. As it is mentioned in the last chapter, there is a dichotomy in modern survey research, the camp of *describers* and the camp of *modelers*. However, at the heart of modern opinion polling, for both *describers* and *modelers*, is representative sampling, built around the goal that every individual in a particular target population (e.g., registered or likely U.S. voters) has the same probability of being sampled. Non-representative sampling has fallen out of favor among pollsters as a result of its inherent bias. I will show that, using an example of a highly-biased poll on US presidential election conducted on Xbox gaming platform, that hierarchical sampling can be used to remedy the bias and help extract useful information from non-representative polls. This chapter is based on a published paper [69], and a collaboration with David Rothschild, Sharad Goel and Andrew Gelman.

3.1 Representative vs Non-representative Sampling

The wide-scale adoption of representative polling can largely be traced to a pivotal polling mishap in the 1936 U.S. presidential election campaign. During that campaign, the popular magazine *Literary Digest* conducted a mail-in survey that attracted over two million responses, a huge sample even by modern standards. The magazine, however, incorrectly predicted a landslide victory for Republican candidate Alf Landon over the incumbent Franklin Roosevelt. Roosevelt, in fact, decisively won the election, carrying every state except for Maine and Vermont. As pollsters and academics have since pointed out, the magazine's pool of respondents was highly biased: it consisted mostly of auto and telephone owners as well as the magazine's own subscribers, which underrepresented Roosevelt's core constituencies [63]. During that same campaign, pioneering pollsters, including George Gallup, Archibald Crossley, and Elmo Roper, used considerably smaller but representative samples to predict the election outcome with reasonable [26]. Accordingly, non-representative or "convenience sampling" rapidly fell out of favor with polling experts. Methods used for sampling have evolved over time, from address-based, in-home interview sampling in the 1930s to random digit dialing after the growth of landlines and cellphones; nevertheless, leading polling organizations continue to put immense effort into obtaining representative samples.

Two recent trends spur the interest for non-representative polls. First, representative sampling is not nearly as representative as its name suggests, and it is becoming less so. Random digit dialing (RDD), the standard method in modern representative polling, has suffered increasingly high non-response rates, both due to the general public's growing reluctance to answer phone surveys, and expanding technical means to screen unsolicited calls [36]. By one measure, RDD response rates have decreased from 36% in 1997 to 9% in 2012 [37]. With such low response rates, even if the initial pool of targets is representative, those who ultimately answer the phone and elect to

respond are almost certainly not, calling into question the statistical benefits of such an approach. Related to dropping response rates is a corresponding increase in cost, in both time and money, as one needs to contact more and more potential respondents to find one willing to participate. The second trend driving this research is that with recent technological innovations, it is increasingly convenient and cost-effective to collect large numbers of highly non-representative samples via online surveys. What took several months for the *Literary Digest* editors to collect in 1936 can now take only a few days and can cost just pennies per response. The challenge, of course, is to extract meaningful signal from these unconventional samples.

It is worth noting that the so-called "Big Data" is more often than not a convenient sample, with potentially huge selection bias. Without adequately addressing this issue first, any conclusion drawn from big data analysis might be misleading.

3.2 Xbox Data

The analysis is based on an opt-in poll continuously available on the Xbox gaming platform during the 45 days preceding the 2012 U.S. presidential election. Each day, three to five questions were posted, one of which gauged voter intention with the standard query, "If the election were held today, who would you vote for?". Respondents were allowed to answer at most once per day. The first time they participated in an Xbox poll, respondents were additionally asked to provide basic demographic information about themselves, including their sex, race, age, education, state, party ID, political ideology, and for whom they voted in the 2008 presidential election. In total, 750,148 interviews were conducted with 345,858 unique respondents—over 30,000 of whom completed five or more polls—making this one of the largest ever election panel studies.

Despite the large sample size, the pool of Xbox respondents is far from representative of the voting population. Figure 3.1 compares the demographic composition

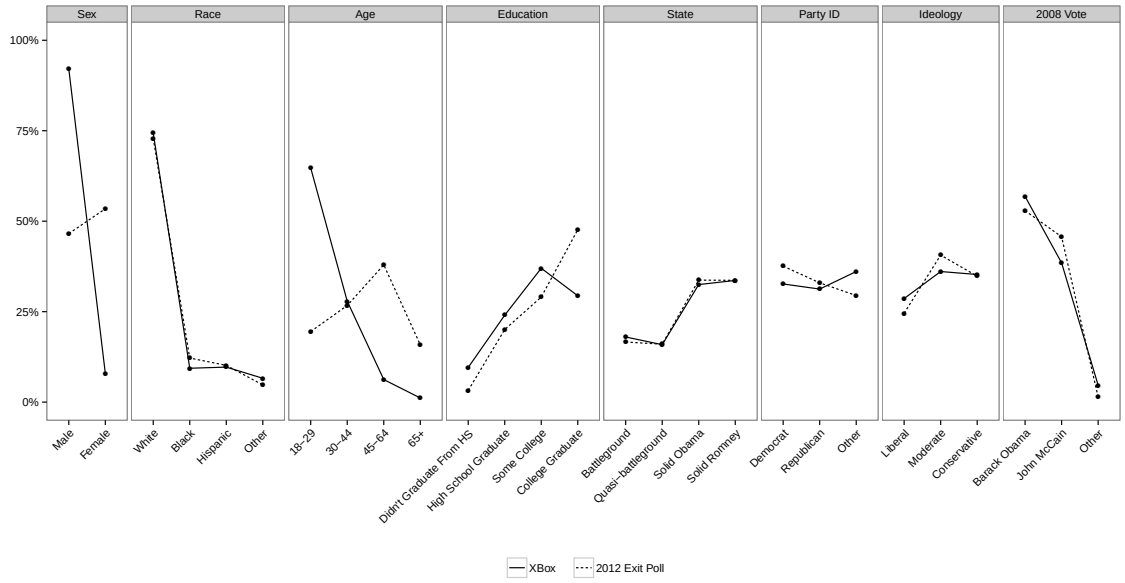

Figure 3.1: A comparison of the demographic, partisan, and 2008 vote distribution in the Xbox dataset and the 2012 electorate (as measured by adjusted exit polls). The sex and age distributions, as one might expect, exhibit considerable differences.

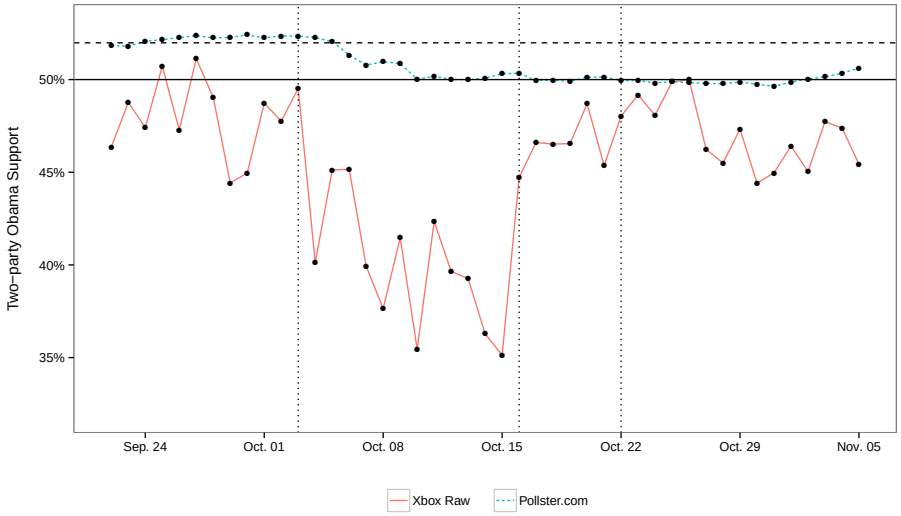

Figure 3.2: Daily (unadjusted) Xbox estimates of two-party Obama support during the 45 days leading up to the 2012 presidential election, which suggest a landslide victory for Mitt Romney. The dotted blue line indicates a consensus average of traditional polls (the daily aggregated polling results from Pollster.com), the horizontal dashed line at 52% indicates the actual two-party vote share obtained by Barack Obama, and the vertical dotted lines give the dates of the three presidential debates.

of the Xbox participants to that of the general electorate, as estimated via the 2012 national exit poll. For grouping states into different categories based on contestedness of the race, please refer to the Appendix. The most striking differences are for age and sex. As one might expect, young men dominate the Xbox population: 18-to-29-year-olds comprise 65% of the Xbox dataset, compared to 19% in the exit poll; and men make up 93% of the Xbox sample but only 47% of the electorate. Political scientists have long observed that both age and sex are strongly correlated with voting preferences (Kaufmann and Petrocik 1999), and indeed these discrepancies are apparent in the unadjusted time-series of Xbox voter intent shown in Figure 3.2. In contrast to estimates based on traditional, representative polls (indicated by the dotted blue line in Figure 3.2), the uncorrected Xbox sample suggests a landslide victory for Mitt Romney, reminiscent of the infamous *Literary Digest* error.

3.3 Estimating voter intent with multilevel regression and poststratification

3.3.1 Multilevel regression and poststratification

To transform the raw Xbox data into accurate estimates of voter intent in the general electorate, I make use of the rich demographic information that respondents provide. In particular I *poststratify* the raw Xbox responses to mimic a representative sample of likely voters. Poststratification is a popular method for correcting for known differences between sample and target populations [41]. The core idea is to partition the population into cells (e.g., based on combinations of various demographic attributes), use the sample to estimate the response variable within each cell, and finally to aggregate the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population. Using y to indicate the outcome of interest, the poststratification estimate is defined by,

$$\hat{y}^{\text{PS}} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

where $\hat{y}_j$ is the estimate of y in cell j, and N_j is the size of the j-th cell in the population. An estimate of y can be analogously derived at any subpopulation level s (e.g., voter intent in a particular state) by

$$\hat{y}_s^{\text{PS}} = \frac{\sum_{j \in J_s} N_j \hat{y}_j}{\sum_{j \in J_s} N_j}$$

where J_s is the set of all cells that comprise s. As is readily apparent from the form of the poststratification estimator, the key is to obtain accurate cell-level estimates, as well as estimates for the cell sizes.

One of the most common ways to generate cell-level estimates is to simply average sample responses within each cell. If within a cell the sample is drawn at random from the larger population, this yields an unbiased estimate. However, this assumption of cell-level simple random sampling is only reasonable when the partition is sufficiently fine; on the other hand, as the partition becomes finer, the cells become sparse, and the empirical sample averages become unstable. I address these issues by instead generating cell-level estimates via a regularized regression model, namely multilevel regression.

This combined model-based poststratification strategy, known as multilevel regression and poststratification (MRP), has been used to obtain accurate small-area subgroup estimates, such as for public opinion and voter turnout in individual states and demographic subgroups [23, 40, 24].

More formally, applying MRP in this setting comprises two steps. First a Bayesian hierarchical model is fit to obtain estimates for sparse poststratification cells; second, one averages over the cells, weighting by a measure of forecasted voter turnout, to get state and national-level estimates. Specifically, I generate the cells by considering all possible combinations of sex (2 categories), race (4 categories), age (4 categories), education (4 categories), state (51 categories), party ID (3 categories), ideology (3

categories) and 2008 vote (3 categories), which partition the data into 176,256 cells. {All demographic variables are collected prior to respondents' first poll, alleviating concerns that respondents may adjust their demographic responses to be inline with their voter intention (e.g., a new Obama supporter switching his or her party ID from Republican to Democrat). I fit two, nested multilevel logistic regressions to estimate candidate support in each cell. The first of the two models predicts whether a respondent supports a major-party candidate (i.e., Obama or Romney), and the second predicts support for Obama given that the respondent supports a major-party candidate. Following the notation of [22], the first model is given by

$$\begin{aligned}\Pr(Y_i \in \{\text{Obama, Romney}\}) = \\ \text{logit}^{-1}\big(&\alpha_0 + \alpha_1(\text{state last vote share}) \\ &+ a^{\text{state}}_{j[i]} + a^{\text{edu}}_{j[i]} + a^{\text{sex}}_{j[i]} + a^{\text{age}}_{j[i]} + a^{\text{race}}_{j[i]} + a^{\text{party ID}}_{j[i]} + b^{\text{ideology}}_{j[i]} + b^{\text{last vote}}_{j[i]}\big)\end{aligned} \tag{3.1}$$

where α_0 is the fixed baseline intercept, and α_1 is the fixed slope for Obama's fraction of two-party vote share in the respondent's state in the last presidential election. The terms $a^{\text{state}}_{j[i]}$, $a^{\text{edu}}_{j[i]}$, $a^{\text{sex}}_{j[i]}$ and so on—which in general is denote by $a^{\text{var}}_{j[i]}$—correspond to varying coefficients associated with each categorical variable. Here the subscript $j[i]$ indicates the cell to which the i-th respondent belongs. For example, $a^{\text{age}}_{j[i]}$ takes values from $\{a^{\text{age}}_{18-29}, a^{\text{age}}_{30-44}, a^{\text{age}}_{45-64}, a^{\text{age}}_{65+}\}$ depending on the cell membership of the i-th respondent. The varying coefficients $a^{\text{var}}_{j[i]}$ are given independent prior distributions

$$a^{\text{var}}_{j[i]} \sim N(0, \sigma^2_{\text{var}}).$$

To complete the full Bayesian specification, the variance parameters are assigned a hyperprior distribution

$$\sigma^2_{\text{var}} \sim \text{inv-}\chi^2(\nu, \sigma^2_0),$$

with a weak prior specification for the remaining parameters, ν and σ_0. The benefit of using a multilevel model is that estimates for relatively sparse cells can be improved through "borrowing strength" from demographically similar cells that have richer data. Similarly, the second model is defined by

$$\begin{aligned}\Pr(Y_i = \text{Obama} \,|\, Y_i \in \{\text{Obama, Romney}\}) = \\ \text{logit}^{-1}\big(\beta_0 + \beta_1(\text{state last vote share}) \\ + b^{\text{state}}_{j[i]} + b^{\text{edu}}_{j[i]} + b^{\text{sex}}_{j[i]} + b^{\text{age}}_{j[i]} + b^{\text{race}}_{j[i]} + b^{\text{party ID}}_{j[i]} + b^{\text{ideology}}_{j[i]} + b^{\text{last vote}}_{j[i]}\big)\end{aligned} \tag{3.2}$$

and

$$\begin{aligned} b^{\text{var}}_{j[i]} &\sim N(0, \eta^2_{\text{var}}), \\ \eta^2_{\text{var}} &\sim \text{inv-}\chi^2(\mu, \eta^2_0). \end{aligned}$$

Jointly, Eqs. (3.1) and (3.2) define a Bayesian model that describes the data. Ideally, a fully Bayesian analysis would be performed to obtain the posterior distribution of the parameters. However, for computational convenience, I use the approximate marginal maximum likelihood estimates obtained from the `glmer()` function in the `R` package `lme4` [5].

Having detailed the multilevel regression step, I now turn to poststratification, where cell-level estimates are weighted by the proportion of the electorate in each cell and aggregated to the appropriate level (e.g., state or national). To compute cell weights, cross-tabulated population data is needed. One commonly used source for such data is the Current Population Survey (CPS); however, the CPS does not includes some key poststratification variables, such as party identification. I thus instead use exit poll data from the 2008 presidential election. Exit polls are conducted on election day outside voting stations to record the choices of exiting voters, and they are generally used by researchers and news media to analyze the demographic

breakdown of the vote (after a post-election adjustment that aligns the weighted responses to the reported state-by-state election results). In total, 101,638 respondents were surveyed in the state and national exit polls. I use the exit poll from 2008, not 2012, because this means that in theory the method as described here could have been used to generate real-time predictions during the 2012 election campaign. Admittedly, this approach puts my prediction at a disadvantage since the demographic shifts of the intervening four years cannot be captured. While combining exit poll and CPS data would arguably yield improved results, for simplicity and transparency I exclusively use the 2008 exit poll summaries for poststratification.

3.3.2 National and State Voter Intent

Figure 3.3 shows the adjusted two-party Obama support for the last 45 days of the election. %The daily voter intents for two-party Obama support at the national level are %illustrated in Figure 3.3. Compared with the uncorrected estimates in Figure 3.2, the MRP-adjusted estimates yield a much more reasonable timeline of Obama's standing over the course of the final weeks of the campaign. With a clear advantage at the beginning, Obama's support slipped rapidly after the first presidential debate—though never falling below 50%—and gradually recovered, building up a decisive lead in the final days.

On the day before the election, the estimate of voter intent is off by a mere 0.6 percentage points from the actual outcome (indicated by the dotted horizontal line). Voter intent in the weeks prior to the election does not directly equate to an estimate of vote share on election day—a point I return to later. As such, it is difficult to evaluate the accuracy of the full time-series of estimates. Nonetheless, note that the estimates are not only intuitively reasonable, but that they are also inline with prevailing estimates based on traditional, representative polls. In particular, the estimates roughly track—and are even arguably better than—those from Pollster.com, one of the leading poll aggregators during the 2012 campaign.

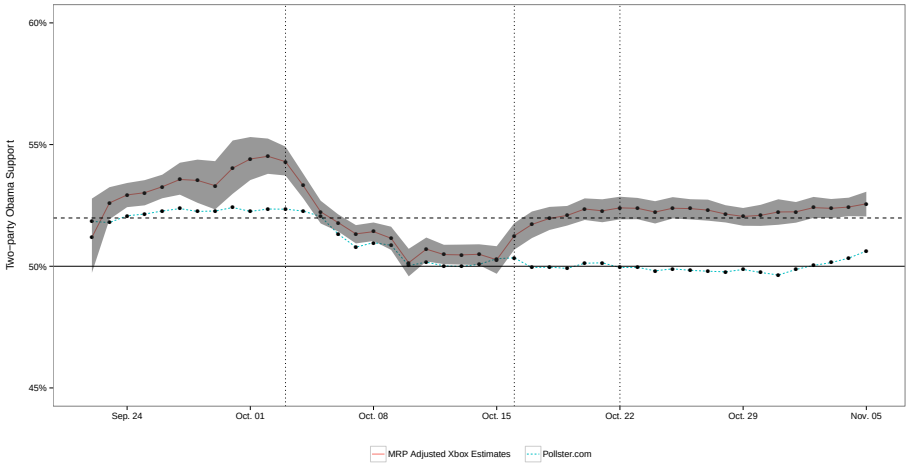

Figure 3.3: National MRP-adjusted voter intent of two-party Obama support over the 45-day period and the associated 95% confidence bands. The horizontal dashed line indicates the actual two-party Obama vote share. The three vertical dotted lines indicate the presidential debates. Compared with the raw responses in Figure 3.2, the MRP-adjusted voter intent is much more reasonable, and voter intent in the last few days is very close to the actual outcome. For comparison, the daily aggregated polling results from Pollster.com, shown as the blue dotted line, are further away from the actual vote share than the estimates generated from the Xbox data in the last few days.

National vote share receives considerable media attention, but state-level estimates are particularly relevant for many stakeholders given the role of the Electoral College in selecting the winner [54]. forecast the joint probability of victory for each candidate in Forecasting state-by-state races is a challenging problem due to the interdependencies in state outcomes, %and the joint electoral votes has not yet become the standard forecast the logistical difficulties of measuring state-level vote preference, and the effort required to combine information from various sources [42]. The MRP framework, however, provides a straightforward methodology for generating state-level results. Namely, I use the same cell-level estimates employed in the national estimate, as generated via the multilevel model in Eqs. (3.1) and (3.2), and I then poststratify to each state's demographic composition. In this manner, the Xbox responses can be used to construct estimates of voter intent over the last 45 days of the campaign for all 51 Electoral College races.

Figure 3.4 shows two-party Obama support for the 12 states with the most electoral votes. The state timelines share similar trends (e.g., support for Obama dropping after the first debate), but also have their own idiosyncratic movements, an indication of a reasonable blend of national and state-level signals. To demonstrate the accuracy of the MRP-adjusted estimates, I plot, in dotted blue lines in Figure 3.4, the estimates generated by Pollster.com, which are broadly consistent with the state-level MRP estimates. Moreover, across the 51 Electoral College races, the mean and median absolute errors of the estimates on the day before the election are just 2.5 and 1.8 percentage points, respectively.

3.3.3 Voter intent for demographic subgroups

Apart from Electoral College races, election forecasting often focuses on candidate preference among demographic subpopulations. Such forecasts are of significant importance in modern political campaigns, which often employ targeted campaign strategies [31]. In the highly non-representative Xbox survey, certain subpopulations

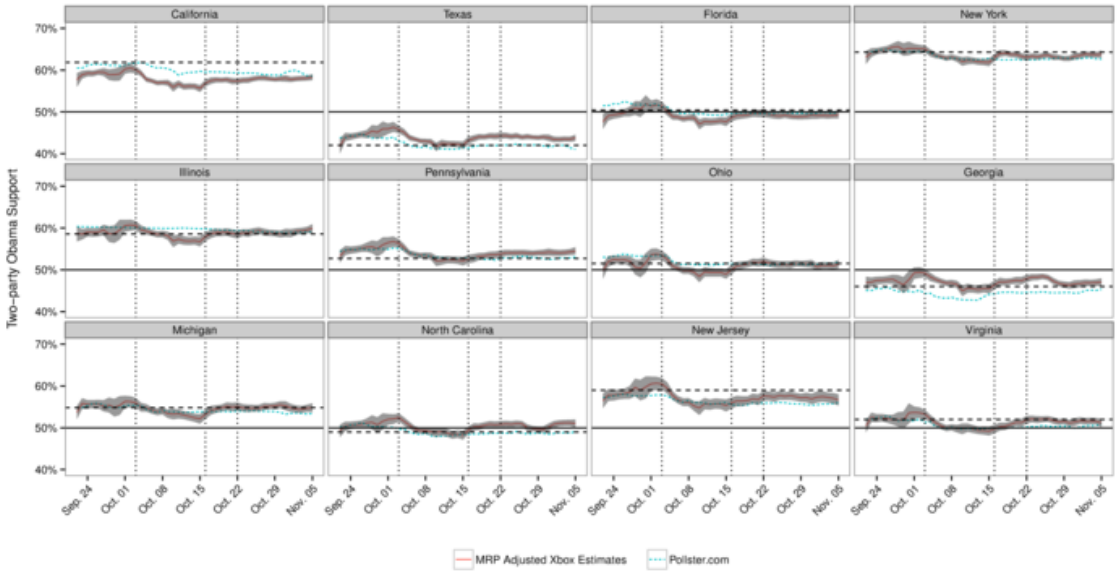

Figure 3.4: MRP-adjusted daily voter intent for the 12 states with the most electoral votes, and the associated 95% confidence bands. The horizontal dashed lines in each panel give the actual two-party Obama vote shares in that state. The mean and median absolute errors of the last day voter intent across the 51 Electoral College races are 2.5 and 1.8 percentage points, respectively. The state-by-state daily aggregated polling results from Pollster.com, given in the dotted blue lines, are broadly consistent with the estimates from the Xbox data.

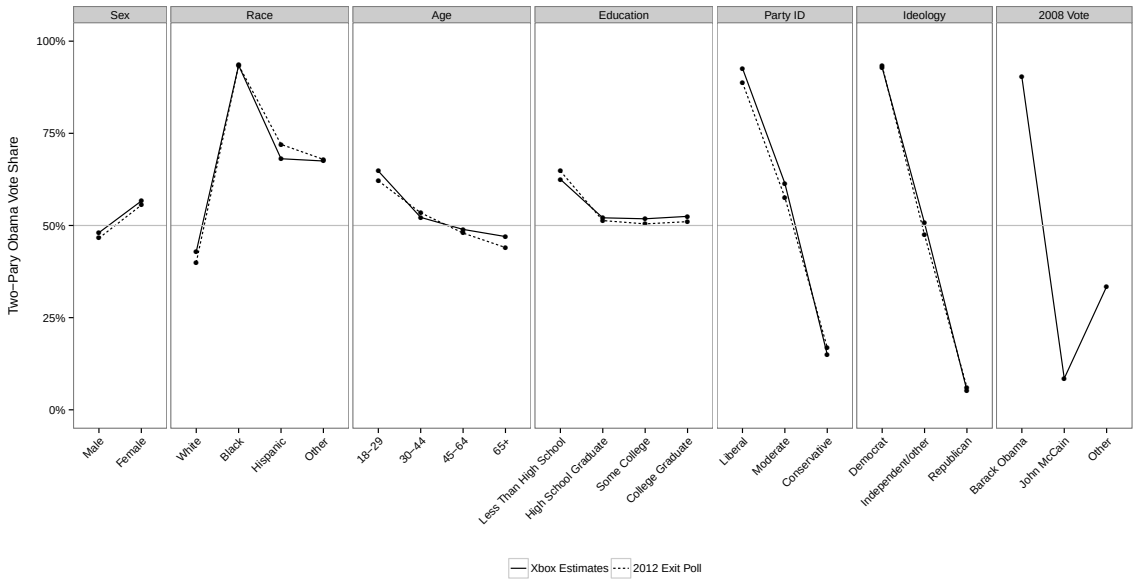

Figure 3.5: Comparison of two-party Obama vote share for various demographic subgroups, as estimated from the 2012 national exit poll and from the Xbox data on the day before the election.

are heavily underrepresented and plausibly suffer from strong self-selection problems. This begs the question, how accurate the estimates for older women based on a platform that caters to mostly young men?

It is straightforward in MRP to estimate voter intent among any collection of demographic cells: I again use the same cell-level estimates as in the national and state settings, but poststratify to the desired target population. For example, to estimate voter intent among women, the poststratification weights are based on the relative number of women in each demographic cell. To illustrate this approach, I compute Xbox estimates of Obama support for each level of the categorical variables (e.g., males, females, white, black, etc.) on the day before the election, and compare those with the actual voting behavior of those same groups as estimated by the 2012 national exit poll. As seen in Figure 3.5, the Xbox estimates are remarkably accurate, with a median absolute difference of 1.5 percentage points between the Xbox and the

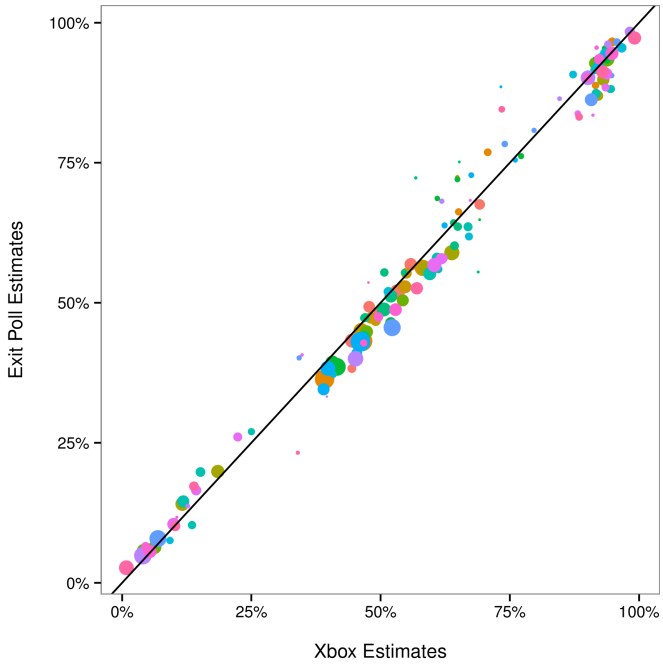

Figure 3.6: Two-party Obama support as estimated from the 2012 national exit poll and from the Xbox data on the day before the election, for various two-way interaction demographic subgroups (e.g., 65+ year-old women). The sizes of the dots are proportional to the population sizes of the corresponding subgroups. Subgroups within the same two-way interaction category (e.g., age by sex) have the same color.

exit poll numbers. Note that Respondents' 2008 vote was not asked on the 2012 exit poll, so I exclude that comparison from Figure 3.5.

Not only do the Xbox data facilitate accurate estimation of voter intent across these single-dimensional demographic categories, but they also do surprisingly well at estimating two-way interactions (e.g., candidate support among 18–29 year-old Hispanics, and liberal college graduates). Figure 3.6 shows this result, plotting the Xbox estimates against those derived from the exit polling data for each of the 149 two-dimensional demographic subgroups. Note that state contestedness is excluded from the two-way interaction groups since the 2012 state exit polls are not yet available, and the 2012 national exit poll does not have enough data to reliably estimate state interactions; 2008 vote is also excluded, as it was not asked in the 2012 exit poll. The "other" race category was also dropped as it was not consistently defined across the Xbox and exit poll datasets. Most points lie close to the diagonal, indicating that the Xbox and exit poll estimates are in agreement. Specifically, for women who are 65 and older—a group whose preferences one might a priori believe are hard to estimate from the Xbox data—the difference between Xbox and the exit poll is a mere one percentage point (49.5% and 48.5%, respectively). Across all the two-way interaction groups, the median absolute difference is just 2.4 percentage points. As indicated by the size of the points in Figure 3.6, the largest differences occur for relatively small demographic subgroups (e.g., liberal Republicans), for which both the Xbox and exit poll estimates are less reliable. For the 30 largest demographic subgroups, Figure 3.7 lists the differences between Xbox and exit poll estimates. Among these largest subgroups, the median absolute difference drops to just 1.9 percentage points.

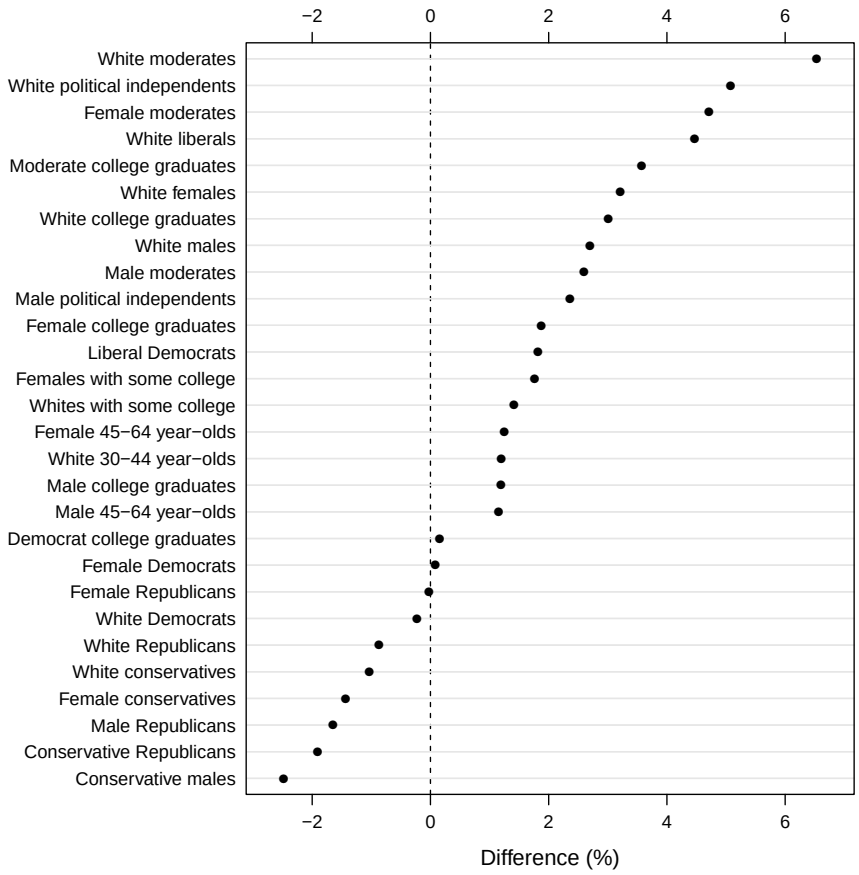

Figure 3.7: Differences between the Xbox MRP-adjusted estimates and the exit poll estimates for the 30 largest two-dimensional demographic subgroups, ordered by the difference. Positive values indicate the Xbox estimate is larger than the corresponding exit poll estimate. Among these 30 subgroups, the median and mean absolute differences are 1.9 and 2.2 percentage points, respectively.

3.4 Forecasting Election Day Outcome

3.4.1 Converting Voter Intent to Forecasts

As mentioned above, daily estimates of voter intent do not directly correspond to estimates of vote share on election day. There are two key factors for this deviation. First, opinion polls (both representative and non-representative ones) only gauge voter preference on the particular day when the poll is conducted, with the question typically phrased as, "if the election were held today." Political scientists and pollsters have long observed that such stated preferences are prone to several biases, including the anti-incumbency bias, in which the incumbent's polling numbers tend to be lower than the ultimate outcome [10], and the fading early lead bias, in which a big lead early in the campaign tends to diminish as the election gets closer [19]. Moreover, voters' attitudes are affected by information revealed over the course of the campaign, so preferences weeks or months before election day are at best a noisy indicator of one's eventual vote. Second, estimates of vote share require a model of likely voters. That is, opinion polls measure preferences among a hypothetical voter pool, and are thus accurate only to the extent that this pool captures those who actually turn out to vote on election day. Both of these factors introduce significant complications in forecasting election day outcomes.

To convert daily estimates of voter intent to election day predictions—which I hereafter refer to as the voter intent—I compare daily voter intent in previous elections to the ultimate outcomes in those elections. Specifically, I collected historical data from three previous U.S. presidential elections, in 2000, 2004, and 2008. For each year, I obtained top-line (i.e., not individual-level) national and state estimates of voter intent from all available polls conducted in those elections. The polling data are obtained from Pollster.com and RealClearPolitics.com. From this collection of polling data, I then constructed daily estimates of voter intent by taking a moving

average of the poll numbers, in a similar manner to the major poll aggregators. Note that I rely on traditional, representative polls to reconstruct historical voter intent; in principle, however, I could have started with non-representative polls if such data were available in previous election cycles.

I next infer a mapping from voter intent to election outcomes by regressing election day vote share on the historical time-series of voter intent. The key difference between the approach in this chapter and previous related work [19, 53] is that I explicitly model state-level correlations, via nested national and state models and correlated error terms. Specifically, I first fit a national model given by

$$y_e^{\mathrm{US}} = a_0 + a_1 x_{t,e}^{\mathrm{US}} + a_2 |x_{t,e}^{\mathrm{US}}| x_{t,e}^{\mathrm{US}} + a_3 t x_{t,e}^{\mathrm{US}} + \eta(t,e)$$

where y_e^{US} is the national election day vote share of the incumbent party candidate in election year e, $x_{t,e}^{\mathrm{US}}$ is the national voter intent of the incumbent party candidate at t days before the election in year e, and $\eta \sim N(0, \sigma^2)$ is the error term. Both y_e^{US} and $x_{t,e}^{\mathrm{US}}$ are offset by 0.5, so the values run from \$-\$0.5 to 0.5 rather than 0 to 1. The term involving the absolute value of voter intent pulls the vote share prediction toward 50%, capturing the diminishing early lead effect. I do not include a main effect for time since it seems unlikely that the number of days until the election itself contributes to the final vote share directly, but rather time contributes through its interaction with the voter intent (which it is include in the model).

Similarly, the state model is given by

$$y_{s,e}^{\mathrm{ST}} = b_0 + b_1 x_{s,t,e}^{\mathrm{ST}} + b_2 |x_{s,t,e}^{\mathrm{ST}}| x_{s,t,e}^{\mathrm{ST}} + b_3 t x_{s,t,e}^{\mathrm{ST}} + \varepsilon(s,t,e)$$

where $y_{s,e}^{\mathrm{ST}}$ is the election day state vote share of the state's incumbent party candidate at day t, $x_{s,t,e}^{\mathrm{ST}}$ is the state voter intent at day t, and ϵ is the error term. The outcome $y_{s,e}^{\mathrm{ST}}$ is offset by the national projected vote share on that day as fit with the national calibration model, and $x_{s,t,e}^{\mathrm{ST}}$ is offset by that day's national voter

intent. Furthermore, I impose two restrictions on the magnitude and correlation structure of the error term $\varepsilon(s,t,e)$. First, since the uncertainty naturally decreases as the election gets closer (as t becomes smaller), I apply the heteroscedastic structure $\text{Var}(\varepsilon(s,t,e)) = (t+a)^2$, where a is a constant to be estimated from the data. Second, the state-specific movements within each election year are allowed to be correlated. For simplicity, and as in [11], I assume these correlations are uniform (i.e., all pairwise correlations are the same), which creates one more parameter to be estimated from the data. I fit the full calibration model with the `gls()` function in the `R` package `nlme` [49].

In summary, the procedure for generating election day forecasts proceeds in three steps:

1. Estimate the joint distribution of state and national voter intent by applying MRP to the Xbox data.

2. Fit the nested calibration model described above on historical data to obtain point estimates for the parameters, including estimates for the error terms.

3. Convert the distribution of voter intent to election day forecasts via the fitted calibration model.

3.4.2 National and state election day forecasts

Figure 3.8 plots the projected vote shares and pointwise 95% confidence bands over time for the 12 states with the most electoral votes. Though these time-series look quite reasonable, it is difficult to assess their accuracy as there are no ground truth estimates to compare with in the weeks prior to the election. As a starting point, I compare the state-level estimates to those generated by prediction markets, which are widely considered to be among the most accurate sources for political predictions [54, 73]. For each state, prediction markets produce daily probabilities of victory.

Though Figure 3.8 plots the forecasts in terms of expected vote share, this estimation procedure in fact yields the full distribution of outcomes, and so I can likewise convert my estimates to probabilistic forecasts. Figure 3.9 shows this comparison, where the prediction market estimate is derived by averaging the two largest election markets, Betfair and Intrade. My probabilistic estimates are largely consistent with the prediction market probabilities. In fact, for races with little uncertainty (e.g., Texas and Massachusetts), the Xbox estimates do not seem to suffer from the long-shot bias common to prediction markets [53], and instead yield probabilities closer to 0 or 1. For tighter races, the Xbox estimates—although still highly correlated with the prediction market probabilities—look more volatile, especially in the early part of the 45-day period. Since the ground truth is not clearly defined, it is difficult to evaluate which method—Xbox or prediction markets—yields better results. From a Bayesian perspective, if one believes the stability shown by prediction markets, this could be incorporated into the structure of the Xbox calibration model.

With the full state-level outcome distribution, I can also estimate the distribution of Electoral College votes. Figure 3.10 plots the median projected electoral votes for Obama over the last 45-days of the election, together with the 95% confidence band. In particular, on the day before the election, my model estimates Obama had an 88% chance of victory, in line with estimates based on traditional polling data. For example, Simon Jackman predicted Obama had a 91% chance of victory, using a method built from [34]. Zooming in on the day before the election, Figure 3.11 shows the full predicted distribution of electoral votes for Obama. Compared to the actual 332 votes that Obama captured, I estimate a median of 312 votes, with the most likely outcome being 303. Though this distribution of Electoral College outcomes seems reasonable, it does appear to have higher variance than one might expect. In particular, the extreme outcomes seem to have unrealistically high likelihood of occurring, which is likely a byproduct of the calibration model not fully capturing the state-level correlation structure. Nonetheless, given that my forecasts are based

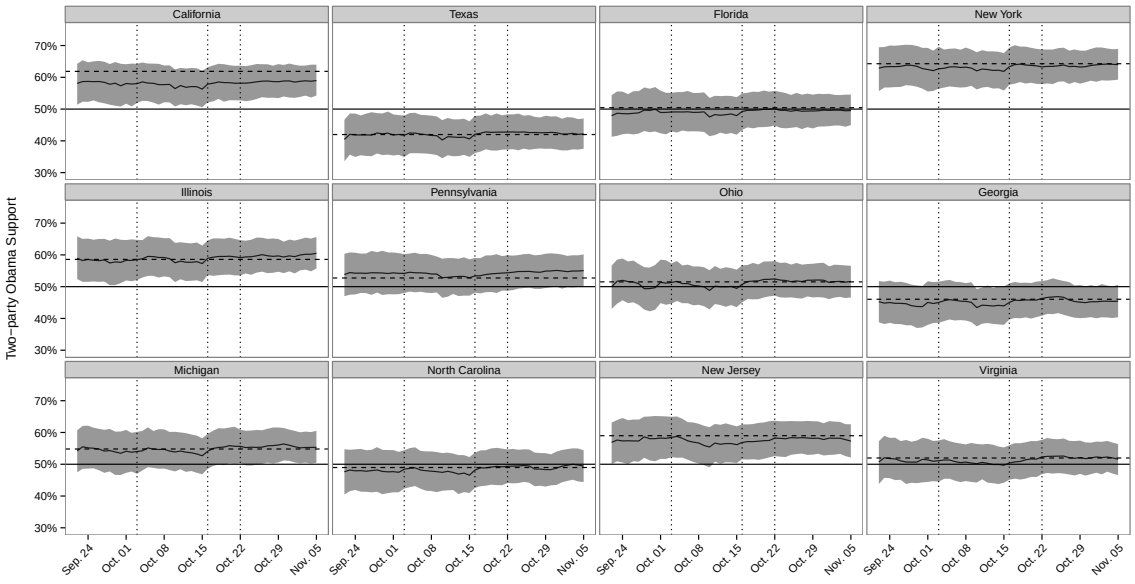

Figure 3.8: Projected Obama share of the two-party vote on election day for each of the 12 states with the most electoral votes, and associated 95% confidence bands. Compared to the MRP-adjusted voter intent in Figure 3.4, the projected two-party Obama support is more stable, and the North Carolina race switches direction after applying the calibration model. Additionally, the confidence bands become much wider and give more reasonable state-by-state probabilities of Obama victories.

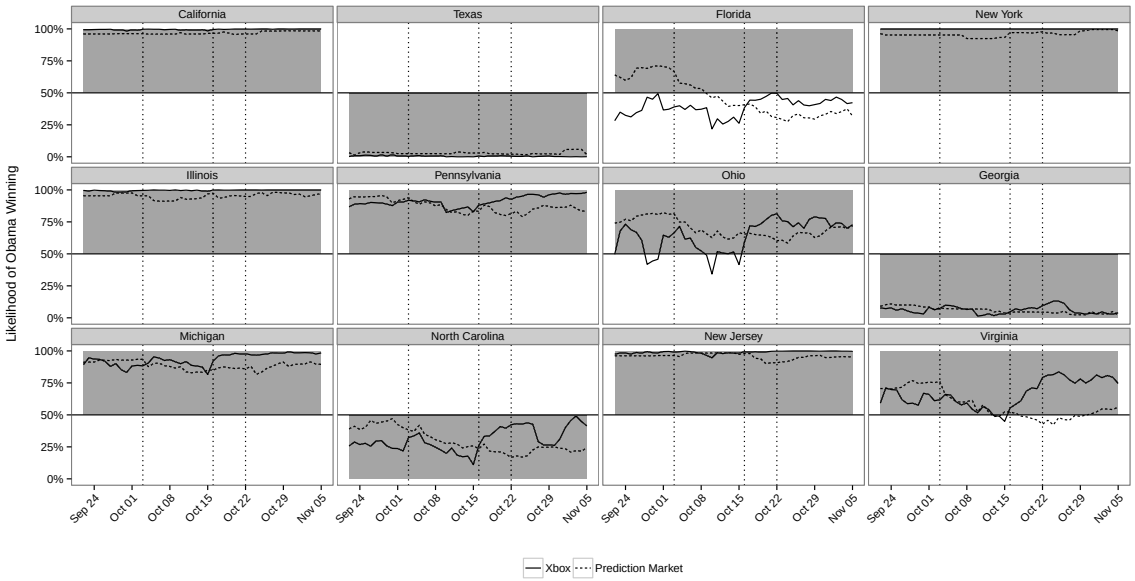

Figure 3.9: Comparison between the probability of Obama winning the 12 largest Electoral College races based on Xbox data and on prediction market data. The prediction market data are the average of the raw Betfair and Intrade prices from winner-take-all markets. The three vertical lines represent the dates of three presidential debates. The shaded halves indicate the direction that race went.

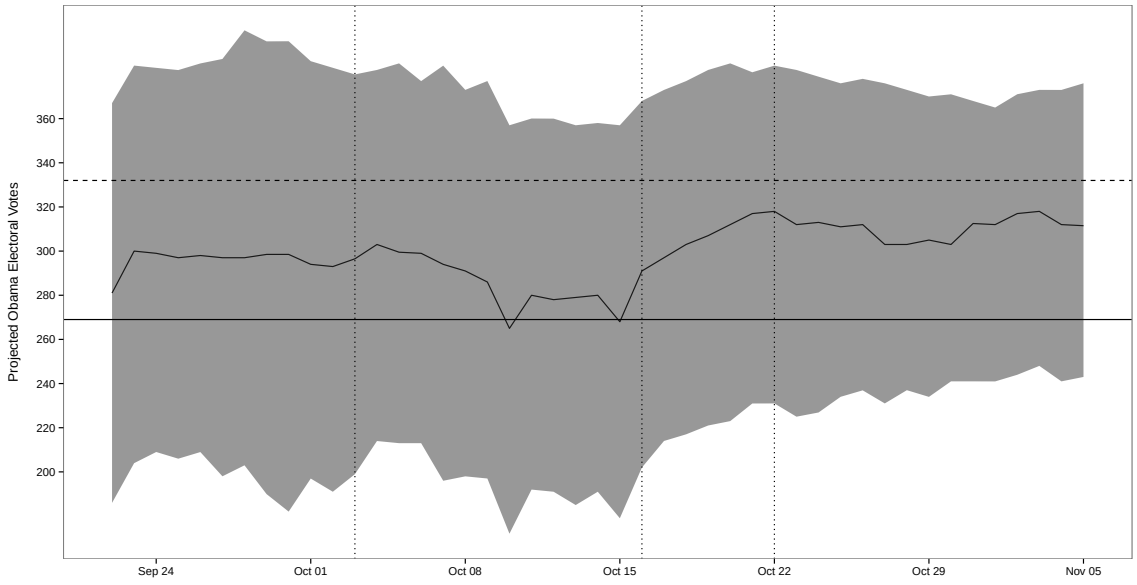

Figure 3.10: Daily projections of Obama electoral votes in the 45-day period leading up to the 2012 election and associated 95% confidence bands. The solid line represents the median of the daily distribution. The horizontal dashed line represents the actual electoral votes, 332, that Obama captured in 2012 election. Three vertical dotted lines indicate the dates of three presidential debates.

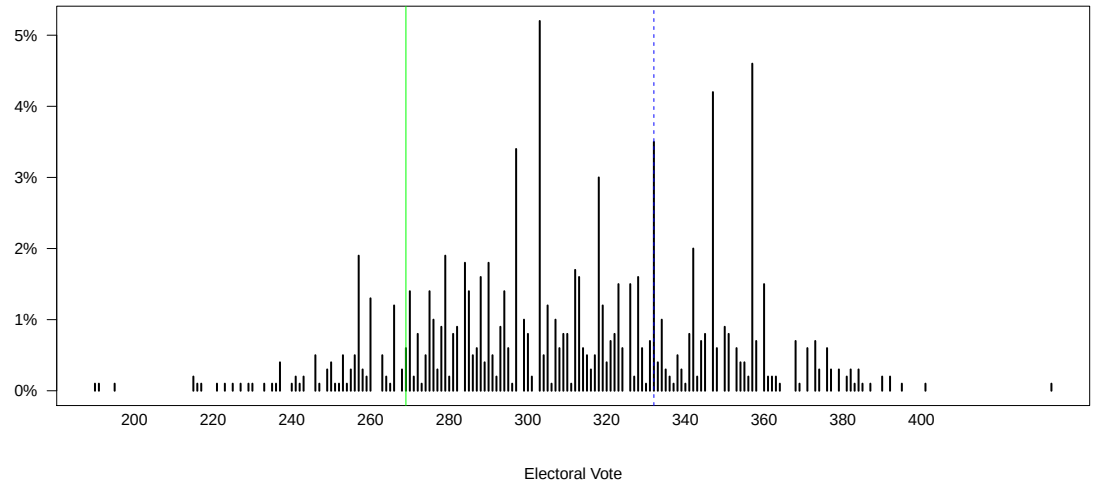

Figure 3.11: Projected distribution of electoral votes for Obama one day before the election. The green vertical dotted line represents 269, the minimum number of electoral votes that Obama needs for a tie. The blue vertical dashed line gives 332, the actual number of electoral votes captured by Obama. The estimated likelihood of Obama winning the electoral vote is 88%.

on a highly biased convenience sample of respondents, the model predictions are remarkably good.

3.5 Conclusion

Forecasts not only need to be accurate, but also relevant, timely, and cost-effective. In this chapter, I construct election forecasts satisfying all of these requirements using extremely non-representative data. Though the data were collected on a proprietary polling platform, in principle one can aggregate such non-representative samples at a fraction of the cost of conventional survey designs. Moreover, the data produce forecasts that are both relevant and timely, as they can be updated faster and more regularly than standard election polls. Thus, the key question—and one of the main contributions of this chapter—is to assess the extent to which one can generate accurate predictions from non-representative samples. Since there is limited ground truth for election forecasts, definitely establishing the accuracy of my predictions is difficult. Nevertheless, I show that the MRP-adjusted and calibrated Xbox estimates are both intuitively reasonably, and are also quite similar to those generated by more traditional means.

The greatest impact of non-representative polling will likely not be for presidential elections, but rather for smaller, local elections and specialized survey settings, where it is impractical to deploy traditional methods due to cost and time constraints. For example, non-representative polls could be used in Congressional elections, where there are currently only sparse polling data. Non-representative polls could also supplement traditional surveys (e.g., the General Social Survey) by offering preliminary results at shorter intervals. General Social Survey, which is . Finally, when there is a need to identify and track pivotal events that affect public opinion, non-representative polling offers the possibility of cost-effective continuous data collection. Standard representative polling will certainly continue to be an invaluable tool for the foreseeable

future. However, 75 years after the *Literary Digest* failure, non-representative polling (followed by appropriate post-data adjustment) is due for further exploration, for election forecasting and in social research more generally.

Part IV

Meta Analysis

Chapter 4

Causal Inference for Multilevel Data with Interference via Gaussian Processes

In this chapter, I discuss causal inference for multilevel data. Multilevel data, as illustrated in the previous chapters, are a mainstay in social, behavioral and medical science. While some of the multilevel data analysis are descriptive, such as the survey examples used in the previous 2 chapters, a large number of multilevel datasets call for causal analysis in nature. Two of the most interesting examples, as will be discussed in this chapter, are meta-analysis and multi-site randomized experiment. Based on the nature of the available data, multilevel causal inference can be dichotomized into two types. In the first type, researchers only have the information of the studies or the sites, rather than the information about individual participants. This corresponds to, for example, the traditional meta-analysis where effect sizes and study characteristics are extracted from published papers. Here, the information available to account for systematic sources of heterogeneity in treatment effects across studies is generally inadequate, and researchers will often estimate a common treatment effect, and possibly the variation in this effects across studies [17, 51]. In the second type,

researchers have access to the individual-participant data. This case is more similar to the multilevel survey data discussed in the previous chapters. Compared with study/site-level multilevel data, Individual participant data (IPD), offer numerous advantages [60, 15], and are becoming increasingly available and common.

The most naïve strategy to handle multilevel dataset is to pool all data together, ignoring the group structure. But average treatment effects frequently vary across studies/sites. Subjects in different studies/sites are often drawn from different populations. For the sake of brevity, I will use study in place of study/site for the remainder of this chapter, but it should be noted that the discussion herein apply to general multilevel data of which meta-analysis is a special case. A viable approach to account for between study heterogeneity in treatment effects is to use covariates, of both individual and study level, together with random effects models [1, 29, 28, 65]. Unfortunately, researchers often overlook sources of between study heterogeneity and use random effect models simply as a tool to fuse a one-number summary of otherwise disparate effects. More careful consideration of sources of heterogeneity is needed to inform analyses and resulting policy recommendations. For example, in a multi-school educational intervention, understanding which demographic groups benefit the most from the intervention, as well as whether certain schools are more receptive to the intervention, might be of primary interest, rather than an overall summary of intervention effects.

To that end, [62] developed an extended potential outcome framework to put multilevel data on a solid causal foundation. Motivated by meta-analyses, the framework explicitly codifies the sources and nature of between study heterogeneity. When extensive information about the subjects is collected, as in IPD meta-analyses, the framework can be used to test, under specified conditions, whether or not particular sources contribute to between study heterogeneity. In meta-analyses based on published data, subjects cannot be linked to covariates that vary within studies. If these characteristics differentiate outcomes and their distribution varies across studies, it

will not be possible to reliably test hypotheses about these sources of variation in general [48]. Nevertheless, the framework may still be used to think more carefully about the sources of heterogeneity and how one might want to conduct and interpret empirical analyses. [62] illustrated the framework with an IPD meta-analysis of the Vioxx clinical trials.

In this chapter I discuss two innovations. One is the extension of the causal framework outlined in [62], specifically in handling interference/peer/neighborhood effect. The other is the use of non-parametric models that explicitly handles heterogeneity across studies based on Gaussian Processes (GP). GP allows for flexible modeling of response functions and admits full probabilistic inference [71]. The second is the relaxation of the so-called Stable Unit Treatment Value Assumption (SUTVA) for multilevel data, incorporating peer influences common in educational settings into the potential outcomes [33] . To illustrate, I reanalyze the dataset from Tennessee Student-Teacher Achievement Ratio study (hereafter Project STAR). Project STAR was a large-scale randomized experiment on the effect of class size on educational outcomes funded by Tennessee legislature. Project STAR was carried out in 79 schools across Tennessee, spanning from 1985 to 1989 and involving about 10,000 students. Upon entry into participating schools in kindergarten through grade 3, each student was randomly assigned to one of three class types, and end-of-year test scores were recorded as outcomes. To date, Project STAR is still the most widely studied educational experiment, and researchers are still perusing its rich data set for insights. [38] pointed on the heterogeneous effects of attending a small class that vary with demographics, however, the existing literature mostly ignore the multi-site structure of the data. Applying GP models with group structure on the Project STAR dataset yields some insights that are not obvious from traditional studies.

This chapter proceeds as follows. I first review the extended potential outcome framework outlined in [62]. Next, I review the basic model setup and inference of Gaussian Processes (GP) model, and discuss how to incorporate multilevel structure

into GP models. Then I discuss Project STAR data, review the previous educational and economic literature on it , and discuss how to adjust the potential outcome framework to handle partial-interference inherent in cluster-randomized multi-site data. Lastly, I present the results from reanalyzing Project STAR data, and discuss insights gained through the potential outcome framework as well as the GP model.

4.1 A Potential Outcome Framework for Multilevel Data

[62] used the potential outcomes framework [44, 55] in which causal effects are defined as within-subject comparisons of outcomes under different treatments, only one of which is observed. The exposition in this section closely follows [62].

4.1.1 Potential Outcomes for a Single Study

The potential outcomes of an individual i under a treatment assignment $\boldsymbol{z}$ is defined as $Y_i(\boldsymbol{z})$. Note that $\boldsymbol{z}$ is a vector whose length is the sample size of the study, since different combination of treatment assignments might affect the potential outcome of individual i. To reduce the complexity, it is common to assume that the stable unit treatment value assumption (SUTVA) hold. SUTVA states that assignments of other unit don't affect the potential outcomes of the unit under considerations, i.e. $Y_i(\boldsymbol{z}) \equiv Y_i(z_i)$. (For cases where SUTVA doesn't hold see [61] and [33]) For a binary treatment $z \in \{0, 1\}$, there are only two potential outcomes for individual i, thus the unit causal effect of treatment z on individual i can be defined as $Y_i(1) - Y_i(0)$. Typically, researchers are interested in estimating quantities such as the population average treatment effect (PATE)

$$E(Y(1) - Y(0)),$$

or the conditional average treatment effect (CATE) corresponding to some covariates X

$$E(Y(1) - Y(0) \mid X = x)$$

An assumption in causal inference that many empirical analyses are based on is the unconfoundedness assumption [52], which states that given a set of observed covariates, treatment assignment Z is independent of the potential outcomes $(Y(0), Y(1))$:

$$Y(0), Y(1) \perp\!\!\!\perp Z \mid X$$

In the case of randomized experiment, the stronger assumption $Y(0), Y(1), X \perp\!\!\!\perp Z$ holds. Under the unconfoundedness assumption,

$$\begin{aligned} & E(Y|X, z = 1) - E(Y|X, z = 0) \\ =& E(Y(1)|X, z = 1) - E(Y(0)|X, z = 0) \\ =& E(Y(1)|X) - E(Y(0)|X) \\ =& E(Y(1) - Y(0)|X) \end{aligned}$$

and thus the causal effect can be identified from the observables.

4.1.2 Extended Potential Outcomes

Let $\mathbf{z} = (z_1, ..., z_n)$ and $\mathbf{s} = (s_1, ..., s_n)$ and let $Y_i(\mathbf{s}, \mathbf{z})$ denote the response subject i would have under the study allocation $\mathbf{s}$ and treatment assignment $\mathbf{z}$. In the most general case, potential outcomes are defined for all treatment by study combinations, the notation and results are easily modified to handle the case where potential outcomes are not defined for all treatment by study combinations, such as when study protocol makes it impossible for certain study-treatment combination.

As it stands, the potential outcomes of subject i may depend on what studies and treatments other subjects are assigned to. This often creates too much complexity and might be reasonably simplified. Following the concept of SUTVA in the single study case, an extended SUTVA for multilevel data can be defined:

Extended stable unit treatment value assumption (ESUTVA): For all possible assignments $\mathbf{z}$ *and allocations* $\mathbf{s}$, $Y_i(\mathbf{s}, \mathbf{z}) = Y_i(s_i, z_i) \equiv Y_i(s, z)$.

Extended SUTVA shares the same caveats as plain SUTVA, which is, when study participants interact, as in social networks, schools and neighborhoods, this assumption may require modification. I will discuss this later.

[62] formalized two sources of between study heterogeneity: the differences in responses of a given unit to the same treatment in different studies, and the assignment mechanism(s) by which treatments and studies are paired with subjects.

First, the notion that a subject's response to a given treatment is the same in all studies, which cannot be properly expressed without the extended potential outcome framework, is implicit in multilevel data analysis where a common treatment effect (or conditional effect) is assumed to hold across studies.

Response consistency assumption for treatment z*: For all* s, s' *and subjects* i, $Y_i(s, z) = Y_i(s', z)$.

[62] also discussed several relaxations of the response consistency assumption that are sufficient for identifying and estimating the effects.

Next is the formalization of the concept of treatment selection within studies and study selection mechanisms. If each study is randomized or has a good amount of covariates, treatment assignment is assumed to be independent of potential outcomes, given covariates $\mathbf{X}$:

Unconfounded treatment assignment within studies given observed covariates: for

every s, and treatment z, $Y(s,z) \perp\!\!\!\perp Z \mid S, \mathbf{X}$.

Unconfounded treatment assignment within studies assumption allows identification of the potential outcome distributions from the observed outcome distributions:

$$F(y \mid Z = z, S = s, \mathbf{X} = \mathbf{x}) = F(y(s,z) \mid Z = z, S = s, \mathbf{X} = \mathbf{x}). \tag{4.1}$$

This assumption is analogous to the ignorability assumption in [52] that is often critical for causal inference in the single study case. But often, different studies sampled from different population in the beginning. For example, schools often represent drastically different student populations, depending on the location and neighborhood. This extra layer of complexity induced by the multilevel structure is the focus here. Assuming that the observed covariates $\mathbf{X}$ may account for differential selection into studies, that is studies and potential outcomes are independent, given $\mathbf{X}$:

Unconfounded study selection, given observed covariates: For all studies s, s' and treatments z, $F(y(s,z) \mid S = s, \mathbf{X} = \mathbf{x}) = F(y(s,z) \mid S = s', \mathbf{X} = \mathbf{x})$.

Like the consistency assumptions, the notion of unconfounded study selection cannot be properly formalized without considering the outcomes subjects would have in studies other than those in which they actually participated.

The main assumptions of the extended potential outcome framework have been laid out. However, it is only informative when one consider them as a whole. Although neither the response consistency assumptions nor the study selection assumption are testable, if both these hold, for every study s in which treatment z is administered, the distributions of the response, conditional on $\mathbf{X}$, are identical:

$$F(y \mid Z = z, S = s, \mathbf{X} = \mathbf{x}) = F(y(s,z) \mid S = s, \mathbf{X} = \mathbf{x}) = F(y(s,z) \mid S = s', \mathbf{X} = \mathbf{x}) \tag{4.2}$$

$$= F(y(s',z) \mid S = s', \mathbf{X} = \mathbf{x}) = F(y \mid Z = z, S = s', \mathbf{X} = \mathbf{x}), \tag{4.3}$$

where the first equality follows from unconfounded treatment assignment within

study, the second from unconfounded study selection, the third from response consistency and the fourth from unconfounded treatment assignment within study again.

Thus, if the observations indicate (4.3) fails to hold, at least one of the two assumptions, unconfounded treatment assignments within studies and unconfounded study selection, fails to hold. The practical implication for a analyst, therefore, is first to assess whether one of the two assumptions could be supported in this particular situation, and if so, (4.3) serves as a test of whether the other assumption hold and thus leads to a better understanding of the source of the effect heterogeneity

4.2 Multilevel Causal Inference via GP

4.2.1 Non-parametric Modeling for Causal Inference

Traditionally, causal inference using potential outcomes focuses on two questions. Modeling of the treatment assignment process $p(z \mid x)$, also known as the propensity score, and modeling of the scientific process of how responses relate to treatment and covariates $E(y \mid z, x)$, also known as the response surface [56]. A myriad of methods based on the either treatment assignment mechanism (e.g., propensity score matching), or response surface modeling (e.g., regression), or combination of these two (e.g., the doubly-robust method), has been proposed for causal inference of observational data.

Recently, following the advances in Bayesian non-parametric models, [30] proposed a model that focuses on accurately estimating the response surface using flexible Bayesian Additive Regression Trees (BART) [13]. Besides the well-known benefits of being robust to model misspecifications and being able to capture highly non-linear and interaction patterns, Bayesian non-parametric models provide well-calibrated probabilistic intervals to convey inferential uncertainty.

4.2.2 Gaussian Processes

Gaussian Processes (GP) have become a popular tool for nonparametric regression. A random function f follows a GP process with kernel κ if any finite-dimensional marginal of f follows a Gaussian distribution, i.e.

$$f(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}}), \forall\ \boldsymbol{x} \in \mathbb{R}^d \text{ and } d$$

where $\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}}$ is the covariance matrix of kernel κ, i.e.,

$$\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}} = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_d) \\ \dots & \dots & \dots & \dots \\ k(x_d, x_1) & k(x_d, x_2) & \dots & k(x_d, x_d) \end{pmatrix}.$$

A large part of the popularity of GP stems from the fact a GP model can be interpreted as a generalization of linear regression with Gaussian errors, the predominant model for parametric regression. In fact, according to Mercer's Theorem [71], the kernel κ can be decomposed

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i^{\intercal}(x')$$

where λ_i and ϕ_i are the respective eigenvalues and eigenfunctions of kernel κ with respect to a measure μ, i.e.,

$$\int k(x, x') \phi_i(x)\, d\mu(x) = \lambda \phi_i(x'),$$

Then GP can be interpreted as a basis expansion method that maps input x to an infinite dimensional space via the infinite series of functions $\{\phi_i(x)\}_{i=1}^{\infty}$. For example, the square exponential kernel, defined as

$$\kappa(x_i, x_j) = \sigma^2 \exp\left(-\frac{1}{2l^2}(x_i - x_j)^2\right),$$

is equivalent to doing a Bayesian linear regression on an infinite amount of basis functions of the form

$$\{\phi_c(x) = \exp\left(-\frac{(x-c)^2}{2l^2}\right) \mid c \in \mathbb{R}\}$$

, with the prior on the coefficients being $N(0, \sigma^2)$ [71]. Since the eigenfunctions of the kernel encodes the structure of the basis expansion, judiciously choosing the kernel κ is the most important part of a GP model.

4.2.3 Inference for Standard GP

The standard GP model for N observation pairs $(y_i, \boldsymbol{x}_i)_{i=1}^N$ is

$$y_i \mid f \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2)$$
$$f \sim GP(0, k)$$

For a given kernel κ, the marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}(0, K_{\boldsymbol{x},\boldsymbol{x}} + \sigma^2 I_N)$$

where $K_{\boldsymbol{x},\boldsymbol{x}}$ is the Gram matrix of kernel κ whose entries are $k(x_i, x_j)$. With some algebra, the predictive distribution at new points $\boldsymbol{X}^\star$ can be derived as

$$\begin{aligned}\boldsymbol{y}^\star \mid \boldsymbol{X}^\star, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}(&K_{\boldsymbol{X}^\star,\boldsymbol{X}}(K_{\boldsymbol{X},\boldsymbol{X}} + \sigma^2 I_N)^{-1}\boldsymbol{y}, \\ &K_{\boldsymbol{X}^\star,\boldsymbol{X}^\star} - K_{\boldsymbol{X}^\star,\boldsymbol{X}}(K_{\boldsymbol{X},\boldsymbol{X}} + \sigma^2 I_N)^{-1}K_{\boldsymbol{X}^\star,\boldsymbol{X}}^\intercal)\end{aligned} \tag{4.4}$$

For inference on hyperparameters, e.g., parameters governing the kernels, a standard practice is to maximize the log marginal likelihood

$$\begin{aligned}\log p(\boldsymbol{y} \mid \boldsymbol{X}, \theta) &= \log \int p(\boldsymbol{y} \mid f, \boldsymbol{X}, \theta) p(f)\, df \\ &\propto -\left[\boldsymbol{y}^\intercal (K_{\boldsymbol{X},\boldsymbol{X}}(\theta) + \sigma^2 I_N)^{-1}\boldsymbol{y} + \log\det(K_{\boldsymbol{X},\boldsymbol{X}}(\theta) + \sigma^2 I_N)\right]\end{aligned}$$

and plug in the MAP (maximum a posteriori) $\hat{\theta}$ into the predictive distribution of new points $\boldsymbol{X}^{\star}$.

Despite the simplicity of the procedure for GP inference, the main difficulty lies in the matrix inversions required for both estimating hyperparameters and predicting the responses at new points, which involves $\mathcal{O}(N^3)$ time complexity with N being the number of observations. datasets that have more than several thousand observations are already prohibitively expensive for computation. In those cases, a number of approximation methods such as low-rank approximations of the Gram kernel matrix (Nyström method) [70], and judicious selections of subset of observations [4] are often recommended.

4.2.4 Machine Learning, Predictions and Potential Outcomes

GP belongs a large but miscellaneous collection of supervised machine learning algorithms, whose shared goal is to approximate the functional form of the relationship between outcomes and predictors through a data-driven approach. Different machine learning algorithms have different inspirations and fit different situations. For example, LASSO works best for high-dimensional feature space and thus is ideal for causal inference in those setting (genomics and online experiment) [64], and Regression Trees have intuitive interpretations and handle heterogeneous effects well [3]. The advantages of using GP in causal inference, comparatively, is the coherent theoretical framework that yields fully probabilistic inference and the ease of optimizing the kernels. The main disadvantage of GP, however, is scalability: GP with sample size on the scale of tens of thousands can often prove to be computationally intractable.

Predictions at unobserved X's are the central goals in GP models, as illustrated in Eq. (4.4). This works well with the potential outcome framework, since the definitions of causal effects are based on "predictive" values at counter-factual X's. So instead of estimating the coefficients of the treatment indicators, which is the norm of causal inference in traditional parametric regressions, the predictive distributions of the a

large swarm of potential outcomes could be obtained. It is then trivial to a variety of effects of interests with corresponding uncertainty, which might have been elusive under parametric regressions.

4.2.5 GP with Multilevel Structure

Expositionally, GP is most often described as taking inputs that are unstructured and continuous, since kernel functions naturally admit these type of inputs. In the case of inputs with multilevel structures, some careful design of the kernel functions is needed. One approach is to frame this question as a multi-task learning problem, in the sense that outcomes in different studies are deemed as different tasks [6, 75]. In multi-task learning, each task has its own kernel function; moreover, there are correlations between tasks. In this setting, it is common to model these two parts separately and then combine them with matrix operations. For example, assuming a shared within task kernel $\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}}$ and a between-task part $\boldsymbol{U}_{t,t}$, the finite dimensional marginals of the vector-valued random function $\boldsymbol{f}$ is matrix normal distributed

$$\mathrm{vec}\boldsymbol{f}(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}} \otimes \boldsymbol{U}_{t,t})$$

where $\otimes$ denotes the Kronecker product.

This structure is often called kernel separability [2]. Assuming separability can significantly reduce the dimensionality of the problem, and properties of the Kronecker product can be used to make the inference efficient [72]. However, this approach works best for the case of "complete design", which means that the observations take on every combination of (pre-defined) predictors' values. This assumption is reasonable in areas such as computer experiments and robotics, where experiments can be artificially planned and thus data can be collected at pre-specified values of X's across different experiment replications. But it is virtually impossible in fields such as education and public health, in which different schools or experiment cites are unlikely to have study participants or subjects with identical demographics.

The way to incorporate multilevel structure into GP in this work is to add the study indicator as an additional covariate. Take the square exponential kernel for example, assume that the kernel takes in d-dimensional covariates

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma^2 \exp\left(-\frac{1}{2}\sum_{k=1}^{d}\frac{(x_{ik} - x_{jk})^2}{l_k^2}\right),$$

it is simple to add the discrete group indicator term as the $(d+1)^{\text{th}}$ covariate, by defining $x_{i,d+1} - x_{j,d+1} = 1$ if i and j are in the same study, and $x_{i,d+1} - x_{j,d+1} = 0$ otherwise (I will call this the delta metric hereafter). So for observations from the same group, e.g., the same school or the same clinical site, this extra term for group indicator would be zero, and thus has no impact on the correlation structure. On the other hand, if the two observations are from different groups, no matter which groups they are from, the impact to the correlation structure would be the same according to the model. Furthermore, it is often important to add group-level predictors in hierarchical models [22], as it can account for variations that cannot be explained by categorical group membership. Admittedly, a more sophisticated approach that better takes into account the group structure would be welcome, and will be the focus of future research. However, the approach of adding group indicators is straightforward, yet effective, as will be seen in the Project STAR example below.

In square exponential kernels, the lengthscale l_k governs the correlation scale in input dimension κ and the magnitude σ^2 controls the overall variability of the process. Thus the magnitudes of the lengthscale l_k can be used for feature selection; the larger the lengthscale, the more important the corresponding feature.

4.2.6 Between-Study Heterogeneity with GP

In parametric models, such as the example of analyzing Vioxx data in [62], a sequence of models of increasing complexity (e.g., adding treatment-study interactions) can be tested to determine whether between-study heterogeneity exists in the data. In non-parametric modeling, however, high-order interaction terms are automatically

included. One might ask whether the model can be used to assess the existence of heterogeneity. There are two solutions. The first is to use the so-called automatic relevance determination (ARD) feature of the Square Exponential kernel in GP, in particular, the lengthscale hyperparameter l_{d+1} corresponding to the study indicators. If the estimate of l_{d+1} is very large, then the study indicator is irrelevant, which suggests a homogeneous effects across studies. The second is to do a formal model comparison between a GP with the study indicators and a GP without. For example, one can use hold-out predictions to compare two GP's. So in the event of no between-study heterogeneity (after accounting for other predictors), following either of these two model comparison approaches can lead to a simplified GP model without the multilevel structure.

4.3 Revisiting Project STAR

In this section I revisit Project STAR, which studied the effect of early grade class sizes on student achievement. Project STAR, conducted in Tennessee, was a state-wide randomized experiment applied to over 10,000 students from 79 schools that last for 4 years. Each student was randomly assigned to one of three class sizes (13 to 15 students, 22 to 25 students, and 22 to 25 students with a paid teaching aid). End-of-year test scores were used to assess the performance of those students in the areas of math, reading and study skills. Classroom teachers were also randomly assigned to the classes they would teach. The interventions were initiated as the students entered school in kindergarten and continued through third grade, based on the common belief that early intervention has persistent effects well into later lives of the students. Due to its richness and well-crafted design, Project STAR is the most widely studied education experiment in history. Several studies [38, 74] found that attending a small class led to higher end-of-year test scores. However, [39] found the effects on test scores faded out by the eighth grade. Interestingly, later research

showed that even though the effects on test scores didn't persist, attending a small class at an early age increased high school completion rates [21], the likelihood of taking college entrance exams [39] and even adult earnings [12].

4.3.1 Project STAR Design

I briefly recap the design of Project STAR.([74, 38, 21] provide a comprehensive summary.) In 1985-1986 school year, over 6,000 kindergarten students in 79 participating schools were randomly assigned to three class types. Students remained in the same class type through grade 3. Over the course, there were substantial attrition due to students moving away from participating schools or retained in grade. In addition, there were additions as new students entered participating schools in grades 1-3. These new students were randomized into three class types upon entry. Not surprisingly, there were violations from the experimental protocols, as students moved from large to small classes and vice versa. The standard approach in literature, which is also adopted for the analysis herein, is intent-to-treat analysis, i.e., based on initial assignment rather than actual attendance.

Characteristics such as gender, ethnicity, receiving free lunch or not are included in the dataset; furthermore, the general school-level economic conditions is measured by the proportion of students receiving free lunch. As for types of treatment, as I mentioned, there are three types of treatment, small-size classroom, regular-size classroom and regular with a paid teaching aid. I combine regular and regular with an aid as an single control arm, following the standard practice of the literature. At the end of each academic year,students were administered the grade-appropriate standard tests in math and reading. There were used as the educational outcomes for analysis. There are multiple outcomes of interests in the dataset, since very years standardized test were taken for multiple subjects. For the sake of simplicity, I only study one outcome, the end-of-year standardized math test score for students participated in Project STAR in grade 1. Thus all predictors are taken from the students' first grade,

such as grade 1 treatment assignment, grade 1 teacher's experience, etc.

4.3.2 Partial Interference

The first problem, if the potential outcome framework is used for Project STAR, is that the extended SUTVA is highly susceptible: since the intervention was implemented in a classroom environment, peer interactions are highly likely to affect student performance. The potential outcome for a student is not just affected by which class size type was her assigned to, but also what kinds of classmates were she surrounded in. The extended SUTVA needs to be relaxed here.

There is a sizeable literature on relaxing SUTVA in the single study case. [61] studied the effects of moving vouchers on household mobility and defined the concept of "partial interference", i.e., the potential outcomes of a subject are affected by the treatment assignments of other subjects, but only in a small "neighborhood". Similarly, [33] studied the effects of kindergarten retention policy with a dataset including multiple schools. Starting with the full potential outcomes under all possible treatment assignments $\boldsymbol{z}$ and school selections $\boldsymbol{s}$ $Y_i(\boldsymbol{z}, \boldsymbol{s})$, they assume no interference between schools, thus if student i is assigned to school $\boldsymbol{s}_i$ and $\boldsymbol{z}_{\boldsymbol{s}_i}$ denotes the treatment assignments of all the students in school $\boldsymbol{s}_i$, the potential outcome could be written as

$$Y_i(\boldsymbol{z}, \boldsymbol{s}) \equiv Y_i(z_i, \boldsymbol{z}_{-i}, \boldsymbol{s}) = Y_i(z_i, \boldsymbol{z}_{\boldsymbol{s}_i}).$$

[33] further simplified the potential outcome by assuming that the effect of treatment assignments of other students in the same school $\boldsymbol{z}_{\boldsymbol{s}_i}$ is only through a function $\nu(\boldsymbol{z}_{\boldsymbol{s}_i})$

$$Y_i(\boldsymbol{z}, \boldsymbol{s}) = Y_i(z_i, \nu(\boldsymbol{z}_{\boldsymbol{s}_i})). \tag{4.5}$$

In particular, they chose ν to be a binary function denoting high/low retention rate of school $\boldsymbol{s}_i$.

However, since the goal of [33] was on the heterogeneity of effects across schools, their notations didn't admit questions such as how a student would fare were she to

go to another school. Thus their analysis focused on causal questions conditioned on school selection. Assuming no interference between schools I can relax the extended SUTVA assumption

$$Y_i(\boldsymbol{z}, \boldsymbol{s}) \equiv Y_i(z_i, s_i, \boldsymbol{z}_{-i}, \boldsymbol{s}_{-i}) = Y_i(z_i, s_i, \nu(\boldsymbol{z}_{s_i}))$$

where $\boldsymbol{z}_{s_i}$ denotes the treatments assigned to the group of students who go to school s_i under the school selection $\boldsymbol{s}$. Many types of ν can be formulated, reflecting the theoretical conceptions of the influence of peer effects in classroom. One possible choice in the example of Project STAR is the proportion of students in the assigned school receiving free lunch. Furthermore, formulation 4.5 assumes no impact of school/class characteristics, which is quite strong in the education setting. In addition to peer effect, students' exam scores are also likely to be affected by factor such as the level of teacher experience in each school. I later use this formulation in the analysis.

4.3.3 Response Consistency and School Selections

Aside from interference, other key assumptions outlined in the previous sections are unlikely to hold either. First, the consistency assumption, i.e., a student would have the same potential outcomes were she to go to another school, also seems unlikely based on evidence from educational research [45, 25]. Moreover, it is unlikely that all schools are sampled from the same population of students, since the school vary considerably in terms of proportions of students receiving free lunch, running from as low as 31% to as high as 96%. So although within each school, class size treatments are randomized, the full set of potential outcomes, for all combinations of treatments by studies, are not independent of the treatment. More acutely, based on the covariates information available, including student-level (gender, race and receiving free lunch) as well as school-level (free lunch proportion), the assumption that there is no school-selection after conditioning on those covariates seems strenuous at best. To make the no school-selection assumption plausible, I select a subset of the data, specifically

inner city school with a free-lunch rate above 80%. It is then more reasonable to assume that given the covariates, all schools under considerations are sampling from the same population of students. This also reduces the data size to a manageable range for computational considerations (15 schools with 1,300 students). However, the downside of this treatment is that the results obtained in this analysis is not directly comparable with the literature, most of which analyzed the data from all participating schools in Project Star.

4.3.4 Model and Results

I use a GP model that has the score potential outcomes as the outputs. The predictors include school ID (15 inner-city schools), proportion of free lunch students in that school, treatment assignment (2 treatments), student gender (male and female), student ethnicity (minority and non-minority), student receiving lunch or not and teacher's experience (years of teaching). The potential outcome formula is given by

$$\text{score} = f(\text{gender}, \text{ethnicity}, \text{free lunch}, \text{treatment}, \text{teacher experience}, \tag{4.6}$$

$$\text{school ID}, \text{school free lunch proportion}) \tag{4.7}$$

It should be pointed out that gender, ethnicity and free lunch are fixed demographic covariates for a given individual, whereas treatment, teacher experience, school ID and school free lunch proportion can be altered and these lead to the potential outcomes. I call the second types of variables *intervention variables*, since they potentially can serve as interventions. Altering one or more of the intervention variables lead to counter-factuals, such as for a given student what her test score would be if she went to school A with 75% free lunch proportion, were assigned to a small class had a teacher with 10 year of experience. While in Project STAR, class size is the actual intervention, the school ID is the potential source of heterogeneity, and teacher experience and school free lunch proportion are covariates.

To fit a GP model, it is implicitly assumed that the *Unconfounded Treatment*

Assignment within Studies and *Unconfounded Study Selection* hold (given the covaraites). Empirically it is not unreasonable because 1) within each school it is a randomized experiment and 2) the data is further restricted to a homogeneous group of schools (inner city schools with a free-lunch rate above 80%). Furthermore, if the Response Consistency Assumption holds, then

$$f(\text{school A}, :) = f(\text{school B}, :), \forall \text{ School A, B}$$

where : represent keeping all other predictors the same. Since f is modeled as a GP with square exponential kernel, this means that School ID drops out of the GP kernel, or equivalently the lengthscale corresponding to School ID $l_{\text{school ID}}$ is estimated to be a very large number.

All of the above predictors are put into a square exponential kernel. I also introduce acronyms SPoFL (School Proportion of Free Lunch) and YoE (Teacher Years of Experience). Computations are conducted using the MATLAB toolbox `GPStuff` [66].

4.3.4.1 Individual Treatment Effects Heterogeneity across Schools

The first class of causal effects I discuss concerns the individuals. For an idealized student with a fixed demographic profile, e.g., a minority female with low social-economic status (receiving free lunch), how much the small class effect varies across schools. Mathematically, it can be defined as

$$f(\text{Student P}, \text{Small Class}, \text{School A}, \text{YoE}, \text{SPoFL}) - f(\text{Student P}, \text{Regular Class}, \text{School A}, \text{YoE}, \text{SPoFL})$$

In terms of education policy, this can help education administrators and researchers better understand the effect heterogeneities. [46] found little evidence that the small class size effect varies by social-economic status and ethnicity, through adding them as interaction terms in regression models. Here I present a non-parametric approach.

In terms of the potential outcome formula, I calculate the individual causal effects of small class size of a pupil with fixed demographic profiles (gender, ethnicity and social-economic status through receiving free lunch or not). It is also important to fix the other intervention variables in the potential outcome formula—teacher experience and school level proportion of the free lunch—to have a well-defined intervention. These variables are kept at the median level at respective schools, which represents the typical conditions the student would be exposed to if she were to attend this school.

Corresponding to the pupil (female, minority, and from a low social-economic status family), there are 2×15 different potential outcomes, for all treatment by school combinations. For each school, there is a two-dimensional vector corresponding to small-size and regular-size outcomes. Using Gaussian Process, the posterior distribution of the 2-dimensional vector can be derived in the form of a bi-normal distribution

$$\begin{pmatrix} f(x_{\text{trt}}) \\ f(x_{\text{ctrl}}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} a^2 & \rho ab \\ \rho ab & b^2 \end{pmatrix} \right)$$

Then the treatment effect has the distribution $N(\mu_1 - \mu_2, a^2 + b^2 - 2\rho ab)$.

Fig. 4.1 gives the results for the idealized pupil. Clearly, the treatment effect for this pupil is not homogeneous across different schools. In fact, the effects could have reversed signs for some schools, which seem to contradict with common accepted conclusion,. It is worth noting that, however, those schools with negative small class effects have either a small sample size (40-60 students) or a less experienced teacher body (1-5 years of teaching experience). It also needs to be pointed out that the sample investigated here is a subpopulation of the Project STAR data. This heterogeneity might be due to a number of reasons, such as the demographic compositions, the teachers' general level of experience and particular level of experience with specific demographic groups, etc. Identifying the mechanism behind the varying receptiveness of the intervention would be important for the next step in the study. And the education policy makers would need to take this into account had they faced the

problem of assigning the pupils to schools.

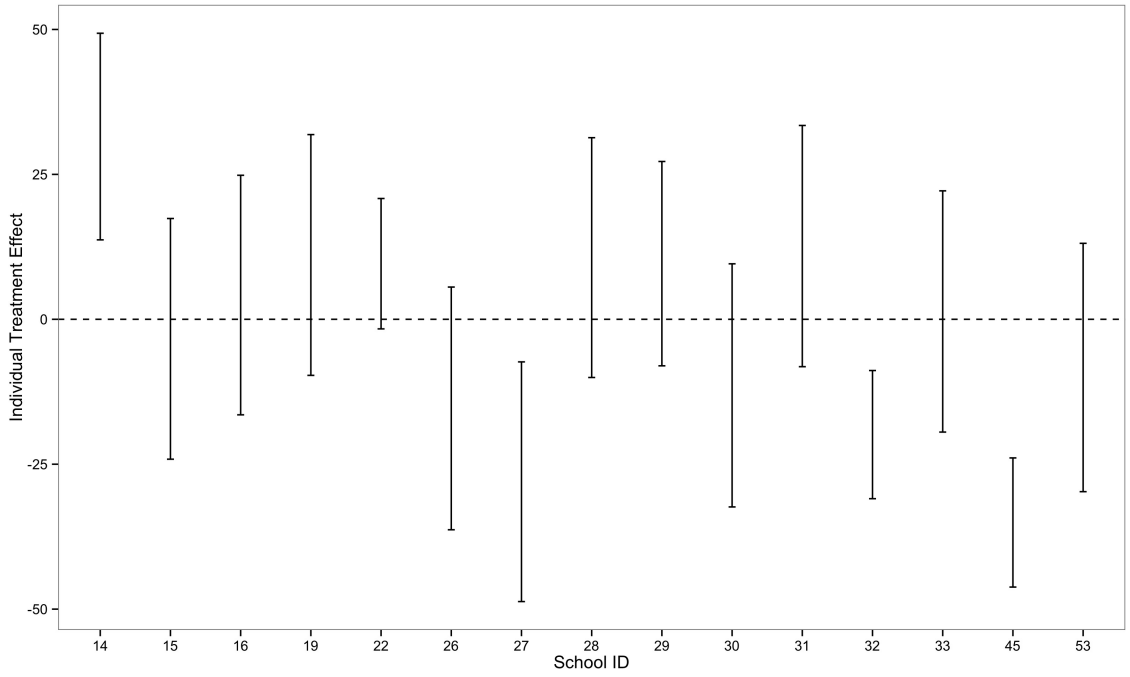

Figure 4.1: The causal effect of attending a small-size class for an pupil (female, minority, and from a low social-economic status family), if she were to attend each of the 15 schools included in the dataset. The error bars represent one standard deviation of the posterior distribution while the center is the posterior mean. Almost all of them overlap with 0. There are also large amount of heterogeneities across schools.

4.3.4.2 Average Treatment Effects within Schools

The second class of causal effect I discuss here is the average treatment effects within schools. In this case, the potential outcomes of students were they assigned to different schools are of no interest to the researchers. Instead, the quantity of interest is, for fixed demographic profile of each school, how much the school district can benefit on

average from small class size. Mathematically, it can be defined as

$$E_{\mathrm{P}\in\mathrm{A}}[f(\text{Student P}, \text{Small Class}, \text{School A}, \text{YoE}, \text{SPoFL}) - f(\text{Student P}, \text{Regular Class}, \text{School A}, \text{YoE}, \text{SPoFL})]$$

This is sometimes the relevant policy question to ask when the whole school as a unit is the focus. This analysis can identify those schools not receptive to the small class effects and thus save potential waste of resources in implementing class size adjustments. [38] examine the school-by-school heterogeneity by fitting a separate regression model for each school, and use it to verify the robustness of the overall treatment effect rather than to estimate conditional effects.

To calculate average treatment effect within a school, for each pupil in the school, we find the posterior means of the potential outcome function setting small class size treatment at 1 and 0 respectively, and then calculate the difference. The average of this difference across all pupils in the school gives us the average treatment effects within the school. Similarly, I calculate the average treatment effects for all schools, and present the results in Fig. 4.2. Note that this graph conveys a different causal interpretation compared with Fig. 4.1. Here we are looking at the conditional average effect of attending a small size class for the representative population of each school. Clearly, for some schools, the positivity of small class size effect is highly susceptible, and it raises concerns about the effectiveness of this treatment in those school districts. The implications for education policy makers is to take more considerations in deciding whether to carry on full-scale reform in those school districts.

4.4 Discussion

In this chapter, I discuss an extended potential outcome framework originally built for meta-analysis and extend it to more general multilevel data. Compared with the classic potential outcome framework, a plethora of counterfactuals with group structures need to be created to handle multi-level data. I then introduce a GP-based

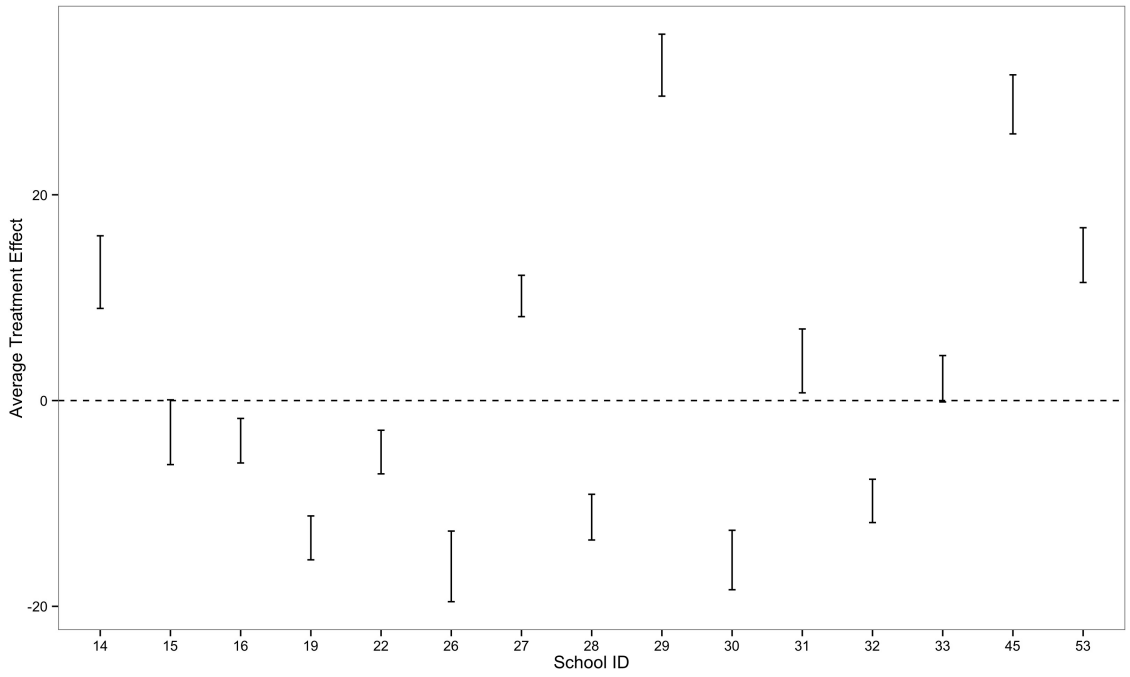

Figure 4.2: The average causal effect of attending a small-size class for students attending each of the 15 schools included in the data. The error bars represent one standard deviation of the posterior distribution while the center is the posterior mean. Although the consensus of literature is that small class effect size is unequivocal, analysis here shows the variations are rather large, with some schools actually having negative treatment effects.

approach. The main advantages of GP, and in general non-parametric methods, is the fidelity of inferential uncertainty. The central question to the extended potential outcome framework is how to incorporate the group structure. The design of Project STAR dictate that a relaxation of the extended SUTVA is needed to accommodate peer influence common in educational study. A partial interference structure is proposed and the functional form of the peer influence is discussed. The full potential outcomes can be easily derived from the posterior distributions of Gaussian Processes, and it allows a great amount of richness for estimating the causal quantify of interest. I illustrate this point with the example of Project STAR data.

There are still a lot of caveats for the use of GP in multilevel causal analysis. First of all, I only use a small fraction of the data available from Project STAR. There are a lot of research on speeding up GP models, and I hope advances on this front will enable a full analysis of the Project STAR data to uncover more interesting patterns. Second, I didn't fully explore the full modeling capability of GP. It is possible that there are more appropriate types of kernels or combinations of kernels than the ubiquitous square exponential kernels used in this chapter. However, this chapter is just a first step. I am working on a full-scale paper on exploring GP models in tackling structured causal inference.

Chapter 5

Bibliography

[1] Murray Aitkin. Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, 18(17-18):2343–2351, 1999.

[2] Mauricio A Alvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *arXiv preprint arXiv:1106.6251*, 2011.

[3] Susan Athey and Guido Imbens. Machine learning methods for estimating heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*, 2015.

[4] Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 70(4):825–848, 2008.

[5] Douglas Bates, Martin Maechler, and Ben Bolker. *lme4: Linear mixed-effects models using S4 classes*, 2013. R package version 0.999999-2.

[6] Edwin V Bonilla, Kian M Chai, and Christopher Williams. Multi-task gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2007.

[7] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-

validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

[8] Prabir Burman, Edmond Chow, and Deborah Nolan. A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358, 1994.

[9] Matthew K. Buttice and Benjamin Highton. How does multilevel regression and poststratification perform with conventional national surveys? *Political Analysis*, 21(4):449–467, 2013.

[10] James E Campbell. *The American campaign: US presidential campaigns and the national vote*, volume 6. Texas A&M University Press, 2008.

[11] M Keith Chen, Jonathan E Ingersoll, and Edward H Kaplan. Modeling a presidential prediction market. *Management Science*, 54(8):1381–1394, 2008.

[12] Raj Chetty, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. How does your kindergarten classroom affect your earnings? evidence from project STAR. Technical report, National Bureau of Economic Research, 2010.

[13] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, pages 266–298, 2010.

[14] Yeojin Chung, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, 78(4):685–709, 2013.

[15] Harris Cooper and Erika A Patall. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, 14(2):165, 2009.

[16] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

[17] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

[18] Vincent Dorie. *blme: Bayesian Linear Mixed-Effects Models*, 2013. R package version 1.0-1.

[19] Robert S Erikson and Christopher Wlezien. Are political markets really superior to polls as election predictors? *Public Opinion Quarterly*, 72(2):190–215, 2008.

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

[21] Jeremy D Finn, Susan B Gerber, and Jayne Boyd-Zaharias. Small classes in the early grades, academic achievement, and graduating from high school. *Journal of Educational Psychology*, 97(2):214, 2005.

[22] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[23] Andrew Gelman, David K. Park, Boris Shor, and Jeronimo Cortina. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do, second edition*. Princeton University Press, 2009.

[24] Yair Ghitza and Andrew Gelman. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776, 2013.

[25] Robert James Gordon, Thomas J Kane, and Douglas Staiger. *Identifying effective teachers using performance on the job*. Washington, DC: Brookings Institution, 2006.

[26] Harold F Gosnell. How accurate were the polls? *Public Opinion Quarterly*, 1(1):97–105, 1937.

[27] Robert M Groves. *Survey errors and survey costs*, volume 536. John Wiley & Sons, 2004.

[28] Julian Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A*, 172(1):137–159, 2009.

[29] Julian Higgins, Anne Whitehead, Rebecca M Turner, Rumana Z Omar, and Simon G Thompson. Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine*, 20(15):2219–2241, 2001.

[30] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 2011.

[31] D Sunshine Hillygus and Todd G Shields. *The persuadable voter: Wedge issues in presidential campaigns*. Princeton University Press, 2009.

[32] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

[33] Guanglei Hong and Stephen W Raudenbush. Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, 101(475), 2006.

[34] Simon Jackman. Pooling the polls over an election campaign. *Australian Journal of Political Science*, 40(4):499–517, 2005.

[35] Satyen Kale, Ravi Kumar, and Sergei Vassilvitskii. Cross-validation and mean-square stability. In *Innovations in Computer Science,*, pages 487–495. Tsinghua University Press, 2011.

[36] Scott Keeter, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. Gauging the impact of growing nonresponse on estimates from a national rdd telephone survey. *Public Opinion Quarterly*, 70(5):759–779, 2006.

[37] Andrew Kohut, Scott Keeter, Carroll Doherty, Michael Dimock, and Leah Christian. Assessing the representativeness of public opinion surveys. *Pew Research Center for The People & The Press*, 15(May):2012, 2012.

[38] Alan B Krueger. Experimental estimates of education production functions. Technical report, National Bureau of Economic Research, 1997.

[39] Alan B Krueger and Diane M Whitmore. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *The Economic Journal*, 111(468):1–28, 2001.

[40] Jeffrey R Lax and Justin H Phillips. How should we estimate public opinion in the states? *American Journal of Political Science*, 53(1):107–121, 2009.

[41] Roderick JA Little. Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88(423):1001–1012, 1993.

[42] Kari Lock and Andrew Gelman. Bayesian combination of state polls and election forecasts. *Political Analysis*, 18(3):337–348, 2010.

[43] B. Mandelbrot. On the language of taxonomy: An outline of a "thermostatistical" theory of systems of categories with willis (natural) structure. In Colin Cherry, editor, *Information Theory—Third London Symposium*, pages 135–145, 1955.

[44] Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Roczniki Nauk Rolniczych Tom X [in Polish]*, 1923. translated in *Statistical Science*, 5, 465-480.

[45] Barbara Nye, Spyros Konstantopoulos, and Larry V Hedges. How large are teacher effects? *Educational evaluation and policy analysis*, 26(3):237–257, 2004.

[46] Barbara A Nye, Larry V Hedges, and Spyros Konstantopoulos. Do the disadvantaged benefit more from small classes? evidence from the tennessee class size experiment. *American Journal of Education*, pages 1–26, 2000.

[47] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[48] Eva Petkova, Thaddeus Tarpey, Lei Huang, and Liping Deng. Interpreting meta-regression: application to recent controversies in antidepressants efficacy. *Statistics in Medicine*, 32(17):2875–2892, 2013.

[49] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2012. R package version 3.1-104.

[50] Phillip N. Price, Anthony V. Nero, and Andrew Gelman. Bayesian prediction of mean indoor radon concentrations for minnesota counties. *Health Physics*, 71:922–936, 1996.

[51] Stephen W Raudenbush. Analyzing effect sizes: Random-effects models. In Harris Cooper, Larry V Hedges, and Jeffrey C Valentine, editors, *The Handbook of Research Synthesis and Meta-Analysis*, pages 295–316. New York: Russell Sage Foundation, 2009.

[52] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[53] David Rothschild. Forecasting elections comparing prediction markets, polls, and their biases. *Public Opinion Quarterly*, 73(5):895–916, 2009.

[54] David Rothschild. Combining forecasts: Accurate, relevant, and timely, 2013. Working paper.

[55] Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral Statistics*, 2(1):1–26, 1977.

[56] Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469), 2005.

[57] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*, volume 12. Cambridge University Press, 2003.

[58] Matthias W Seeger. Cross-validation optimization for large scale structured classification kernel methods. *Journal of Machine Learning Research*, 9:1147–1178, 2008.

[59] Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

[60] Mark C Simmonds, Julian PT Higgins, Lesley A Stewart, Jayne F Tierney, Mike J Clarke, and Simon G Thompson. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials*, 2(3):209–217, 2005.

[61] Michael E Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

[62] Michael E Sobel, David B Madigan, and Wei Wang. Meta-analysis: a causal framework, with application to randomized studies of vioxx. *Psychometrika*, 2016. forthcoming.

[63] Peverill Squire. Why the 1936 literary digest poll failed. *Public Opinion Quarterly*, 52(1):125–133, 1988.

[64] Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. Heterogeneous treatment effects in digital experimentation. *arXiv preprint arXiv:1412.8563*, 2014.

[65] Catrin Tudur Smith, Paula R Williamson, and Anthony G Marson. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine*, 24(9):1307–1319, 2005.

[66] Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. Gpstuff: Bayesian modeling with gaussian processes. *The Journal of Machine Learning Research*, 14(1):1175–1179, 2013.

[67] Aki Vehtari and Janne Ojanen. A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

[68] Wei Wang and Andrew Gelman. Difficulty of selecting among multilevel models using predictive accuracy. *Statistics and Its Interface*, 7:1–8, 2014.

[69] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2014.

[70] Christopher Williams and Matthias Seeger. Using the nystrom method to speed up kernel machines. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, volume EPFL-CONF-161322, pages 682–688, 2001.

[71] CKI Williams and CE Rasmussen. *Gaussian processes for machine learning*. Cambridge: MIT Press, 2006.

[72] Andrew Wilson, Elad Gilboa, John P Cunningham, and Arye Nehorai. Fast kernel learning for multidimensional pattern extrapolation. In *Advances in Neural Information Processing Systems*, pages 3626–3634, 2014.

[73] Justin Wolfers and Eric Zitzewitz. Prediction markets. Technical report, National Bureau of Economic Research, 2004.

[74] Elizabeth R Word et al. The state of Tennessee's Student/Teacher Achievement Ratio (STAR) project: Technical report (1985-1990). Technical report, Tennessee State Department of Education, 1990.

[75] Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine learning*, pages 1012–1019. ACM, 2005.

Appendix A

Grouping of States by Contestedness

For ease of interpretation, in Figure 3.1 states are grouped into 4 categories: (1) battleground states (Colorado, Florida, Iowa, New Hampshire, Ohio, and Virginia), the five states with the highest amounts of TV spending plus New Hampshire, which had the highest per-capita spending; (2) quasi-battleground states (Michigan, Minnesota, North Carolina, Nevada, New Mexico, Pennsylvania, and Wisconsin), which round out the states where the campaigns and their affiliates made major TV buys; (3) solid Obama states (California, Connecticut, District of Columbia, Delaware, Hawaii, Illinois, Maine, Maryland, Massachusetts, New Jersey, New York, Oregon, Rhode Island, Vermont, and Washington); and (4) solid Romney states (Alabama, Alaska, Arizona, Arkansas, Georgia, Idaho, Indiana, Kansas, Kentucky, Louisiana, Mississippi, Missouri, Montana, Nebraska, North Dakota, Oklahoma, South Carolina, South Dakota, Tennessee, Texas, Utah, West Virginia, and Wyoming).