ABSTRACT

Title of dissertation: GENERALIZED OBSERVED BEST PREDICTION WITH EMPIRICAL BAYES PARAMETRIC BOOTSTRAP MODEL BUILDING

William R. Waldron, Doctor of Philosophy, 2020

Dissertation directed by: Professor Partha Lahiri
Department of Mathematics &
Joint Program in Survey Methodology

The observed best predictor (OBP) has been recently offered as a more robust alternative to the remarkable empirical best linear unbiased predictor (EBLUP). Although the latter has become a pervasive tool among applied statisticians, there are critical reasons why the OBP should almost always be used in conjunction with the EBLUP. In particular, mathematical models are often oversimplified or misspecified, lacking key predictors within the available set of data. For more complex models such as time-series applications, model robustness becomes even more imperative. We will provide some results related to the OBP theory and introduce a generalized, or weighted version of the OBP for different loss functions. This will first be defined on the Fay-Herriot model and then extended to the General Linear Mixed model. Finally, we will apply the best predictive estimator (BPE) to both parameter coefficients and variance parameters within the Fay-Herriot and cross-sectional time series models.

Model building strategies abound, and have continued to evolve. These are instrumental for applied statisticians and analysts passing judgement on whether statistical models are suitable for drawing conclusions or producing official estimates. A number of methodologies and approaches have been developed to consider this critical question of model selection and diagnostics. We endeavor to view this problem from the perspective of empirical Bayes (EB) - in a similar fashion as the EBLUP. As such, we define and develop an EB parametric bootstrap approach not only to estimate mean squared error, but also for finding the best model from a set of candidates (e.g., variable selection). This could be done for general criteria by considering leave-one-out predictive distributions. Once a viable model is selected, we can continue the model-building process by performing appropriate validation. Thus, the method is not only versatile, but has some computational advantages over other model building strategies.

GENERALIZED OBSERVED BEST PREDICTION WITH EMPIRICAL BAYES PARAMETRIC BOOTSTRAP MODEL SELECTION AND DIAGNOSTICS

by

William Rene Waldron

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Partha Lahiri, Chair/Advisor
Professor Paul J. Smith
Professor Robert Fay
Professor Yan Li
Professor Cinzia Cirillo
Professor Takumi Saegusa

© Copyright by
William R. Waldron
2020

Dedication

Dedicated to my wife, Irada, whose encouragement, patience, and love made this research possible.

Acknowledgments

First and foremost, I want to express gratitude to my advisor Partha Lahiri for his guidance and direction during this entire process. He has a most remarkable work ethic, which I found inspiring. His professionalism and ability to concentrate are two other traits I hope to emulate as I begin a new chapter of my academic and professional career.

I also want to thank the doctoral committee, for their insightful and penetrating questions during my dissertation defense. Their level of engagement, and attention to crucial details with regard to the research was surprising and unexpected. I received the final phase of my formal education during that conversation, I certainly benefited in tangible and memorable ways from their knowledge and wisdom. I thank each of you for your contributions. In particular, Robert Fay, whose eponymous model is constantly referenced in the dissertation, made lasting impressions that will impact the direction and quality of my future work.

Finally, I want to thank the incredible faculty at the University of Maryland Dept. of Mathematics. They are certainly very dedicated not only to the study of statistics, but to pedagogy, as well. I have always found them receptive to student questions and patient with their technical explanations. In particular, I'd like to recognize Paul Smith, Eric Slud, and Abram Kagan. Not only were they superb lecturers, but their perspectives on statistical theory had a profound impact in shaping my scientific knowledge and intuition.

Table of Contents

Chapter 1

Introduction

1.1 Small Area Estimation

Historically, national statistical institutes and the broader survey industry have been reluctant to adopt modeling solutions over traditional design-based methods. Nevertheless, the techniques known collectively as *Small Area Estimation* have burgeoned in the past 20 years. There is growing popularity and acceptance for these methods within both official statistics and the private sector due to their effectiveness and utility. This is particularly true under the advent of greater amounts of auxiliary data becoming available. Small area methods encompass a wide variety of techniques for estimating population parameters. They have expanded to cover a variety of circumstances that arise in practice, yet there is always a common thread: limited information from the primary source containing the outcome variable(s).

What exactly constitutes a small area? Firstly, it need not be based on a strictly geographical designation, the small *area* may represent any domain in which the direct estimate is unreliable. Secondly, one may also ask: *What determines an adequate sample size for an estimate to be reliable*? Professional opinion pollsters for the 2020 election will sample about 1,000 people for estimates to attain a margin of error of 3% for dichotomous responses under a 95% confidence level. In some applications, the effective sample size can be diminished from the effects of clustering, underlying

probabilities of selection, and other design effects. When such a high level of precision is untenable or unwarranted, some authors recommend a sample size of fifty as a general lower bound to maintain quality estimates from a sample (Lohr (2010)). Others advocate thirty (the so-called thirty rule) as the minimum practical sample size (the "thirty" is derived from the notion that the t-distribution with 30 degrees of freedom begins to approximate the standard normal distribution). In practice, the necessary amount of sample is a function of the variance in the underlying population: the greater the variation, the larger the samples required for stable estimates. Conversely. when a population parameter is relatively stable, fewer samples will suffice.

Direct Estimates and Design-Based Estimation

In general, we are striving to obtain the most precise estimates which are within our resources to acquire. Thus, the most desirable solution is to obtain more sample points and proceed to use a direct domain estimator, or **direct estimate**. This is a survey estimator based solely on sample points collected within the domain and is the most traditional form of survey estimation. This is also referred to as *design*-based, where the only randomness is derived from the selection of one particular sample over the $\binom{N}{n}$ possible samples of size n from a population of size N. Various methods, some of which could be argued to be a *model*, have been developed to improve design-based estimation. Different sampling approaches (e.g., systematic, probability proportional to size) and sampling stages can be more efficient and advantageous than simple random sampling. Stratification permits the survey administrator to exercise more control over the areas where the sample is to be located - and stratification generally can only *improve* sample inference. The use of sample weights can ensure estimators remain unbiased and can incorporate known population covariate informa-

tion (e.g., race/ethnicity, age group, gender) to provide additional power when projecting survey estimates onto the population. Moreover, sample weighting can also adjust the survey estimates to correct for any non-response bias. Indeed, the basic expansion estimator $\hat{Y} = \sum_{i=1}^{n} w_i y_i$ is really a model which can admit a good deal of information within its weights, although the variability of the weights (especially zero weights or very large weights) can be a concern when too much information is attempted to be placed within the weights. There is a trade-off between bias and variance, therefore, when adding more control total constraints to the weights w_j. When $w_j = 1/\pi_j$, the inverse of the probability of selection π_j of the j^{th} sample point, the expansion estimator is called the Horvitz-Thompson estimator and its properties have been well-studied in design-based literature (see Särndal et al. (1992) or Cochran (1977) for details).

These accommodations with direct domain estimation have really become standard professional practice amongst survey administrators and analysts. Those who are able obtain viable survey estimates of acceptable quality using available resources do not require further recourse. When circumstances are not so, there are some limitations in how far purely design-based methods can further reduce design effects beyond the precision from a simple random sample, which is a function only of sample size. The classical texts on design-based estimation include Hansen et al. (1953), Kish (1965), and Cochran (1977). More recent treatments can be found in Särndal et al. (1992), Lohr (2010), and Fuller (2009).

Another direct estimator which takes advantage of domain-specific auxiliary information is that of the generalized regression estimator, or GREG. This is a regression based estimator that

updates the sample weights w_i of the basic expansion estimator from $\hat{Y} = \sum_{i=1}^{n} w_i y_i$ to obtain new weights w_i^* for the revised expansion estimator $\hat{Y}_{GR} = \sum_{i=1}^{n} w_i^* y_i = \hat{Y}_{GR}(\mathrm{y})$. This can be done under the presence of known population totals $X = (X_1, X_2, \ldots, X_p)'$ which are available for p characteristics, and when this information is also observed during data collection from each respondent j, taking the form x_j. The GREG estimator is rather versatile, as the weights w_j^* are independent of y and can be applied to any response variable. The expansion estimator is linear in the sense that for any two survey variables y and v, $\hat{Y}_{GR}(\mathrm{y+v}) = \hat{Y}_{GR}(\mathrm{y}) + \hat{Y}_{GR}(\mathrm{v})$. In fact, GREG estimation (sometimes called the *calibration* estimator) satisfies two properties related to iterative proportional fitting (IPF, or *raking* based on Deming & Stephen (1940)). Weighted marginal totals for the observed x_i's are summable to the population and the GREG minimizes a distance function between the original weights w_j and the updated weights w_j^* (in the sense of *chi-squared* distances $\sum_{i=1}^{n} c_j (w_j - w_j^*)^2 / w_j$ for some c_j).

First Steps: Synthetic Estimation

When sample increases are impractical or cost prohibitive, alternative methods may achieve greater reliability than that of direct estimators. If we are able to demonstrate that some auxiliary information is strongly correlated with our outcome of interest, we may be willing to accept the model bias incurred from assuming a mathematical relationship between characteristics in the underlying population. In this way, the small area estimate can "borrow strength" to improve the *precision* of the estimate at the expense of some *accuracy* (i.e., bias) deemed to be negligible.

The first modern example given in the survey methods literature regarding the use of these techniques is the 1945 Radio Listening Survey conducted by the Census Bureau at the behest of the Federal Communications Commission (Hansen et al. (1953)). A large mail survey was sent out to about a thousand households in each of 500 county areas. They were asked for voluntary responses regarding their radio listening habits. A much smaller subsample was taken from among 85 county areas. These households were visited by a personal interviewer to measure their radio listening. A model was fit within the 85 counties between the interview responses and the mail survey. Once a viable mathematical formula was determined (in the form of a ratio estimator), it was leveraged to obtain interviewer estimates within the non-sampled 500 - 85 = 415 county areas, even though no interviews were actually conducted there. The correlation between the mail survey estimates and the in-person survey estimates was measured at 0.70. Such a high correlation is a good indication of the potential effectiveness of the ratio estimator. This was applied to the results of the mail survey in the remaining 415 county areas, to obtain survey responses without additional personal interviews.

The ratio estimator in the Radio Listening Survey is an example of a *synthetic* estimator. This is an indirect domain estimator derived from reliable direct survey estimates collected over a broader domain. It relies upon the assumption that the characteristics in the larger area are consistent with those of the smaller area, permitting the extrapolation to the smaller domain. In fact, synthetic estimation can even be applied to non-sampled areas given the availability of accurate area-specific auxiliary data (Rao & Molina (2015)). The GREG estimator described above has been modified to be used as a synthetic estimator. When only a single population characteristic is observed, the synthetic GREG estimator reverts to a ratio estimator. The National Center for Health Statistics

(NCHS) helped pioneer these methods in 1968 by deriving a special case of the synthetic GREG estimator to obtain state-level measurements for disability and other health outcomes using the National Health Interview Survey (NHIS).

Another important synthetic estimator is the *structure-preserving* estimator, or SPREE. This is described as below in Rao & Molina (2015). Purcell & Kish (1980) investigated the One- and Two-Step variations of the SPREE. Suppose that Decennial Census counts N_{iab} were available for a sub-population in the i^{th} area for category a (e.g., citizenship status) of an outcome variable of interest, and category b from a closely related characteristic (e.g., age group). Suppose these two variables have A and B total such categories, respectively. The unknown current counts are M_{iab}, but we are really only interested in estimating $M_{ia\cdot} = \sum_{b=1}^{B} M_{iab}$. For the One-Step SPREE, reliable estimates over large areas are available for the interaction between the outcome and auxiliary variable at all levels: $\hat{M}_{\cdot ab}$. Similar to the GREG, the SPREE will minimize a set of chi-square distances to obtain an estimator for the current year for the small area counts over the main outcome: $\hat{M}_{ia\cdot} = \sum_{b=1}^{B} \frac{N_{iab}}{N_{\cdot ab}} M_{\cdot ab}$. In this case, the SPREE is the sum of census-derived ratios applied to current year interaction totals.

When reliable small area totals $\hat{M}_{i\cdot\cdot}$ are also available, then the Two-Way SPREE is essentially iterative proportional fitting with the $\hat{M}_{i\cdot\cdot}$ as the initial "weights". The SPREE estimator in both cases preserves the marginal totals from the more reliable direct survey estimators taken across broader areas of the population. Moreover, the one-way, two-way, and even three-way interactions between the variables and areas in the Census year and the analysis year are preserved, thus giving

the eponym. That is, for all areas i and i' and a and a':

$$\frac{\hat{M}_{iab}}{\hat{M}_{i'ab}} = \frac{N_{iab}}{N_{i'ab}} \qquad \text{and} \qquad \frac{\hat{M}_{iab}\hat{M}_{i'a'b}}{\hat{M}_{i'ab}\hat{M}_{ia'b}} = \frac{N_{iab}N_{i'a'b}}{N_{i'ab}N_{ia'b}}.$$

Composite Estimators and James-Stein

Small area estimators basically comprise averages between unstable, yet unbiased, direct estimators and possibly very biased synthetic estimators with lower design variance. Once the specific versions of the direct and synthetic estimators have been established, then all that remains is to determine the composite weight. Although it may seem there is a good deal of subjectivity involved, composite estimators apparently arise very naturally within many intuitive model designs. For a direct estimate of small area i, $\hat{y}_i$, and a synthetic estimator $\tilde{y}_i$, the composite estimate is given by:

$$y_i^C = \phi_i \hat{y}_i + (1 - \phi_i)\tilde{y}_i,$$

where $0 < \phi_i < 1 \ \forall i = 1, 2, \ldots m$. Various methods have been attempted to find the optimal value for the ϕ_i's in the sense of minimizing the mean-squared error in *each* area. Despite sound origins, a number of such algorithms have been somewhat unstable and therefore less desirable. Conversely, sample size based derivations for ϕ_i are less volatile but lack the optimality considerations. Purcell & Kish (1979) found a constant value for $\phi_i \equiv \phi$ which was fixed across all areas that was stable and optimal *across* areas. That is, it minimized the MSE across all areas, rather than individually being optimal for any particular area. This leads us to the *celebrated* James-Stein (JS) estimator (James & Stein (1961)), which astonished the scientific community by producing an

example in which the maximum likelihood estimator was inadmissible. Efron & Morris (1973) later showed JS was equivalent to an empirical Bayes solution. This will be discussed in greater detail in Chapter 2.

The Fay-Herriot Model and its Extensions

The landmark paper by Fay & Herriot (1979) is really the cornerstone and the beginning of small area estimation as a separate discipline. This is now referred to as the *Area Level Model* to be used when direct estimates for small domains are readily available and with design variance assumed to be known. Its power and simplicity has had a remarkable impact on applied statistics, fostering Bayesian methods into prominence within survey applications.

The Fay-Herriot model spurred an enormous amount of research branches as it was extended in different directionsso it could be utilized in a variety of scenarios. For example, the model could be generalized from scalar estimation to vector estimation. It is commonplace for investigators to measure multiple outcomes within each domain - surveys tend to have multiple questions for an acquired respondent. When these independent variables within the domain are correlated, there can be efficiency gains in *joint* estimation as opposed to separate individual models. A *multivariate* Fay-Herriot model was introduced by Fay (1987) and further developed by Datta, Fay, & Ghosh (1991).

The same survey estimate could be correlated across domains, as well. Area-level estimates in geographically proximal areas tend to exhibit *spatial* correlation, which if exploited could be advantageous to improving the reliability of estimates. Cressie (1991) used a spatial small area

model to adjust for undercounts in the Decennial Census. Similar spatial models have been used to measure soil erosion in Iowa lakes (Petrucci & Salvati (2006)) and measure per capita income in “local economic regions" in the Tuscany region in central Italy (Pratesi, M., Salvati (2008)).

Battese et al. (1988) constructed a model related to the Fay-Herriot area level model which could instead be applied directly to respondent-level data. This is now referred to as the *Unit Level* model within the SAE literature. Indeed, this was also an important breakthrough in the evolution of the SAE theory because it permitted the methods to be used on more granular source data when it is available. Datta & Ghosh (1991) extended the unit level model to the general linear mixed model case.

SAE researchers began to handle practical issues arising in real-world applications. Methods were created for both area and unit level models to address typical problems in linear models: non-normal error terms, count and binary data with different linking functions, and problems of model selection diagnostics. Prasad & Rao (1990) developed a formula for mean-squared error that deconstructed the total MSE into three components, that coming from the first level sampling variance, error introduced from estimating unknown parameter coefficients, and finally error coming from estimating the unknown *variance* components.

The process of "benchmarking" was taken up to force the small area estimates to be consistent with reliable direct estimates. That is to say, the modeled small area totals should be equivalent to the direct estimate of the total measured on the aggregated areas. Benchmarking techniques remain

an active area of research, given their complexity in trying to retain the optimality aspects of the small area estimates. Pfefferman & Barnard, C. (1991) demonstrated the impossibility in obtaining linear unbiased predictors for each small area that satisfy the benchmarking property.

1.2 Linear Mixed Model Theory

The foundations of small area estimation are based upon the theory of mixed linear models. This was developed in the context of animal breeding applications in rural Iowa in the early 1950s. In a series of papers between 1950 and 1975, Henderson stated and solved the mixed model equations and provided a technique to handle unbalanced data. Furthermore, he expressed his solutions in a matrix format that avoided difficult matrix inversions and were more easily computable before the dawn of computers. We also mention that he introduced three different methods to estimate unknown variance components (called methods I, II, and III). Finally, Henderson derived the best linear unbiased predictor (BLUP) by assuming normality and maximizing a *pseudo*-likelihood function to obtain the BLUP. See Henderson (1950), Henderson (1953), and Henderson (1975) for details. Robinson (1991) provides an excellent exposition of the evolution of mixed models and the BLUP. An account of Henderson's biography and contributions to statistics is given by Searle (1990).

Searle et al. (2006) laments the problems being encountered by the late 1960's in using ANOVA to estimate variance components. This included negative estimates, lack of distributional properties, and no clear manner to compare ANOVA between models. Henderson's Method I was manageable, but the advent of computers made maximum likelihood an attractive concept. The

seminal paper by Hartley and J.N.K. Rao (1967) provided a means to calculate variance components for a wide class of mixed effects models, with or without covariates, both balanced and unbalanced. It was Thompson (1962) who first suggested maximizing the portion of the likelihood invariant to the location parameters - and this came to be known as the restricted maximum likelihood estimator, or REML. These remain the most popular methods of variance component estimation, although they are quite similar and both susceptible to the same computational challenges. Indeed, they both still permit zero and negative estimates. Other methods of have been proffered, including minimum norm quadratic unbiased estimates (MINQUE) and Bayesian approaches. Although MINQUE is attractive since there are no iterations or distributional assumptions (just need to solve a system of equations), Searle et al. (2006) faults the subjective nature of the *a priori* values needed to provide estimates. As for pure Bayes methods, Monte Carlo Markov Chain procedures (MCMC) can be used to obtain variance parameter estimates comparable to those obtained with REML and ML.

1.3 Repeated Surveys over Time

National surveys measuring important economic and health behaviors are often repeated on an annual basis in order to monitor trends and changes in the population. Sample designs will ensure that survey estimates for the current year have adequate precision at the *national* level. On the other hand, there will always be smaller domains without sufficient sample size for viable measurement within a single year. Small area researchers have found ways for analysts to address low sample sizes and improve the measurement within the *current* year. However, each method will have advantages and trade-offs, so the selection of the most desirable will remain dependent upon

individual scenarios and the objectives of a given study.

The most straightforward way to accomplish improved inference for smaller areas is to simply append data sets from multiple years into a single dataset. Obviously, this requires a willingness to dilute the "currency" of the present year with information from prior years. The NHIS produces state-level health outcome estimates by pooling together three survey years (see NCHS (2018)). This procedure assumes the survey questions remain unchanged from year to year and that no individuals were present in multiple sample years.

In the Current Population Survey (CPS), participants are invited to report their status in multiple time iterations of the survey (thus saving some costs associated with data collection). From the standpoint of variance estimation, it is easier to compute the standard errors when each survey has the same sample design, and this is usually the case. Korn & Graubard (1999) provide instruction for appropriate adjustments under various scenarios under the Taylor series variance estimation approach (e.g., same strata, different primary sampling units). Rizzo et al. (2008) provided guidance on how to combine yearly datasets using jackknife replicate variance estimation using different statistical software packages. The National Survey of Drug Use and Health (NSDUH) will pool together two years of data, but will also utilize information from other surveys and any administrative sources. This is sometimes referred to as *data integration*, and represents a closely related area of specialized survey research. The focus of this chapter, however, will be borrowing strength across both time and other sample areas, from the *same* survey.

Pooling together datasets from multiple years does have certain challenges. The procedure may be likened to moving averages: where estimates from individual years are averaged together. Even when the mean average (equal weights for all years) is not taken, and more recent time estimates are given greater weight in the average, the meaning and interpretation is clouded. The impact of the most current estimate is dampened. Even if the samples themselves are independent, there should be some information contained in the estimates from prior years. Measurements taken in the remote past will have less bearing on the current year, but more recent estimates do shed light on what is happening currently. How to harness that information is the question posed in time series applications.

The U.S. Bureau of Labor Statistics and Statistics Netherlands were two national statistical institutes to begin using time series methodology for computing official statistics, see Tiller & Evans (2018) and Van der Brakel & Roels (2010). Small area time series were considered under specific autoregressive conditions by Pfefferman & Burck (1990). More general small area models were developed by Rao & Yu (1994), Datta, Lahiri, Maiti (2002), and You (2008). Comparisons between such combined time series cross-sectional models against state-space models were made by Balabay (2016) using the Dutch Travel Survey.

One not uncommon sampling strategy for repeated cross-sectional surveys involves retaining respondents for multiple measurements over time. These *panels* can significantly reduce data collection costs when purely independent samples are not necessary with every survey iteration. For example, Nielsen Television and Radio panels will keep individuals in the survey for periods of up

to two years. The CPS, administered by the U.S. Census Bureau for the Bureau of Labor Statistics (BLS), follows a design where a person's employment status is first measured for four consecutive months. They then exit the panel for eight months, and finally reenter for four more months. The Canadian Labor Force Survey maintains respondents for a continuous six-month period. Because the same individuals are in successive surveys, there would be a degree of autocorrelation present in the sample. Pooling datasets together achieves less of a benefit because of diminished effective sample size.

Depending on the amount of turnover, panel data can have even more dependence between successive samples than may be present in the underlying population. This property can improve estimates of change, however, as there will be less variation in the sample due to the selection of new respondents. Panel methods are also advantageous because they permit spatial correlations. Incorporating the correlation structures in both space and time can improve the reliability of small area estimates by utilizing more information. Indeed, under positive correlations, the covariance matrices of the space and time components can scale down terms in the mean squared error. The PANEL Procedure in SAS Software (SAS V9 (2020)) contains many variations on Two-Way Random effects models, including models where the random errors have some additional autocorrelation structure presumed. There is extensive literature in econometrics on this subject (e.g., Judge et al. (1985) or Hsiao (2014)), In contrast to small area models, the corresponding economical time series models will typically not account for sampling variances.

1.4 Model Selection and Diagnostics

The uncertainty involved with regard to model assumptions is often overlooked when interpreting statistical models. Decision makers should be fully aware of the subjective choices made during the model development process. There is the question of whether to assume normality versus some other non-normal distribution. The dependent variable could be transformed to make the distributional assumptions more reasonable. Should the model incorporate covariates as fixed effects vs. random effects? There is the question as to whether heteroskedasticity should be considered versus homoskedasticity. Have we selected the best linking function? Do we have an appropriate regression equation (linear, quadratic, etc.)? Finally, there is the issue of variable selection. These are all aspects which a comprehensive model selection process should address.

Once the appropriate model has been selected, it is incumbent to ensure that the model has a satisfactory goodness-of-it. This notion is different from selecting the *best* model out of a class of models: conceivably all the candidate models could fit the data poorly. Model diagnostics will include a variety of plots and tests to indicated a sound model fit. QQ-plots can verify distributional assumptions, while residual plots can help detect outliers, verify constant variation, and observe the independence of residual terms. Standard regression texts include sections on such diagnostics. See for instance Stapleton (1995), Rencher (1999), or Hosmer & Lemeshow (2000).

1.5 The Bootstrap Technique

The original bootstrap proposed by Efron (1979) was *non-parametric* in nature and consisted of resampling from the original dataset, although similar ideas had been presented in the statistical literature; see Hall (2003) for a review on the pre-history of bootstrap. It has since been widely adopted as an alternative to asymptotic procedures, especially with the availability of high-speed computing. It is often used to calculate variances and mean-squared errors. There is even the concept of the *double* bootstrap, where the procedure is repeated for bias reduction. The non-parametric bootstrap consists of taking M sub-samples from the dataset, with replacement, and recomputing all the statistics of interest in each of the M sub-samples, which are then combined. Many different variations of the bootstrap have since been produced. See the monograph by Efron and Tibshirani (1993) for more information.

The *parametric* bootstrap (e.g., see Li Lahiri (2010)), on the other hand, is more Monte Carlo in its application. An explicit probability model is assumed, and new data are generated on the basis of the parameter estimates from the original dataset. Once the data are completely regenerated, new parameter estimates are recalculated from the generated data. The process is repeated under different randomization seeds many times, and a distribution is built from these results. The bootstrap has been indispensable in estimating mean-squared errors, but we will see that it can have even greater utility for model building.

With regard to small area modeling, estimation of mean-squared estimates continued to be problematic. The foundations for applying the parametric bootstrap for small area applications were

outlined in Butar (1997). Butar & Lahiri (2003) proposed bootstrap samples to examine the accuracy of empirical Bayes for small area characteristics. Hall & Maiti (2006) introduced additional parametric bootstrap methodology that could be applied generally to more small area problems. Chatterjee, Lahiri, & Li (2008) proposed a parametric bootstrap to model a centered and scaled EBLUP, and noted the high degree of accuracy with simulation results demonstrating its benefits over competing analytical-based expressions. Saegusa et al. (2020) considered the parametric bootstrap to create confidence intervals for the multivariate Fay-Herriot model.

1.6 Outline of Thesis

In Chap. 2, we explore the relationship between the James-Stein estimator, the observed best predictor, and the empirical best predictor. We prove some results relating to the Bayes' risk (i.e., mean squared error) and frequentist risk. We define the generalized observed best predictive estimator, and compute several examples of best predictive estimators using data from the Small Area Income and Poverty Estimate (SAIPE) program administered by the U.S. Census Bureau.

In Chap. 3, we develop an empirical Bayes parametric bootstrap (EBPB) procedure for use in model selection and diagnostics. We describe the predictive distributions and leave-one-out Bayesian cross-validation approach. Some closed-form expressions for the predictive EBPB distribution for the Fay-Herriot model are calculated. These bootstrap predictive distributions are then

computed on the SAIPE data and used for variable selection. Finally, we define empirical Bayes residuals and compute them on the SAIPE data, as well.

In Chap. 4, we describe the Rao-Yu (RY) model, and develop expressions for the best predictive estimators for the model regression coefficients, along with numerical procedures for estimating the variance components and the autocorrelation parameter. We apply the model to unemployment estimates from the CPS and detail the OBP version of the RY model. We obtain a closed-form expression for $\hat{\rho}$ and build a simulation study to observed its performance. Finally, we posit the use of the empirical Bayes parametric bootstrap predictive distributions for determining between the Rao-Yu and Random Walk models.

In Chap. 5, we outline the areas for future research.

Chapter 2

Generalization of the Observed Best Predictor

2.1 Introduction

The best predictive estimator (BPE) was introduced by Jiang, Nguyen, & Rao (2011) as an alternative to maximum likelihood estimation for parameter coefficients β in linear regressions for area-level models. While $\hat{\beta}^{MLE}$ was plugged into the best predictor (BP) to obtain the best *linear unbiased* predictor (BLUP), the BPE of β (that is, $\tilde{\beta}^{BPE}$) could instead be plugged into the best predictor to obtain the *observed* best predictor (OBP). Under the correctly specified model, the BLUP is an optimal estimator. However, under misspecified regression models, the OBP could potentially outperform the BLUP. The derivation of the BPE is indeed *independent* of model assumptions. It attempts to define an estimator for β that can reduce the mean-squared prediction error (MSPE) under the expectation of the true underlying model. Therefore, for regression models that may be misspecified, the BPE and OBP could be used as a more robust alternative to standard small area estimation methods - in particular, it could be used in the Fay-Herriot model. Finally, the adjective *observed* derives from the fact that the optimization involved in obtaining the BPE does not minimize the *actual* MSPE, but the integrand within the expectation under the true model - this is the observed MSPE.

Observed best prediction is a general methodology that could be applied to variance com-

ponents, as well. Jiang, Nguyen, & Rao (2011) developed machinery for estimating the variance components in the Fay-Herriot model, and extended their results to the General Linear Mixed Model. In the former model, their estimator for the model variance did not have a closed-form solution, and relied upon numerical approximation, instead. In this case, the empirical best linear unbiased predictor (EBLUP) is also replaced by the OBP. Jiang, Nguyen, & Rao (2011) gave important theoretical results for the general linear mixed model, showing the MSPE of any empirical best predictor (EBP) is comprised of three components: a constant term followed by two variable terms. One term is minimized when the OBP is used, while the other term is minimized by the BLUP. They also showed their estimator was asymptotically equivalent to the EBLUP and under certain regularity conditions, is $\sqrt{m}$-*consistent*, where m is the number of domains.

Jiang, Nguyen, Rao (2015) developed the OBP for unit-level small area models originally defined in Battese et al. (1988). Chen, Jiang, & Nguyen (2015) defined an observed best predictor for count data. Benchmarking techniques for use with the OBP were investigated by Bandyopadhyay (2017). A more recent review of the OBP approach was given by Rao (2018).

Chapter Outline

(2.1) Introduction.

(2.2) Review of the Observed Best Predictor for the Fay-Herriot Model. Technical background is provided on the Fay-Herriot model and the observed best predictor.

(2.3) Relationship between the Observed Best Predictor and the James-Stein Estimator.

(2.4) Generalized Observed Best Predictor for the Fay-Herriot Model. We introduce the weighted best predictive estimator, leading to the generalized observed best predictor.

(2.5) Small Area Income and Poverty Estimates Data Analysis. The weighted best predictor and generalized observed best predictor are applied to the Small Area Income and Poverty Estimates (SAIPE) state-level data.

(2.6) Simulation Study. A simulation study is conducted using the SAIPE data, where the true area population means are known.

(2.7) Concluding Remarks.

2.2 Review of the Observed Best Predictor for the Fay-Herriot Model

Fay and Herriot (1979) considered the following model which serves as the basis for small area area-level modeling. Jiang, Rao, and Nguyen (2011) proposed a competing model, which did not assume an explicit relationship between the covariates and the expected value of the small area means. We shall refer to this model as the *Robust Fay-Herriot*, and we follow a similar notation for expectation between competing models.

Definition 2.1. Fay-Herriot Model.

Suppose $y_1, y_2, \ldots, y_m$ were assumed to be generated according to the following hierarchical model. For $i = 1, 2, \ldots, m$,

$$\begin{aligned} &\textbf{Level 1}. \quad y_i|\theta_i \overset{ind}{\sim} N(\theta_i, \psi_i), \\ &\textbf{Level 2}. \quad \theta_i \overset{iid}{\sim} N(x_i\beta, A), \end{aligned} \tag{2.1}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is a vector of known area-specific covariates, β is a vector of parameter coefficients, ψ_i is the known sampling variance for the i^{th} area, and A is of positive value.

Definition 2.2. Robust Fay-Herriot Model.

Now suppose $y_1, y_2, \ldots, y_m$ were actually generated according to this hierarchical model. For $i = 1, 2, \ldots, m$,

$$\begin{aligned} &\textbf{Level 1}. \quad y_i|\theta_i \overset{ind}{\sim} N(\theta_i, \psi_i), \\ &\textbf{Level 2}. \quad \theta_i \overset{ind}{\sim} N(\mu_i, A), \end{aligned} \tag{2.2}$$

where $\mu_i \neq x_i'\beta$ is the *true* mean value of θ. Following Jiang, Nguyen, & Rao (2011), we differentiate between the expectation from the assumed model and the true model by using the notation $E_M(\cdot)$ and $E(\cdot)$, respectively.

When β and A are known, the well-known solution to the Fay-Herriot model to estimate

the small area means θ_i is the best predictor of θ_i expressed as

$$\hat{\theta}_i^{BP} = B_i y_i + (1 - B_i) x_i' \beta,$$

where $B_i = B_i(A) = A/(A + \psi_i)$. When the hyperparameters are known, this is the *best* estimator in the sense of having the lowest variance and is obtained through $\hat{\theta}_i^{BP} = E(\theta_i|y)$.

When β is unknown and A is known with the errors e_i and ν_i normal, identically distributed, and mutually independent, then the best predictor $E(\theta_i|y)$ replaces the unknown β with $\hat{\beta}^{BLUE} = (\boldsymbol{X'V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'V}y$, the best linear unbiased estimator of β. This next estimator for θ_i is referred to as the BLUP, the best linear unbiased predictor, and remains optimal:

$$\hat{\theta}_i^{BLUP} = B_i y_i + (1 - B_i) x_i' \hat{\beta}_{BLUE}.$$

Finally, when the linking variance parameter A is unknown, a plug-in estimator can be used for A and the resulting predictor is referred to as the empirical best linear unbiased predictor, denoted by $\hat{\theta}_i^{EBLUP}$. Under this substitution, optimality properties are diminished. However, under REML estimation for A, the EBLUP was shown to be asymptotically optimal, as $m \to \infty$, see Jiang (2007). Whether likelihood-based estimates, MOM, or some other technique is used to estimate the variance components, the resulting formula can be expressed as:

$$\hat{\theta}_i^{EBLUP} = \hat{B}_i y_i + (1 - \hat{B}_i) x_i' \hat{\beta}_{BLUE},$$

this time with $\hat{B}_i = B_i(\hat{A}) = \hat{A}/(\hat{A} + \psi_i)$.

The optimality considerations for the BP, BLUP, and the EBLUP presume the model $Ey_i = x_i\beta$, where the unqualified expectation $E(\cdot)$ is taken to be with respect to the true underlying

population distribution. In most cases this is not known with certainty, as there may exist other latent variables in the model which have not been observed. When the objective is to predict the small area means $\theta_i = x_i\beta + \nu_i$, a reasonable approach is to consider another estimator that performs better when there is uncertainty with regards to full availability of prescient variables to the model. Another desirable property of such an estimator would be to have little loss in predictive power compared to the EBLUP when the model was actually correct. As discussed in the chapter introduction, Jiang, Nguyen, & Rao (2011) proposed the observed best predictor, $\tilde{\theta}_i^{OBP}$ given by

$$\tilde{\theta}_i^{OBP} = (1 - \tilde{B}_i)y_i + \tilde{B}_i x_i'\tilde{\beta}_{BPE},$$

now with $\tilde{B}_i = B(\tilde{A}^{BPE}) = \tilde{A}^{BPE}/(\tilde{A}^{BPE} + \psi_i)$. The hyperparameters β and A are estimated according to the so-called *best predictive estimator* and are computed to minimize the mean-squared error between the estimator and θ. The expectation E is taken under the *true* model configuration $\mu = [x_i' \; x_i^{+\prime}s]\begin{pmatrix}\beta \\ \alpha\end{pmatrix}$, where x_i^+ denotes a set of unknown variable information and α corresponds to the latent parameter coefficients which cannot be estimated. Note that this derivation presumes that the covariates $x_{i1}, x_{i2}, \ldots, x_{ip}$ are indeed correct, while in general some of the covariates may be extraneous.

In contrast to the likelihood-based approaches, the joint estimation vector $(\tilde{A}, \tilde{\beta})_{BPE}$ and the two stand-alone estimators $\tilde{\beta}^{BPE}$ and $\tilde{A}^{BPE}$ satisfy:

$$(\tilde{A}, \tilde{\beta})_{BPE} = \arg\min_{\beta, A} Q(A, \beta),$$
$$\tilde{A}_{BPE} = \arg\min_{A} Q(A),$$
$$\tilde{\beta}_{BPE} = \arg\min_{\beta} Q(\beta),$$

where $Q(.) = Q(A, \beta)$ is defined in Definition 2.5 such that MSPE$(\hat{\theta}) = E(\hat{\theta}_i^{BP} - \theta_i)^2 = EQ(A, \beta)$. The function Q is the *observed* mean square prediction error, and minimizing its value can reduce the *actual* MSPE. When either A or β are fixed we relax notation and use $Q(\beta)$, or $Q(A)$, respectively. It can be expedient to replace the joint minimum $(A, \beta)^{BPE}$ by fixing A first and obtaining the best predictor for β as a function of A. Finally, A is replaced with an estimator independent of β, resulting in the usage of $\tilde{A}_{BPE}$ and $\tilde{\beta}_{BPE} = \tilde{\beta}_{BPE}(\tilde{A}_{BPE})$.

2.3 Relationship between the Observed Best Predictor and the James-Stein Estimator

The following classical example was considered by James & Stein (1961), where they showcased an estimator that everywhere dominates the maximum likelihood estimator (MLE) in terms of overall mean-squared error across all observations, when the number of observations is greater than or equal to 3. This finding was rather counterintuitive to mainstream thinking, and the result is sometimes called *Stein's Phenomenon*. The problem was revisited by Efron & Morris (1973), where the astonishing outcome was shown to be an empirical Bayes solution, giving the estimator of James and Stein greater theoretical justification and understanding. We now consider this problem again in the context of the best predictive estimator to find a surprising new connection.

Example 2.3. Let $y_1, y_2, \ldots, y_m$ be generated according to the following hierarchical model.

For $i = 1, 2, \ldots, m$,

$$\begin{aligned} &\textbf{Level 1}. \quad y_i|\theta_i \overset{ind}{\sim} N(\theta_i, 1), \\ &\textbf{Level 2}. \quad \theta_i \overset{iid}{\sim} N(0, A). \end{aligned} \tag{2.3}$$

We are interested in finding estimators for each of the θ_i's. This problem can be viewed in the context of the Fay-Herriot model (Definition 2.1), with $x_i\beta = 0$ and $\psi_i = 1$, for i=1, . . . , m.

Case 1. A is known (BP solution). Letting $\tau = \frac{1}{1+A}$, the theory of mixed linear models gives the best predictor as the posterior mean, $\hat{\theta}_i^{BP} = E_M(\theta_i|y)$, where the expectation E_M is with respect to the random variable θ_i conditional on $y = (y_1, y_2, \ldots, y_m)'$ based on the model (2.3).

Proposition 2.4. Best Predictor for θ when A is known and $\tau = \frac{1}{1+A}$,

$$\hat{\theta}_i^{BP} = (1 - \tau)y_i.$$

Proof. Since θ_i and $y_i|\theta_i$ are both normal random variables, the normal theory asserts that y_i is also unconditionally normal. Note that the expectation of y with respect to model (**??**) is $E_M y_i = E_M\{E_M(y_i|\theta_i)\} = E_M\theta_i = 0$. Using the law of total variance, the variance of y_i, also with respect to model (2.3), is $V_M(y_i) = E_M\{V_M(y_i|\theta_i)\} + V_M\{E_M(y_i|\theta_i)\} = E(1) + V_M(\theta_i) = 1 + A$. Thus, $y_i \sim N(0, 1 + A)$. Now let f(·) be a generic function denoting the probability density function of its argument (strictly speaking, this might be denoted by $f_M(\cdot)$, but we relax this notation and use $f(\cdot)$, instead). From the Bayes' theorem, the conditional density function of θ_i given y_i, denoted as

$f(\theta_i|y_i)$, can be expressed as

$$
\begin{aligned}
f(\theta_i|y_i) &= \frac{f(y_i|\theta_i) \times f(\theta_i)}{f(y_i)} \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[(y_i-\theta_i)^2\right\} \times \frac{1}{\sqrt{2\pi}A} \exp\left\{-\frac{\theta_i^2}{2A}\right\} \times \sqrt{2\pi}(A+1) \exp\left\{\frac{y_i^2}{2(A+1)}\right\} \\
&= \frac{A+1}{\sqrt{2\pi}A} \exp\left\{-\frac{1}{2}[y_i^2 - 2y_i\theta_i + \theta_i^2 + \theta_i^2/A - \frac{y_i^2}{A+1}]\right\} \qquad (2.4) \\
&= \frac{A+1}{\sqrt{2\pi}A} \exp\left\{-\frac{1}{2}[\theta_i^2(1+\frac{1}{A}) - 2\theta_i y_i + y_i^2(1-\frac{1}{1+A})]\right\} \\
&= \frac{1+A^{-1}}{\sqrt{2\pi}} \exp\left\{-\frac{1+A^{-1}}{2}[\theta_i^2 - \frac{2\theta_i y_i}{1+A^{-1}} + y_i^2(\frac{A}{A+1})^2]\right\}.
\end{aligned}
$$

We recognize this distribution as that of a normal random variable with conditional mean $\frac{A}{A+1}y_i$ and variance $\frac{A}{A+1}$: $\theta_i|y_i \overset{ind}{\sim} N\left(\frac{A}{A+1}y_i, \frac{A}{1+A}\right)$. Moreover, note that $E_M(\theta_i|y_i) = (\frac{A}{A+1})y_i = (1-\frac{1}{1+A})y_i = (1-\tau)y_i$. Therefore, the best predictor when A is known is $\hat{\theta}_i^{BP} = (1-\tau)y_i = (1-\frac{1}{A+1})y_i$. □

Let us consider the extreme cases for values of A. When A=0, then θ_i has zero variance and is considered a constant identically equal to zero, $\theta_i \equiv 0$. Consequently, the best predictor also becomes just zero, $\hat{\theta}_i^{BP} = 0$ and is no longer a function of the data when A=0. Conversely, when A is very large, then there is less auxiliary information about θ_i being equivalent to zero. If our prior information about θ_i being close to zero is unreliable, then it's not sensible to deviate much from the original data y_i, and so it turns out that $\hat{\theta}_i^{BP} \approx y_i$ when A is large. Otherwise, moderate values of A induce the shrinkage effect on y_i that pushes the best predictor closer to to the presumed mean $x_i\beta$ (zero in this example).

Example 2.3 *(continued)*. **Case 2. A is unknown (EBP solution)**. The basic approach in small area estimation is to obtain an estimator for any nuisance parameters and to plug them into the best predictor, $\hat{\theta}_i^{BP}$. It is helpful to emphasize the dependency of the best predictor as a function of A, besides being a function of y: $\hat{\theta}_i^{BP} = \hat{\theta}_i^{BP}$(y, A) = $\hat{\theta}_i^{BP}$(A). Now let $\hat{A}_{MLE}$ denote the maximum likelihood estimator for A. We proceed to utilize $\hat{\theta}_i^{BP}(\hat{A}_{MLE}) = \hat{\theta}_i^{EBP} = \hat{\theta}_i^{EBP}(y, \hat{A}_{MLE})$ as the most viable plug-in estimator. Let $f(y_i; A)$ denote the marginal probability density of y_i with parameter A, and $L(y; A) = L(A)$ is the likelihood function of A with respect to the distribution of y.

Lemma 2.5. Calculation of $\hat{A}_{MLE}$.

$$\hat{A}_{MLE} = \frac{||y||^2}{m} - 1.$$

Proof.

$$\begin{aligned} log(L(A)) = log(\prod_{i=1}^{m} f(y_i; A)) = -\frac{m}{2} log(2\pi) - \frac{m}{2} log(1+A) - \frac{1}{2} \sum_{i=1}^{m} \frac{y_i^2}{A+1}, \\ \frac{\partial}{\partial A} log(L(A)) = -\frac{m}{2} \frac{1}{1+A} + \frac{1}{2} \frac{||y||^2}{(A+1)^2} = 0 \implies A = \frac{||y||^2}{m} - 1. \end{aligned} \tag{2.5}$$

□

Corollary 2.6. The empirical best predictor of θ_i, $\hat{\theta}_i^{EBP}$ is given by:

$$\hat{\theta}_i^{EBP} = \left(1 - \frac{m}{||y||^2}\right) y_i.$$

Proof. Taking $\hat{\theta}_i^{BP}(A)$ from Proposition 2.4 and plugging in the MLE for A obtained in Lemma 2.5 yields $\hat{\theta}_i^{EBP} = \hat{\theta}_i^{BP}(\hat{A}_{MLE}) = (1 - \tau(\hat{A}_{MLE}))y_i = (1 - \frac{1}{1+\hat{A}_{MLE}})y_i = (1 - \frac{m}{||y||^2})y_i$. □

Remark on Notation. When the variance parameter is unknown, the plug-in estimator based on the best predictor is known as the *empirical* best predictor, and *BP* becomes *EBP*. The estimator is no longer *best*, but under normality and $\hat{A}^{REML}$ as the plug-in, the EBP is asymptotically optimal (e.g., see Jiang (2007)). When the parameter coefficient vector β is also unknown, the EBP is known as the EBLUP, or empirical best linear unbiased predictor. The adjectives "linear" and "unbiased" refer to the estimator for β, which usually takes the form of the least squares estimator, and satisfies both conditions.

Example 2.3 (*continued*). **Case 2. A is unknown (OBP solution)**. Now suppose that the model specified in (1) was not correct, and that $X_i\beta = \mu_i \neq 0$. Then the true underlying model could be given by:

$$\begin{aligned} &\textbf{Level 1}. \quad y_i|\theta_i \overset{ind}{\sim} N(\theta_i, 1), \\ &\textbf{Level 2}. \quad \theta_i \overset{ind}{\sim} N(\mu_i, A). \end{aligned} \tag{2.6}$$

We denote expectation under the true distribution (2.6) as the unqualified operator $E(\cdot)$. We will see that the same solution to Corollary 2.6 can be obtained through consideration of the mean squared prediction error, or MSPE. This is not the case, in general. Note that the MSPE takes expectation with respect to (2.6) under *all* random variables in the random vectors θ and y.

Definition 2.7. (*Observed MSPE*). Any function $Q(||y||, \tau)$ of the data y and the unknown pa-

rameters satisfying the equality below may be called an *observed* mean squared prediction error.

$$MSPE(\hat{\theta}) = \sum_{i=1}^{m} E(|\hat{\theta}_i - \theta_i|^2) = E(Q(||y||, \tau)]$$

Finding the nuisance parameters under the true expectation which minimize such functions Q is the key to OBP theory. The expectation above and what follows are with respect to the true underlying distribution (2.2). First, we can observe that $E(y_i - \theta_i)^2 = E\left\{E(y_i - \theta_i)^2|\theta)\right\} = E(1) = 1$. Next, $Ey_i^2 = E\left\{E(y^2|\theta)\right\} = E(\theta_i^2 + 1) = \mu_i^2 + A + 1$. Finally, we note that $E(y_i\theta_i) = E\left\{E(y_i\theta_i|\theta)\right\} = E\theta_i^2 = \mu_i^2 + A$. Following Jiang, Rao, and Nguyen (2011), we compute the observed MSPE for $\hat{\theta}_i^{BP}$ yielding

$$\begin{aligned}
MSPE(\hat{\theta}_i^{BP}) &= \sum_{i=1}^{m} E(|(1-\tau)y_i - \theta_i|^2) \\
&= E\sum_{i=1}^{m}(y_i - \theta_i - \tau y_i)^2 \\
&= E\sum_{i=1}^{m}\left\{(y_i - \theta_i)^2 - 2\tau y_i(y_i - \theta_i) + \tau^2 y_i^2\right\} \\
&= E\sum_{i=1}^{m}\left\{(y_i - \theta_i)^2 - 2\tau(A + \mu_i^2 + 1 - \mu_i^2 - A) + \tau^2 y_i^2\right\} \\
&= E\left\{\sum_{i=1}^{m}\left[1 - 2\tau + \tau^2 y_i^2\right]\right\}.
\end{aligned} \tag{2.7}$$

It should be noted that we are free to simplify the function Q as much as possible, so long as it remains equal in expectation to the MSPE. With A and thus $\tau = \tau(A)$ unknown, we can derive a suitable estimator of A (or equivalently of $\tau = 1/(1 + A)$). The function Q=$Q(||y||; \tau)$ may be simplified:

Proposition 2.8. $Q(||y||;\tau)$ is minimized by $\tau = \frac{m}{||y||^2}$.

Proof.

$$\begin{aligned} Q(||y||;\tau) &= \sum_{i=1}^{m}\left[\tau^2 y_i^2 - 2\tau + 1\right] = \tau^2||y||^2 - 2m\tau + m \\ &= \tau^2||y||^2 - 2\tau||y||\frac{m}{||y||} + \left(\frac{m}{||y||}\right)^2 + m - \left(\frac{m}{||y||}\right)^2 \\ &= \left(\tau||y|| - \frac{m}{||y||}\right)^2 + m\left(1 - \frac{m}{||y||^2}\right) \\ &\geq m\left(1 - \frac{m}{||y||^2}\right). \end{aligned}$$

We see that minimum of $Q(||y||;\tau)$, (with respect to τ), is attained at $\tilde{\tau}_{BPE} = \frac{m}{||y||^2}$. □

The variance parameter which minimizes the mean-square prediction error is depicted as $\tilde{\tau}_{BPE}$ (it is also convenient to use the notation $\tilde{A}_{BPE}$, whence $\tilde{\tau}_{BPE} = \frac{1}{1+\tilde{A}_{BPE}}$). We also use the following notation for the observed best predictor as $\tilde{\theta}_i^{OBP} = \hat{\theta}_i^{BP}(\tilde{\tau}^{BPE})$. The consequence of Proposition 2.6 is that

$$E[Q(||y||,\tau)] \geq E[Q(||y||,\tilde{\tau}_{BPE})],\ \forall\tau \in (0,1).$$

Ideally, we would like to verify that $E(\hat{\tau}_{BPE})$ is the τ that minimizes $MSPE(\hat{\theta}^{OBP})$. Instead, we show show that $\tilde{\tau}^{BPE} \xrightarrow{P} \tau^*$.

Proposition 2.9. The best predictive estimator of τ, $\tilde{\tau}^{BPE}$, converges in probability to the minimizer

τ^* of the mean-squared prediction error of the best predictor: $\text{MSPE}(\hat{\theta}^{BP})$,

$$\tilde{\tau}^{BPE} \xrightarrow{P} \tau^*/$$

Proof.

$$\begin{aligned} Pr\left\{\left|\tilde{\tau}^{BPE} - \tau^*\right| > \epsilon\right\} &= Pr\left\{\left|\frac{m}{||y||^2} - \frac{m}{\sum_{i=1}^m(\mu_i^2 + A + 1)}\right| > \epsilon\right\} \\ &= Pr\left\{\left|\frac{1}{\frac{1}{m}||y||^2} - \frac{1}{\frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1)}\right| > \epsilon\right\} \\ &= Pr\left\{\left|\frac{\frac{1}{m}||y||^2 - \frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1)}{(\frac{1}{m}||y||^2)(\frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1))}\right| > \epsilon\right\}. \end{aligned} \tag{2.8}$$

$$\begin{aligned} Pr\left\{\left|\tilde{\tau}^{BPE} - \tau^*\right| > \epsilon\right\} &= Pr\left\{\left|\frac{\frac{1}{m}||y||^2 - \frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1)}{(\frac{1}{m}||y||^2)(\frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1))}\right| > \epsilon\right\} \\ &\leq Pr_\Omega\left\{\left|\frac{\frac{1}{m}||y||^2 - \frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1)}{C_m}\right| > \epsilon\right\} \\ &+ Pr_{\Omega^C}\left\{\left|\frac{\frac{1}{m}||y||^2 - \frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1)}{(\frac{1}{m}||y||^2)(\frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1))}\right| > \epsilon\right\} \\ &= Pr_\Omega\left\{\left|\frac{1}{m}||y||^2 - \frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1)\right| > C_m\epsilon\right\} \\ &+ Pr_{\Omega^C}\left\{\left|\frac{\frac{1}{m}||y||^2 - \frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1)}{(\frac{1}{m}||y||^2)(\frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1))}\right| > \epsilon\right\} \\ &\leq E\left\{\left|\frac{1}{m}||y||^2 - \frac{1}{m}\sum_{i=1}^m(\mu_i^2 + A + 1)\right|\right\} + \delta \\ &= 0 + \delta. \end{aligned} \tag{2.9}$$

Where we have used Markov's Inequality in the last inequality of (2.8) and because under the true model, $E||y||^2 = \sum_{i=1}^m(\mu_i^2 + A + 1)$. Since δ was arbitrary we have shown the result. □

The OBP and EBP are then *both* given by

$$\tilde{\theta}_i^{OBP} = \left(1 - \frac{m}{||y||^2}\right) y_i.$$

These compare very closely with the solution given by James and Stein in 1961:

$$\hat{\theta}_i^{JS} = \left(1 - \frac{m-2}{||y||^2}\right) y_i.$$

It is instructive to examine the frequentist risk of the OBP and compare it with that of the James-Stein estimator. We will need the following lemma and corollary, a proof of which can be found in Rao and Molina (2013) and is due to Stein (1961).

Lemma 2.10 (Stein's Lemma). Let $Z \sim N(\mu, 1)$. Then E[h(Z)(Z-μ)] = $E[\frac{\partial h(Z)}{\partial Z}]$ = E[h'(Z)], provided all expectations exists and,

$$\lim_{z \to \infty} h(z) \exp\left\{-\tfrac{1}{2}(z-\mu)^2\right\} = 0.$$

Corollary 2.11. Now let Z = $(Z_1, Z_2, \ldots, Z_m)'$, with $Z_i \overset{ind}{\sim} N(\mu_i, 1)$. Under the same conditions as Lemma 2.1, then E[h(Z)($Z_i - \mu_i$)] = $E[\frac{\partial h(Z)}{\partial Z_i}]$.

Letting $h(y) = \frac{y_1}{||y||^2}$, it is evident that the conditions of Lemma 2.10 hold since h is bounded by zero and one: $0 \leq |h(y)| \leq 1$ and $\lim_{z \to \infty} \exp\left\{-\frac{1}{2}(z-\mu)^2\right\} = 0$. The conditional frequentist risk given θ for the small area means for $\theta = (\theta_1, \theta_2, \ldots, \theta_m)'$ is denoted by $R(\tilde{\theta}^{OBP}, \theta)$ and given by

the sum of conditional expectations $E_\theta(\cdot) = E(\cdot|\theta)$:

$$\begin{aligned}
R(\tilde{\theta}^{OBP}, \theta) &= \sum_{i=1}^{m} R(\tilde{\theta}_i^{OBP}, \theta_i) \\
&= \sum_{i=1}^{m} E_\theta \left[(\tilde{\theta}_i^{OBP} - \theta_i)^2 \right] \\
&= \sum_{i=1}^{m} E_\theta \left((1 - \frac{m}{||y||^2}) y_i - \theta_i \right)^2 \\
&= \sum_{i=1}^{m} E_\theta (y_i - \theta_i)^2 - 2m \sum_{i=1}^{m} E_\theta \left\{ (y_i - \theta_i) \frac{y_i}{||y||^2} \right\} + \sum_{i=1}^{m} E_\theta \frac{y_i^2 m^2}{||y||^4} \\
&= m - 2m E_\theta \; he\{ \sum_{i=1}^{m} \frac{\partial}{\partial y_i} (\frac{y_i}{||y||^2}) \Big\} + m^2 E_\theta \frac{1}{||y||^2} \\
&= m - 2m E_\theta \left\{ \sum_{i=1}^{m} \frac{||y||^2 - 2y_i^2}{||y||^4}) \right\} + m^2 E_\theta \frac{1}{||y||^2} \\
&= m - 2m(m-2) E_\theta \frac{1}{||y||^2} + m^2 E_\theta \frac{1}{||y||^2} \\
&= m - m(m-4) E_\theta \frac{1}{||y||^2}.
\end{aligned} \tag{2.10}$$

We have therefore found the frequentist risk for the observed best predictor. A similar calculation is followed to obtain the freuqentist risk of the James-Stein estimator. We summarize these results as:

Proposition 2.12. The frequentist risk of the OBP(EBP), $\tilde{\theta}^{OBP}$, and the James-Stein estimator, $\hat{\theta}^{JS}$, are given by:

$$\begin{aligned}
R(\tilde{\theta}^{OBP}, \theta) &= m - m(m-4) E_\theta \frac{1}{||y||^2}, \\
R(\hat{\theta}^{JS}, \theta) &= m - m(m-2) E_\theta \frac{1}{||y||^2}.
\end{aligned} \tag{2.11}$$

Under the conditional expectation of the frequentist risk given θ we have the following distribution for the squared norm of y given θ, $||y||^2|\theta$,

$$||y||^2|\theta \sim \chi^2(m, ||\theta||^2).$$

which is the non-central chi-squared distribution with m degrees of freedom and non-centrality parameter $|||\theta||^2$. Unlike the standard chi-squared distribution, the property $E(\frac{1}{U}) \neq \frac{1}{EU}$ for $U \sim \chi^2(m)$ does not generally hold. However, since the function $f(t) = \frac{1}{t}$ is convex on the interval (0,∞), Jensen's inequality shows, $E_\theta \frac{1}{||y||^2} \geq \frac{1}{E_\theta(||y||^2)} = \frac{1}{k} \geq 0$ and so is strictly positive. The frequentist risk from $\hat{\theta}^{MLE}$ = y (just the data itself) is $R(y, \theta) = \sum_{i=1}^{m} E_\theta(y_i - \theta)^2$ = m. Then the OBP dominates the direct estimates (i.e., has a lower risk for all possible values of $\theta \in R^m$) whenever $m \geq 5$. In contrast, the *celebrated* James-Stein estimator has risk $R(\hat{\theta}^{JS}, \theta) = m - m(m-2)E_\theta \frac{1}{||y||^2}$, and so it dominates y for $m \geq 3$.

Despite the James-Stein estimator $\hat{\theta}^{JS}$ having a lower overall risk than either $\tilde{\theta}^{OBP}$ or $\tilde{\theta}^{EBP}$, it is not widely used in computing official statistics. Ignoring the fact that no additional information comes from the auxiliary variables (i.e., $x_i\beta = 0$), we have to acknowledge that the sampling variance is equal across all areas. In practice, sample sizes will vary across areas, leaving some with reliability issues while others will maintain precision direct estimates. This is the fundamental situation which small area estimation addresses.

While having a uniform value for τ across all areas may be optimal for reducing *overall* risk across *all* areas, the value τ may fail to produce minimal risk in *any* individual small area. When

all sampling variances are not equal, there will be some domains which do not necessitate much modeling by virtue of higher sample sizes and greater sample dispersion. Conversely, some domains will have lower reliability, requiring a greater dependency on the modeled value (which is zero in the James-Stein case, and $x_i\beta$ in the general case). This is a bit problematic, because it is important to obtain the best possible prediction for every geographical area or domain. In some applications, there will be important decisions regarding allocation of resources as a direct consequence of the modeling approach. Thus, the James-Stein estimator is not a practical tool for real-world situations where it is preferable to have area-specific modeling in which $\tau_i's$ are not all equal to ensure each area has a reasonably accurate estimate.

Proposition 2.13. The frequentist risk of the James-Stein estimator is lower than that of the Observed Best Predictor when $m \geq 5$:

$$R(\hat{\theta}^{JS}, \theta) < R(\tilde{\theta}^{OBP}, \theta).$$

Proof. $R(\hat{\theta}^{JS}, \theta) = m - m(m-2)E\frac{1}{||y||^2} < m - m(m-4)E\frac{1}{||y||^2} = R(\tilde{\theta}^{OBP}, \theta)$ □

Corollary 2.14. The mean squared prediction error of the James-Stein estimator is lower than that of the Observed Best Predictor when $m \geq 5$:

$$MSPE(\hat{\theta}^{JS}) < MSPE(\tilde{\theta}^{OBP}).$$

Proof. Applying the expectation under (2.2) to the inequality in Proposition 2.13 yields MSPE($\hat{\theta}^{JS}$)

$= E\left\{R(\hat{\theta}^{JS}, \theta)\right\} \leq E\left\{R(\tilde{\theta}^{OBP}, \theta)\right\} = MSPE(\tilde{\theta}^{OBP}, \theta).$ □

Historical Notes.

The estimator $\hat{\theta}^{JS}$ is remarkable because it was shown that for m ≥ 3 that $\hat{\theta}_i^{JS}$ has a lower risk under umfer squared error loss than the MLE (just y_i, in this case) for all possible values of θ_i. Under these simplified conditions, the OBP lies witin the same class of James-Stein estimators defined by $\hat{\theta}_i^{JS} = \left(1 - \frac{a}{||y||^2}\right) y_i$ for some constant a.

The term $E_\theta \frac{1}{||y||^m}$ is cumbersome because it is not necessarily the reciprocal of the expectation of a non-central chi-squared random variable (the property $E\frac{1}{U^2} = \frac{1}{EU^2}$ does hold for standard chi-squared random variables U^2). Stein (1966) proposed the following second-order approximation for the frequentist risk of JS-type estimators using Taylor series:

$$E_\theta(\sum_{i=1}^{m}(\hat{\theta}_i - \theta_i)^2 \approx \frac{1}{a - 2\lambda}\left\{1 + \frac{1}{(a + 2\lambda)^2}\right\},$$

where the conditional expectation E_θ is taken under the presumed model $\lambda = \frac{||y||}{2}$ and with respect to the distribution of the data $y = (y_1, y_2, \ldots, y_m)$'. Bhattacharya (1966) explored the risk under general weights: $(\hat{\theta}_i - \theta_i)'\boldsymbol{W}(\hat{\theta}_i - \theta_i)$, where $\boldsymbol{W}$ may not necessarily be a diagonal matrix. Egerton and Laycock (1982) found closed form solutions for the frequentist risk under various specific conditions.

2.4 Generalized Observed Best Prediction for the Fay-Herriot Model

We will now investigate a new generalized version of the observed best predictor, which we have seen to be closely related to the James-Stein estimator. Having control over the importance weights will provide practitioners with greater flexibility when working with their unknown data than is permitted under the standard observed best predictor. Instead of minimizing the sum of mean squared errors $\sum_{i=1}^{m} E(\hat{\theta}_i^{BP} - \theta_i)^2$ we investigate a weighted mean squared error (WMSE) of some predictor of $\bar{\theta}$ of θ, given by,

Definition 2.15. Weighted Mean Squared Prediction Error.

$$WMSnoE(\bar{\theta}) = \sum_{i=1}^{m} w_i E(\bar{\theta}_i - \theta_i)^2,$$

where $w_i \geq 0 \; \forall i = 1, 2, \ldots m$. The expectation E denotes expectation under the true underlying population distribution, which may differ from the assumed distribution. We will assume the importance weights w_i may possibly be a function of A, $w_i = w_i(A)$ but not of β. There will be compelling reasons for different choices of weights for the original MSE. The WMSE is associated with a function Q_W, whose inputs correspond to the unknown nuisance parameters: β and A.

Definition 2.16. Observed Weighted Mean Squared Prediction Error.

Given an estimator $\hat{\theta}$ and a weighting matrix W, the *observed weighted mean squared error*, $Q_W(\beta, A)$, is any measurable function whose expectation is equivalent to the weighted mean

squared error WMSE

$$EQ_W(A,\beta) = WMSE(\hat{\theta}).$$

Note that Q_W is also dependent on the choice of estimator for θ, $\bar{\theta}$. The basic form of the estimator should be developed up to this point, with the only unknowns corresponding to estimation of nuisance parameters. In small area estimation, the best predictor $\hat{\theta}^{BP}$ is used as the baseline estimator for θ.

Definition 2.17. Weighted Best Predictive Estimator.

Given a vector ϕ of unknown parameters, and a weighted observed MSPE $Q_W(\phi)$, then the *weighted* best predictive estimator of ϕ, denoted by $\tilde{\phi}^{WBPE}$ is defined as

$$\tilde{\phi}^{WBPE} = \arg\min_{\phi} Q_W(\phi).$$

For the Fay-Herriot model, there are three choices for ϕ: $\phi = A$, $\phi = \beta$, or $\phi = (\beta, A)$. We are now in a position to define the generalized observed best predictor, $\tilde{\theta}^{GBP}$, as,

Definition 2.18. Generalized Observed Best Predictor,

$$\tilde{\theta}^{GBP} = \hat{\theta}^{BP}(\tilde{\phi}^{WBPE}).$$

Fay-Herriot Model Case when A is known.

We can let $B_i(A) \equiv B_i = A/[A+\psi_i]$ also be known, and possibly have $w_i \equiv w_i(A)$ so that

$$\begin{aligned} WMSE(\hat{\boldsymbol{\theta}}^{BP}) &= \sum_{i=1}^{m} w_i E(\hat{\theta}_i^B - \theta_i)^2 \\ &= \sum_{i=1}^{m} w_i E(B_i y_i - \theta_i + x_i'\beta(1-B_i))^2 \\ &= \sum_{i=1}^{m} w_i E(B_i y_i - \theta_i)^2 - 2\sum_{i=1}^{m} w_i x_i'\beta(1-B_i)E(B_i y_i - \theta_i) + \sum_{i=1}^{m} w_i (x_i'\beta)^2(1-B_i)^2. \end{aligned} \tag{2.12}$$

The first term in the last identity does not depend on β, and for the second term we have that $E(B_i y_i - \theta_i) = (1-B_i)E(y_i)$. So that,

$$WMSE(\hat{\boldsymbol{\theta}}^{BP}) = E\left\{ \sum_{i=1}^{m} w_i (B_i y_i - \theta_i)^2 - 2\sum_{i=1}^{m} w_i y_i x_i'\beta(1-B_i)^2 + \sum_{i=1}^{m} w_i (x_i'\beta)^2(1-B_i)^2 \right\}. \tag{2.13}$$

It is enough to consider the expression above since its expectation will still be equivalent to the WMSE. Rather than minimizing the expected value E, which is unknown, minimizing under the integrand yields the following objective function of β:

$$\begin{aligned} Q(\beta) &= \sum_{i=1}^{m} w_i(A)\left[(1-B_i(A))^2 (x_i'\beta)^2\right] - 2\sum_{i=1}^{m} w_i(A)\left[(1-B_i(A))^2 x_i'\beta y_i\right] \quad (*) \\ &= \beta' \boldsymbol{X}'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}\boldsymbol{X}\beta - 2y'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}\boldsymbol{X}\beta, \end{aligned} \tag{2.14}$$

where $\boldsymbol{X} = (x_i')_{1\le i\le m}$, y $= (y_i)_{1\le i\le m}$, $\boldsymbol{\Gamma} = diag(1-B_i)$ and $\boldsymbol{W} = diag(w_i)_{1\le i\le m}$. Then it is easy to

see using calculus that β is maximized by:

$$\frac{d}{d\beta}Q(\beta) = 2\beta' \boldsymbol{X}'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}\boldsymbol{X} - 2y'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}\boldsymbol{X} = 0 \implies \beta' = y'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}\boldsymbol{X})^{-1}.$$

This leads to the following definition of the weighted best predictive estimator for β:

$$\tilde{\beta}_W^{BPE} = (\boldsymbol{X}'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}y \tag{2.15}$$

The optimal BPE estimator is weighted by the diagonal matrix W = W(A) may also be expressed as follows:

$$\hat{\beta}_W^{BPE} = \hat{\beta}_W^{BPE}(A) = \left[\sum_{i=1}^{m} w_i(A)(1 - B_i(A))^2 x_i x_i'\right]^{-1} \sum_{i=1}^{m} w_i(A)(1 - B_i(A))^2 x_i y_i. \tag{2.16}$$

We are thus able to produce custom BPEs of β by choosing the w_i differently. If one chooses $w_i = 1$, or the weighting matrix W is just the identity, W = I_m, we obtain the standard BPE, $\hat{\beta}_{I_m}^{BPE}$ = $\hat{\beta}^{BPE}$. But there are other reasonable estimators, as well. It should be pointed out that the optimality properties of the BPE will be based on what areas are deemed by the practitioner to be more important, and that subjectivity should be considered under direct comparisons to other estimators. The ability to adjust the importance weights gives some flexibility - essentially we are looking at the BPE under a more general loss function. It's prudent to compare the weights applied to the small areas in each of the estimators of β.

Table 2.19. *Estimation Weights by Area, for different estimators of β*

Estimation Weights	
$\hat{\beta}$	Area Weight
$\hat{\beta}^{MLE}$	$\frac{1}{A+\psi_i}$
$\tilde{\beta}^{BPE}$	$\frac{\psi^2}{(A+\psi_i)^2}$
$\tilde{\beta}^{BPE}_W$	$\frac{w_i\psi^2}{(A+\psi_i)^2}$

Example 2.20: Weighted BPE of β (Inverse Variance Weights)

It is interesting to observe that:

$$\hat{\theta}_i^{BP}|\theta \overset{ind}{\sim} N\left(B_i\theta_i + (1-B_i)x_i'\beta,\ B_i^2\psi_i\right).$$

We take the time to verify this. We can derive the covariance between $\hat{\theta}_i^{BP}$ and θ_i by noting that

$$E(\theta_i\hat{\theta}_i^{BP}) = E\left[(x_i'\beta+v_i)(B_iy_i+(1-B_i)x_i'\beta)\right] = B_i(x_i\beta)^2 + B_iA + 0 + (1-B_i)(x_i'\beta)^2 = (x_i\beta)^2 + B_iA$$

$$\implies Cov(\theta_i, \hat{\theta}_i^{BP}) = (x_i'\beta)^2 + B_iA - (x_i'\beta)^2 = B_iA \implies \rho = \frac{B_iA}{B_i\sqrt{A(A+\psi_i)}} = \frac{A}{\sqrt{A(A+\psi_i)}} = \sqrt{B_i}.$$

Now we can express the joint distribution of $\hat{\theta}_i$ and θ_i as,

$$
\begin{aligned}
f(\hat{\theta}_i, \theta_i) &= \frac{1}{2\pi}\sqrt{\frac{1}{AB_i^2(\psi_i + A)}\frac{\psi_i + A}{\psi}} \\
&\times \exp\left\{-\frac{1}{2(\frac{\psi_i}{\psi_i + A})}\left(\frac{(\hat{\theta}_i^{BP} - x_i'\beta)^2}{B_i^2(\psi_i + A)} - 2B_iA\frac{(\hat{\theta}_i^{BP} - x_i'\beta)(\theta_i - x_i'\beta)}{B_i^2A(\psi_i + A)} + \frac{(\theta_i - x_i'\beta)^2}{A}\right)\right\} \\
&= \frac{1}{2\pi}\frac{1}{B_i\sqrt{\psi_i A}} \\
&\times \exp\left\{-\frac{1}{2}\left(\frac{(\hat{\theta}_i^{BP} - x_i'\beta)^2}{B_i^2\psi_i} - 2\frac{(\hat{\theta}_i^{BP} - x_i'\beta)(\theta_i - x_i'\beta)}{B_i\psi} + \frac{(\theta_i - x_i'\beta)^2}{B_i\psi_i}\right)\right\}.
\end{aligned}
\tag{2.17}
$$

Recalling the distribution of $\hat{\theta}_i^{BP} = \theta_i$ is $f(\theta_i) = \sqrt{\frac{1}{2\pi}}\sqrt{\frac{1}{A}}\exp -\frac{1}{2}\left\{\frac{(\theta_i - x_i'\beta)^2}{A}\right\}$, we compute the conditional distribution as the ratio of the density function of θ_i divided by joint density of θ_i and $\hat{\theta}_i$,

$$
\begin{aligned}
f(\hat{\theta}_i, \theta_i)/f(\theta_i) &= \sqrt{\frac{1}{2\pi}}\frac{1}{B_i\sqrt{\psi_i}} \\
&\times \exp\left\{-\frac{1}{2}\left(\frac{(\hat{\theta}_i^{BP} - x_i'\beta)^2}{B_i^2\psi_i} - 2\frac{(\hat{\theta}_i^{BP} - x_i'\beta)(\theta_i - x_i'\beta)}{B_i\psi} + \frac{(\theta_i - x_i'\beta)^2}{B_i\psi_i} - \frac{(\theta_i - x_i'\beta)^2}{A}\right)\right\} \\
&= \sqrt{\frac{1}{2\pi}}\frac{1}{B_i\sqrt{\psi_i}}\exp\left\{-\frac{1}{2}\left(\frac{(\hat{\theta}_i^{BP} - x_i'\beta)^2}{B_i^2\psi_i} - 2\frac{(\hat{\theta}_i^{BP} - x_i'\beta)(\theta_i - x_i'\beta)}{B_i\psi} + \frac{(\theta_i - x_i'\beta)^2}{\psi_i}\right)\right\} \\
&= \sqrt{\frac{1}{2\pi}}\frac{1}{B_i\sqrt{\psi_i}}\exp\left\{-\frac{1}{2B_i^2\psi_i}(\hat{\theta}_i^{BP} - x_i'\beta - B_i(\theta_i - x_i'\beta))^2\right\} \\
&= \sqrt{\frac{1}{2\pi}}\frac{1}{B_i\sqrt{\psi_i}}\exp\left\{-\frac{1}{2B_i^2\psi_i}(\hat{\theta}_i^{BP} - B_i\theta_i - (1 - B_i)x_i'\beta)^2\right\} \\
&\implies \hat{\theta}_i^{BP}|\theta_i \sim N(B_i\theta_i - (1 - B_i)x_i'\beta, B_i^2\psi_i).
\end{aligned}
\tag{2.18}
$$

We can then choose the inverse of the $V(\hat{\theta}_i^B|\theta)$ as an intuitive weight. i.e., $w_i = B_i^{-2}\psi_i^{-1}$. which

is free of θ. Inverse-variance weights are often used when combining multiple random variables so that the weighted sum has a reduced overall variance. This particular choice with weighting matrix $\eta = diag(B_i^{-2}\psi_i^{-1})_{1\leq i\leq m}$ would yield:

$$\tilde{\beta}_{\eta}^{\prime BPE} = \left[\sum_{i=1}^{m} \psi_i x_i x_i'\right]^{-1} \sum_{i=1}^{m} \psi_i x_i y_i, \qquad \eta = diag(B_i^{-2}\psi_i^{-1})_{1\leq i\leq m},$$

free of A. We can see this differs from the best predictive estimator of β under the balanced case: $\psi_i \equiv \psi_0$, $\left[\sum_{i=1}^{m}(\frac{\psi_0}{\psi_0+A})^2 x_i x_i'\right]^{-1} \sum_{i=1}^{m}(\frac{\psi_0}{\psi_0+A})^2 x_i y_i = (\frac{\psi_0}{\psi_0+A})^{-2}[\sum_{i=1}^{m}(x_i x_i')\Big]^{-1} (\frac{\psi_0}{\psi_0+A})^2 \sum_{i=1}^{m} x_i y_i$ $= \left[\sum_{i=1}^{m} x_i x_i'\right]^{-1} \sum_{i=1}^{m} x_i y_i$.

Example 2.21. Weighted BPE of β (Balanced Case). Under the balanced case, $B_1 = B_2 = \ldots = B_m = B$ and there is cancellation of variance ratios:

$$\begin{aligned}
\hat{\beta}_W^{BPE}(A) &= \left[\sum_{i=1}^{m} w_i(A)(1-B_i(A))^2 x_i x_i'\right]^{-1} \sum_{i=1}^{m} w_i(A)(1-B_i(A))^2 x_i y_i \\
&= \frac{1}{(1-B(A))^2}\left[\sum_{i=1}^{m} w_i(A)^2 x_i x_i'\right]^{-1} (1-B(A))^2 \sum_{i=1}^{m} w_i(A) x_i y_i \qquad (2.19)\\
&= \left[\sum_{i=1}^{m} w_i(A) x_i x_i'\right]^{-1} \sum_{i=1}^{m} w_i(A) x_i y_i.
\end{aligned}$$

Compare this to the standard BPE and the MLE of β, in the balanced case they are equal:

$$\hat{\beta}^{MLE} = \hat{\beta}^{BPE} = \left[\sum_{i=1}^{m} x_i x_i'\right]^{-1} \sum_{i=1}^{m} x_i y_i. \qquad (2.20)$$

Fay-Herriot Model Case when A and β are unknown.

We now consider the general case when both A and β are unknown. As with Jiang, Nguyen, & Rao (2011), we feel it is reasonable to estimate A and then use the plug-in estimator $\tilde{\beta}_W^{BPE}(\tilde{A}_W^{BPE})$ for β. To that end, let the vector of unknown parameters by denoted by $\phi_i = (A, \beta)$. Then $\tilde{\theta}_i$ may be considered a function of ϕ, $\tilde{\theta}_i = \tilde{\theta}_i(\phi) = \tilde{\theta}_i(\beta, A)$. Following our approach from when A was known, the expression for MSPE is still the same, except that we are no longer able to drop the first term, since B_i is a function of A. Observe that $E\theta_i^2 = E(\theta_i y_i) = Ey_i^2 - \psi_i$, and that $Ey_i \neq x_i'\beta$ in what follows. As shown in Jiang, Nguyen, & Rao (2011), we have $E(B_i y_- \theta_i) = (B_i - 1)Ey_i$. Then,

$$
\begin{aligned}
WMSE(\hat{\boldsymbol{\theta}}^{BP}) &= E\sum_{i=1}^{m} w_i\Big\{B_i y_i - \theta_i + x_i'\beta(1-B_i)\Big\}^2 \\
&= E\Big\{\sum_{i=1}^{m} w_i(B_i y_i - \theta_i)^2 + 2\sum_{i=1}^{m} w_i(B_i y_i - \theta_i)x_i'\beta(1-B_i) + \sum_{i=1}^{m} w_i(x_i'\beta)^2(1-B_i)^2\Big\} \\
&= E\Big\{\sum_{i=1}^{m} w_i(B_i y_i - \theta_i)^2 + 2\sum_{i=1}^{m} w_i x_i'\beta(1-B_i)E(B_i y_i - \theta_i) + \sum_{i=1}^{m} w_i(x_i'\beta)^2(1-B_i)^2\Big\} \\
&= E\Big\{\sum_{i=1}^{m} w_i(B_i^2 y_i^2 - 2B_i y_i\theta_i + \theta_i^2) - 2\sum_{i=1}^{m} w_i x_i'\beta y_i(1-B_i)^2 + \sum_{i=1}^{m} w_i(x_i'\beta)^2(1-B_i)^2\Big\} \\
&= E\Big\{\sum_{i=1}^{m} w_i\left(B_i^2 y_i^2 - 2B_i(y_i^2 - \psi) + y_i^2 - \psi\right) \\
&\quad - 2\sum_{i=1}^{m} w_i x_i'\beta y_i(1-B_i)^2 + \sum_{i=1}^{m} w_i(x_i'\beta)^2(1-B_i)^2\Big\} \\
&= E\Big\{\sum_{i=1}^{m} w_i\Big[(B_i-1)^2 y_i^2 - (1-2B_i)\psi_i - 2x_i'\beta y_i(1-B_i)^2 + (x_i'\beta)^2(1-B_i)^2\Big\} \\
&= E\Big\{\sum_{i=1}^{m} w_i\Big[(1-B_i)^2(y_i - x_i'\beta)^2 + 2B_i\psi_i - \psi_i\Big]\Big\}.
\end{aligned}
\tag{2.21}
$$

This is different than the derivation in finding $\tilde{\beta}^{BPE}$, we are unable to drop the first term as it

is a function of A through $B_i(A)$. To estimate A, we will seek to minimize:

$$\tilde{Q}(A) = \sum_{i=1}^{m} w_i(A) \left[(1 - B_i(A))^2 (y_i - x_i' \tilde{\beta}^{BPE}(A))^2 + 2B_i(A)\psi_i - \psi_i \right],$$

with respect to A. This can be minimized numerically using any number of iterative scoring algorithms.

The best predictor $\hat{\theta}^{BP} = E(\theta|y)$ is given in matrix form by $\tilde{\theta} = y - \mathbf{\Gamma}(y - \boldsymbol{X}\beta)$, where $\Gamma = diag(1 - B_i)_{1 \le i \le m}$ so that,

$$\begin{aligned}
WMSE &= E\left[(\hat{\theta}^{BP} - \theta)' \boldsymbol{W} (\hat{\theta}^{BP} - \theta)\right] \\
&= E\left[(y - \mathbf{\Gamma}(y - \boldsymbol{X}\beta) - \theta)' \boldsymbol{W} (y - \mathbf{\Gamma}(y - \boldsymbol{X}\beta) - \theta)\right] \\
&= E\left[(y - \theta - \Gamma(y - \boldsymbol{X}\beta))' \boldsymbol{W} (y - \theta - \mathbf{\Gamma}(y - \boldsymbol{X}\beta)\right] \\
&= E\left[(y - \theta)' \boldsymbol{W} (y - \theta)\right] - 2E\left[(y - \theta)' \boldsymbol{W} \mathbf{\Gamma} (y - \boldsymbol{X}\beta)\right] + E\left[(y - \boldsymbol{X}\beta)' \mathbf{\Gamma} \boldsymbol{W} \mathbf{\Gamma} (y - \boldsymbol{X}\beta)\right] \\
&= tr(\boldsymbol{W}\psi) - 2E\left[(y - \theta)' \boldsymbol{W} \mathbf{\Gamma} y\right] + E\left[(y - \boldsymbol{X}\beta)' \mathbf{\Gamma} \boldsymbol{W} \mathbf{\Gamma} (y - \boldsymbol{X}\beta)\right] \\
&= tr(\boldsymbol{W}\psi) - 2tr(\boldsymbol{W}\mathbf{\Gamma}(AI_m + \psi)) + 2E\{\theta' \boldsymbol{W} \mathbf{\Gamma} \theta\} + E\left[(y - \boldsymbol{X}\beta)' \mathbf{\Gamma} \boldsymbol{W} \mathbf{\Gamma} (y - \boldsymbol{X}\beta)\right] \\
&= E\left\{-tr(\boldsymbol{W}\psi) + 2Atr(\boldsymbol{W}\mathbf{\Gamma}) + (y - \boldsymbol{X}\beta)' \mathbf{\Gamma} \boldsymbol{W} \mathbf{\Gamma} (y - \boldsymbol{X}\beta)\right\}.
\end{aligned} \tag{2.22}$$

This is equivalent to minimizing the integrand within the expectation, where we plug in

$$Q(A) = -tr(\boldsymbol{W}\boldsymbol{\psi}) + 2Atr(\boldsymbol{W}\boldsymbol{\Gamma}) + (y - \boldsymbol{X}\tilde{\beta}_W^{BPE})'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}(y - \boldsymbol{X}\tilde{\beta}_W^{BPE})$$
$$= -tr(\boldsymbol{W}\boldsymbol{\psi}) + 2Atr(\boldsymbol{W}\boldsymbol{\Gamma}) + y'\boldsymbol{W}^{1/2}\boldsymbol{\Gamma}\left\{I_m - \boldsymbol{W}^{1/2}\boldsymbol{\Gamma}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Gamma}\boldsymbol{W}^{1/2}\right\}\boldsymbol{\Gamma}\boldsymbol{W}^{1/2}y. \tag{2.23}$$

The solution to this minimizer would be arrived out numerically to obtain the best predictive estimator for A as:

$$\tilde{A}_W^{BPE} = \arg\min_A \text{Q(A)}.$$

Example 2.22. **Discussion**

An interesting use case from above is the case for when W = $\Gamma^{-2} = diag(\frac{(A+\psi_i)^2}{\psi_i^2})_{1\leq i\leq m}$. Then the objective function becomes Q(A) = $-tr(\boldsymbol{\Gamma}^{-2}\psi) + 2Atr(\boldsymbol{\Gamma}^{-1}) + y'\left\{I_m - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right\}y$ The last term does not depend on A, and so therefore we can only consider the first two terms: Q(A) = $-\sum_{i=1}^m(\frac{(A+\psi_i)^2}{\psi_i}) + 2A\sum_{i=1}^m\frac{A+\psi_i}{\psi_i} \implies \frac{d}{dA}Q(A) = -2\sum_{i=1}^m(\frac{A+\psi_i}{\psi_i}) + 2\sum_{i=1}^m\frac{A+\psi_i}{\psi_i} + 2A\sum_{i=1}^m\frac{1}{\psi_i} = 0$ $\implies \tilde{A}_{\Gamma^{-2}}^{BPE} = 0$. This means that there may exist weighting matrices $\boldsymbol{W}$ such that weighted best predictive estimator is zero.

2.5 Small Area Income and Poverty Estimates Data Analysis

It is instructive to demonstrate the proposed tools in this chapter using a well-known dataset in the small area literature. The U.S. Census Bureau's Small Area Income and Poverty Estimates program (SAIPE) produces estimates at the county and state level to measure the total number of people (and children) in poverty along with median household income. In order to

produce this information, the Census Bureau will leverage survey estimates from the American Community Survey, which samples approximately three million addresses per year. These direct estimates are then augmented with administrative data obtained from the Food and Drug Administration, Internal Revenue Service, and other governmental sources as part of the bureau's interagency agreements to acquire such records solely for statistical purposes. Prior to 2005, the SAIPE program relied upon the Current Population Survey (CPS) to monitor poverty in small geographical areas. This is a survey which sampled approximately 100,000 addresses in 2005, and only about 60,000 addresses in 1993. The SAIPE has been widely discussed in the literature on small area estimation, more information can be gleaned from Bell et al. (2016) or directly from the bureau website at https://www.census.gov/programs-surveys/saipe.html. The data used in the analysis below was also analyzed independently by Bell & Franco (2017) and are available at https://www.census.gov/srd/csrmreports/byyear.html.

The following variables were included in a SAIPE date from 1993 comprising poverty estimates for children between 5 and 17 for fifty states and the District of Columbia.

1. **CPS** - direct estimate from the Current Population Survey for the state poverty rate of children ages 5-17

2. **IRSPR** - Poverty rate based on IRS tax data, defined as # Child tax exemptions from impoverished households / # Total child tax exemptions from all households.

3. **IRSNF** - The tax non-filer rate based on IRS tax data, defined as [Population - # Tax Exemptions under Age 65] / Estimated Population of Persons 65+.

4. **FS** - Food Stamp participation rate as measured by the Supplemental Nutritional Assistance Program (SNAP).

5. **GVFSE** - GVF estimates of sampling standard errors from the CPS. These are based on iterative procedures developed by Bell and Otto (1995). The method switches between ML estimates of model parameters vs. the estimation of sampling standard errors.

6. **CENRES** - Residuals from the 1990 census where the full model (with the same covariates) was fit to the same outcome of the poverty rate for children between ages 5 and 17.

The computed statistics in the data analysis included the EBLUP w/ REML(A) and the MLE(β); the OBP w/ BPE for both β and A; and finally, the GBP w/ WBPE for both β and A. Also included in the analysis are comparisons of these three different β's and values for A. The values of the weights in the GBP are of the form $w_i = \frac{\phi_i}{B_i^2 A^2}$. Table 2.23 depicts the various estimators to be used in the analysis. We stress that $\hat{B}_i$ is not estimated separately from $\hat{A}$, so that $\hat{B}_i = B_i(\hat{A})$.

Table 2.23. *Estimators for Nuisance Parameters for Small Area Prediction (and weights, where applicable)*

Estimators and Weights			
$\hat{\theta}_i$	**EBLUP**	**OBP**	**GBP**
β	MLE	BPE	WBPE
A, B_i	REML	BPE	WBPE
w_i	N/A	N/A	$\frac{1}{B_i^2 \psi}$

Parametric Bootstrap for Mean Squared Error Estimation of Small Area Predictions

We wish to evaluate the precision of the small area predictors. In general, computing estimates of mean-squared error can be difficult. In order draw suitable comparisons, we will resort to using a parametric bootstrap approach. The bootstrap is a well-known resampling method which is really indispensable for modern analysts, especially with regard to its utility in estimating variance and because of its non-parametric nature. For our purposes, it is intuitive to use a more Monte-Carlo version of the approach, since our estimators are based on an inherently assumed model. We generate data according to the model (Y_i* and θ_i^*) and then proceed to predict θ_i^* using candidate estimators $\hat{\theta}_i^*$ with y_i^* as inputs, $\hat{\theta}_i^* = \hat{\theta}_i^*(y_i^*)$. The parametric bootstrap MSE estimator is given by:

$$MSPE_B(\hat{\theta}_i) = E_* \left[\hat{\theta}_i^* - \theta_i^*\right]^2,$$

$$MSPE_B(\hat{\theta}) = \frac{1}{m}\sum_{i=1}^{m} MSPE_B(\hat{\theta}_i) = \frac{1}{m}\sum_{i=1}^{m} E_* \left[\hat{\theta}_i^* - \theta_i^*\right]^2,$$

where E_* is expectation with respect to the parametric bootstrap distribution and $\hat{\theta}_i$ can represent any of the estimators discussed in this chapter: $\hat{\theta}_i^{EBLUP}$, $\tilde{\theta}_i^{OBP}$, or $\tilde{\theta}_i^{GBP}$. In practice, we will use Monte Carlo to approximate the above:

$$E_* \left[\hat{\theta}_i^* - \theta_i^*\right]^2 \approx \frac{1}{R}\sum_{r=1}^{R} \left[\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right]^2,$$

where $\theta_i^{(r)}$ is the r^{th} draw from Level 2 of the FH model with estimated β and A; $\hat{\theta}_i^{OBP(r)}$ is the OBP computed from the r^{th} bootstrap sample; R is the number of bootstrap samples.

Now the bootstrap samples are still based on estimators from the data, and will have equivalent

hyperparameters as the underlying mean. We outline the procedure explicitly:

Steps in Parametric Bootstrap MSE Estimation

1. Generate θ_i^* by sampling from a $N(x_i\hat{\beta}, \hat{A})$ distribution.

2. Generate y_i by sampling from a $N(\theta_i^*, \psi_i)$ distribution.

3. Compute $\hat{\theta}_i^*$ using the same estimation process of $\hat{\theta}_i$ on the generated data y^*, with $\hat{\theta}_i^* = \hat{\theta}_i(y^*, x_i)$.

4. Label θ_i^* and $\hat{\theta}_i^*$ as $\hat{\theta}_i^{(1)}$ and $\hat{\theta}_i^{(1)}$. Repeat steps 1-3 R-1 times and relabel.

5. Obtain $\text{MSPE}_B(\hat{\theta}_i) = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right]^2$ and $\text{MSPE}_B(\hat{\theta}) = \frac{1}{mR}\sum_{r=1}^{R}\sum_{i=1}^{m}\left[\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right]^2$.

Theoretical properties of the parametric bootstrap are not known for general parameters $\phi_i = h(\theta_i)$, see Rao and Molina (2015). The motivation behind it is that the bootstrap distributions will converge to the true distribution with large R. Moreover, the law of large numbers gives credence to the notation that the MSPE averaged over all R iterations converges to the true MSPE. While straightforward to implement, the parametric bootstrap is not second order unbiased. Jiang, Nguyen, & Rao (2011) proposed an area-specific MSPE bootstrap having the property of second-order unbiasedness.

Data analysis results with SAIPE 1993, R=200

Table 2.24. Competing Estimators of Parameter Coefficients (SAIPE 1993)

Covariate	$\hat{\beta}^{MLE}$	$\hat{\beta}^{BPE}$	$\hat{\beta}_{W}^{BPE}$
Intercept	-3.84	-3.56	-4.29
CENRES	1.21	2.14	3.02
FS	1.26	1.94	2.25
IRSNF	0.52	0.28	0.24
IRSPR	0.23	0.06	-0.01

From Table 2.24 above we can see there are non-trivial differences in the parameter coefficients across the three classes of parameter estimates. Table 2.25 follows with a display of the visible differences across states in the original CPS Poverty Rates and the three classes of poverty rate prediction: EBLUP, OBP, and GBP.

Table 2.25. Competing Estimators of Small Area Poverty Rates and Variance Components (SAIPE 1993)

State	CPS	$\hat{\theta}_i^{EBLUP}$	$\tilde{\theta}_i^{OBP}$	$\tilde{\theta}_i^{GBP}$	$\hat{B}_i^{REML}$	$\tilde{B}_i^{BPE}$	$\tilde{B}_i^{WBPE}$
AL	20.3	22.5	24.3	23.9	0.13	0.22	0.47
AK	9.4	12.2	9.9	9.2	0.18	0.30	0.56
AZ	22.3	23.9	23.3	23.1	0.11	0.19	0.42
AR	24.6	22.3	23.7	24.7	0.11	0.19	0.42
CA	23.8	22.8	21.8	22.8	0.48	0.63	0.84
CO	11.0	13.2	11.8	11.2	0.21	0.33	0.60
CT	14.8	14.5	16.2	16.8	0.13	0.21	0.45
DE	12.8	13.4	13.7	13.4	0.11	0.18	0.41
DC	49.0	30.8	36.7	41.8	0.04	0.08	0.21
FL	25.9	22.2	22.9	24.4	0.29	0.44	0.70
GA	16.8	21.0	21.4	20.5	0.13	0.22	0.47
HI	13.0	12.7	12.1	11.7	0.13	0.21	0.45
ID	13.6	12.5	11.1	11.4	0.19	0.31	0.58
IL	17.5	17.6	18.3	18.1	0.35	0.50	0.75
IN	10.3	14.0	14.5	13.2	0.20	0.31	0.58

We also show the differences in the shrinkage parameters, B_i, as well as the discrepancies in the estimators of the variance components, A. Finally, we showcase the mean-square prediction errors for the three class of small area means prediction of poverty rates in SAIPE 1993. We also consider the weights $w_i = \psi_i$ and $w_i = 1/\psi_i$.

Table 2.26. Competing Estimators of Variance Components (SAIPE 1993)

Values of $\hat{A}$		
Type	**Value**	% Zeros
$\hat{A}^{REML}$	2.07	14%
$\tilde{A}^{BPE}$	3.87	6%
$\tilde{A}_W^{BPE}(w_i = B_i^{-2}/\psi_i)$.	11.76	0%
$\tilde{A}_W^{BPE}(w_i = 1/\psi_i)$.	1.07	NA
$\tilde{A}_W^{BPE}(w_i = \psi_i)$.	7.99	0%

Table 2.27. Bootstrap Estimator of MSPE (SAIPE 1993)

$MSPE_B(\hat{\theta})$	
Type	**Value**
$\hat{\theta}^{EBLUP}$	7.3
$\tilde{\theta}^{OBP}$	5.4
$\tilde{\theta}^{GBP}(w_i = B_i^{-2}/\psi_i)$	7.2
$\tilde{\theta}^{GBP}(w_i = \psi_i)$	5.1

Discussion. There were noticeable differences in the estimated parameter coefficients, not only between the LSE and BPE, but also between the BPE and WBPE. Estimates of the model variance A were also different, resulting in different amounts in the shrinkage parameter B_i. These values were higher for the GBP, which thus borrowed less strength from the synthetic estimate but

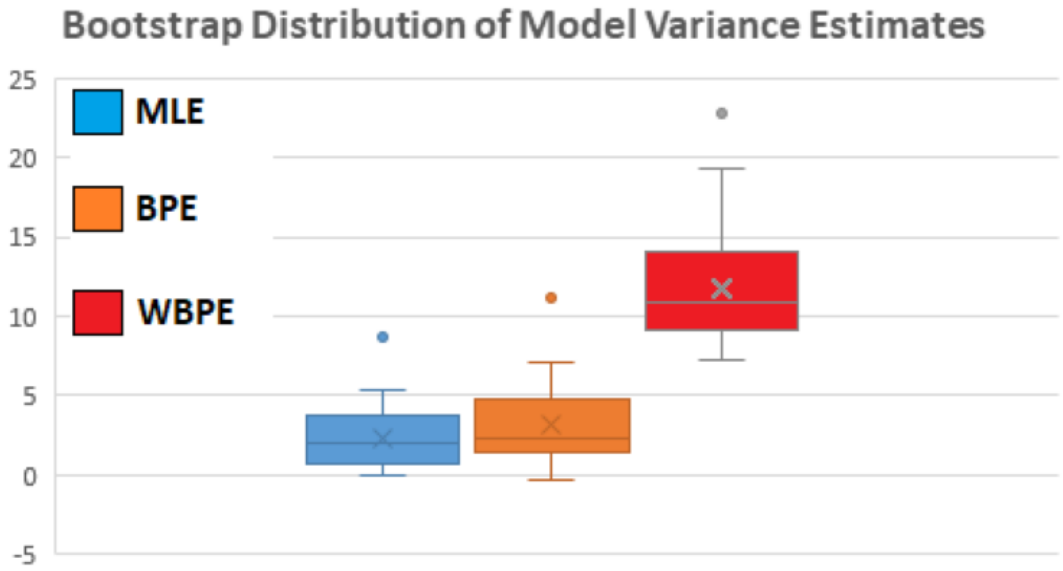
resulted in similar estimated MSPE from the EBLUP. The bootstrap MSPE from the OBP was lower than that of the EBLUP, a possible indication of model misspecification. We did observe some zero estimates from the REML, a phenomenon occurring in many such estimators of variance components, resulting in overshrinkage (i.e., too much weight being placed on the synthetic estimator).

The weighted best predictor of A was much higher than that of the BPE and REML. Our objective is not necessarily getting precision estimates of the model variance. We are willing to permit some overestimation of the model variance to obtain improved small area means. The OBP seems to be an improvement over the EBLUP, while the weighted estimator based on weights $w_i = B_i^{-2}/\psi_i$ had a similar bootstrap MSE without succumbing to over-shrinkage in any of the iterations.

Additional computational runs were made for the alternative weights, $w_i = \psi_i$ and $w_i = 1/\psi_i$. The latter resulted in an estimator for A even lower than REML, causing a large amount of zero estimates when attempting to compute the parametric bootstrap. On the other hand, the choice of weight based on the variance ($w_i = \psi_i$) generated an estimator with a lower bootstrap MSPE than that of the OBP.

Figure 2.28. Bootstrap Histograms of $\hat{A}$ (SAIPE 1993, $w_i = B_i^{-2}/\psi_i$).

Allowing the weight w_i to be a function of A can have some ramifications when trying to find the value of A that optimizes $(\tilde{\theta} - \theta)'\boldsymbol{W}(A)(\tilde{\theta} - \theta)$. An arbitrary weighting function of A could have an adverse effect on the objective function. Caution should be exercised when selecting suitable importance weights. We close this section with Figure 2.28, which depicts the bootstrap histograms of the estimators of the model variance A. The weighted best predictor is the furthest from zero, while REML estimates of A have about 14% of its bootstrap samples being zero.



2.6 Simulation Study

We now undertake a simulation study, still somewhat based on the SAIPE dataset. We will continue to use the same covariates (CENRES, FS, IRSNF, and IRSPR) and sampling variances ψ_i for each state i=1, 2, . . . , m=51. However, we will create arbitrary parameter coefficients, and reconstruct the mean response variable according to those coefficients.

Simulation Regression Function:

$$\boxed{\mu_i = E(\theta_i) = 5 + 3.16 \times CENRES + 7 \times FS + 3.2 \times IRSNF + 5.6 \times IRSPR}.$$

Moreover, we generate random effects according to incremental values of A which loosely range between $\tilde{A}^{BPE}$ and $\bar{\psi} = \frac{1}{m}\sum_{i=1}^{m}\psi_i$. In this fashion, we get a "true" value of the small area means in our simulated population, which we will denote as θ^*. We proceed to generate $\theta^* =$

$\boldsymbol{X}\beta^* + \nu$, where $\nu \sim N_m(0, AI_m)$.

Simulated Values of Model Variance, A	
Simulation #	**Value**
A_1	4
A_2	8
A_3	16

Finally, we generate new y values, denoted by y^* according to $y^* \sim N_m(\theta^*, \psi)$. Now that we have a *known* population θ^*, we can compare the results on the full model, and also under reduced models to observe and compare the robustness of the different estimators computed only from y_* and $\boldsymbol{X}$. Two hundred iterations were taken for Table 29, which includes Monte Carlo errors, with the r^{th} value of θ_i^* and $\hat{\theta}_i^*$ being denoted as $\theta_i^{(r)}$ and $\hat{\theta}_i^{(r)}$, respectively. The *simulation* MSPE is computed similarly in the data analysis, $\text{MSPE}(\hat{\theta}_i) = \dfrac{1}{R}\displaystyle\sum_{r=1}^{R}\left[\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right]^2$ and $\text{MSPE}(\hat{\theta}) = \dfrac{1}{mR}\displaystyle\sum_{r=1}^{R}\sum_{i=1}^{m}\left[\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right]^2$. Subsequent tables had fifty iterations, as that was deemed sufficient earlier runs. Moreover, we also wish to compute the Monte Carlo error, MCE_B, associated with $MSPE_B(\hat{\theta})$. Let $\gamma_i^{(r)} = \left[\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right]^2$ denote the squared difference between the bootstrap estimator of θ_i^r for the i^{th} domain under the r^{th} bootstrap iteration. Then the Monte Carlo error for the i^{th} domain is defined as:

$$MCE(\hat{\theta}_i) = \frac{1}{\sqrt{R}}\sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left[\gamma_i^{(r)} - \bar{\gamma}_i\right]^2},$$

where $\bar{\gamma}_i = \frac{1}{R}\displaystyle\sum_{r=1}^{R}\gamma_i^{(r)}$. Finally, the Monte Carlo errors for each domain i are averaged to obtain a single error values to represent the dataset. This is the Monte Carlo error, which can be associated

with the $\text{MSPE}_B(\cdot)$ for any of the competing estimators.

$$MCE(\hat{\theta}) = \frac{1}{m}\sum_{i=1}^{m} MCE(\hat{\theta}_i).$$

Finally, we define the *weighted* simulation MSPE as below. To differentiate between possibly different weights used for the GBP and WBPE, we will use the notation w_i to represent the WBPE weights, and w_i^* to denote the weighted corresponding to the risk functions below:

$$WMSPE(\hat{\theta}_i) = \frac{1}{R}\sum_{r=1}^{R} w_i \left[\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right]^2$$

$$WMSPE(\hat{\theta}) = \frac{1}{mR}\sum_{r=1}^{R}\sum_{i=1}^{m} w_i^* \left[\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right]^2$$

Simulation Case 1. A known, $w_i^* = \psi_i$. We first consider the case when the model variance A is known to be 16 and let β be unknown. For the generalized OBP, we elect to have the weight defined as a function of the sampling variance as the weight: either $w_i = \psi_i$ or $w_i = 1/\psi_i$, and consider corresponding weighted loss functions: $w_i^* = \psi_i$ and $w_i^* = 1/\psi_i$. These loss functions are applied to both, the small area means and the norm of the estimated parameter coefficient vector.

Results. In Table 2.29, it is clear the that mean square prediction error increases with the model variance A. We also observe the EBLUP has the lowest MSPEs across all variance values, and that

the EBLUP also outperforms in the context of the weighted loss functions.

Table 2.29a. (A known). Simulation results under correctly specified regression model (CENRES, FS, IRSNPR, and IRSNF), R = 200 iterations.

Simulation MSPE by Model Variance and Loss Type Full Model: CENRES, FS, IRSNF, IRSPR									
Loss Functions	Standard Loss $w_i^* \equiv 1$			Variance weights $w_i^* = \psi_i$			Inverse Variance $w_i^* = 1/\psi_i$		
$MSPE(\cdot)$	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**
$\hat{\theta}^{EBLUP}$	3.5	5.1	6.9	48.0	70.1	99.5	0.35	0.50	0.65
$\tilde{\theta}^{OBP}$	4.0	5.5	7.3	55.2	79.1	108.2	0.40	0.53	0.68
$\tilde{\theta}^{GBP}$ $(w_i = \psi_i)$	5.0	6.4	8.0	74.3	96.6	122.9	0.50	0.60	0.72
$\tilde{\theta}^{GBP}$ $(w_i = 1/\psi_i)$	3.6	5.1	7.0	48.3	71.6	100.5	0.36	0.50	0.66
$\|\|\beta\|\|^{LSE}$	2.2	2.6	3.1	25.9	29.8	36.0	0.27	0.32	0.38
$\|\|\beta\|\|^{BPE}$	2.7	3.4	4.1	33.5	39.4	48.5	0.36	0.42	0.51
$\|\|\beta\|\|^{WBPE}$ $(w_i = \psi_i)$	3.6	4.2	5.0	42.6	48.5	58.1	0.45	0.52	0.62
$\|\|\beta\|\|^{WBPE}$ $(w_i = 1/\psi_i)$	2.3	2.7	3.3	26.6	31.5	28.9	0.28	0.33	0.41

Table 2.29b. (A known). Monte Carlo Errors associated with MSPE for the correctly specified model (CENRES, FS, IRSNPR, and IRSNF) for MSPE estimates, R = 200 iterations.

Monte Carlo Error by Model Variance and Loss Type Full Model: CENRES, FS, IRSNF, IRSPR									
Loss Functions	Standard Loss $w_i^* \equiv 1$			Variance weights $w_i^* = \psi_i$			Inverse Variance $w_i^* = 1/\psi_i$		
$MCE(\cdot)$	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**
$\hat{\theta}^{EBLUP}$	0.69	1.07	1.49	13.0	19.9	28.9	0.07	0.10	0.13
$\tilde{\theta}^{OBP}$	0.81	1.17	1.60	15.5	23.2	33.2	0.08	0.11	0.13
$\tilde{\theta}^{GBP}$ $(w_i = \psi_i)$	1.04	1.38	1.78	23.9	31.2	41.3	0.10	0.12	0.14
$\tilde{\theta}^{GBP}$ $(w_i = 1/\psi_i)$	0.72	1.07	1.51	13.0	20.0	29.2	0.08	0.10	0.13

Table 2.29c. (A known). Monte Carlo Errors associated with MSPE for the correctly specified model (CENRES, FS, IRSNPR, and IRSNF) for parameter coefficients, R = 200 iterations.

Monte Carlo Error by Model Variance Full Model: CENRES, FS, IRSNF, IRSPR			
Loss Functions	All Loss Functions $w_i^* \equiv 1$		
$MSPE(\cdot)$	**A=4**	**A=8**	**A=16**
$\|\|\beta\|\|^{LSE}$	2.2	2.5	3.1
$\|\|\beta\|\|^{BPE}$	2.9	3.4	4.1
$\|\|\beta\|\|^{WBPE}$ $(w_i = \psi_i)$	3.6	4.1	4.9
$\|\|\beta\|\|^{WBPE}$ $(w_i = 1/\psi_i)$	2.3	2.7	3.3

This result is somewhat expected, as under the correctly specified model, the EBLUP will provide the lowest mean squared error. It is only under the misspecified model will the best predictive models exhibit advantages over the EBLUP. We are also able to see that under the weight $w_i = 1/\psi_i$, both MSPE values for θ and β approach the same levels as the optimal EBLUP. This is a remarkable result, and parallels might be drawn between weighted least squares and weighted best prediction. Under weighted least squares, the value of the weight would be $\frac{1}{\psi_i + A}$ rather than just $\frac{1}{\psi_i}$, but A will always be unknown in practice and is therefore unsuitable as an importance weight. Even if known, it would only serve as an identical translation applied to areas, and is thus of less value under importance weighting. Finally, a closer inspection reveals that the GBP remains very close to the EBLUP, but is always slightly higher in line with theoretical result of the EBLUP being an optimal estimator under correctly specified models.

Misspecification Models. We now assess simulation under misspecified models, an examination of the robustness of the best prediction algorithms. We will consider the two combinations, (CENRES, FS) as one covariate model, and (IRSNPR, IRSNF) as another model. The simulated population will remain identical, but the prediction models will lack half the underlying variable information.

Table 2.30. (A known). Simulation results with known model variance on misspecified regression model # 1 (Only using IRSNPR and IRSNF), R = 50 iterations.

Simulation MSPE by Model Variance and Loss Type Misspecified Model: Only IRSNF, IRSPR Used									
Loss Functions	Standard Loss			Weighted Loss			Weighted Loss		
	$w_i^* \equiv 1$			$w_i^* = \psi_i$			$w_i^* = 1/\psi_i$		
$MSPE(\cdot)$	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**
$\hat{\theta}^{EBLUP}$	87.7	61.2	38.4	1470	1108	745	8.1	5.0	2.9
$\tilde{\theta}^{OBP}$	84.8	57.6	35.1	1288	914	572	8.6	5.2	3.0
$\tilde{\theta}^{GBP}$ $(w_i = \psi_i)$	91.0	61.9	37.6	1195	861	536	10.1	6.1	3.4
$\tilde{\theta}^{GBP}$ $(w_i = 1/\psi_i)$	91.2	61.3	37.1	1597	1118	697	8.0	5.0	2.8
$\|\|\beta\|\|^{LSE}$	9.1	9.2	9.5	106	108	111	1.1	1.1	1.2
$\|\|\beta\|\|^{BPE}$	14.8	16.8	19.0	173	196	222	1.8	2.1	2.4
$\|\|\beta\|\|^{WBPE}$ $(w_i = \psi_i)$	22.1	23.6	25.2	259	276	298	2.7	2.9	3.2
$\|\|\beta\|\|^{WBPE}$ $(w_i = 1/\psi_i)$	8.6	9.7	11.7	101	113	137	1.1	1.2	1.5

It will be interesting to observe whether the generalized observed best predictor will retain its robustness properties as the OBP. We caution that these results will not be consistent with earlier sections of this chapter, as we have redefined the relationship among variables, and their predictive power is also changed. We expect the OBP and GBP to have lower MSPE than the EBLUP, and we

also expect to so the WBPE's to produce lower MSPE's than other estimators whenever the same weights are used in the loss function (i.e., $w_i \equiv w_i^*$).

Table 2.31. (A known). Simulation results with known model variance on misspecified regression model # 2 (Only using CENRES and FS), R = 50 iterations.

Simulation MSPE by Model Variance and Loss Type Misspecified Model: Only CENRES, FS Used									
Loss Functions	Standard Loss			Weighted Loss			Weighted Loss		
	$w_i^* \equiv 1$			$w_i^* = \psi_i$			$w_i^* = 1/\psi_i$		
$MSPE(\cdot)$	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**
$\hat{\theta}^{EBLUP}$	189.8	122.9	68.3	2405	1641	951	20.5	11.7	6.0
$\tilde{\theta}^{OBP}$	187.2	120.2	68.5	2320	1574	914	21.0	12.0	6.0
$\tilde{\theta}^{GBP}$ $(w_i = \psi_i)$	189.4	121.6	67.3	2493	1633	927	21.7	12.2	6.1
$\tilde{\theta}^{GBP}$ $(w_i = 1/\psi_i)$	193.3	122.5	66.9	2318	1579	920	20.4	11.7	5.9
$\|\|\beta\|\|^{LSE}$	55.2	55.2	55.1	645	645	645	6.9	6.9	6.9
$\|\|\beta\|\|^{BPE}$	56.4	56.8	57.0	659	664	667	7.0	7.1	7.1
$\|\|\beta\|\|^{WBPE}$ $(w_i = \psi_i)$	57.4	57.3	57.1	635	643	651	7.1	7.1	7.1
$\|\|\beta\|\|^{WBPE}$ $(w_i = 1/\psi_i)$	54.3	55.0	55.6	671	670	667	6.7	6.8	6.9

Results. In contrast to the correctly specified model, we note the observed MSPE's are anti-correlated to the model variances. With the higher model variances, more weight is taken from direct estimate as opposed to the biased synthetic estimator. When the synthetic estimator is *poorly* specified, too much bias may be introduced through the composite estimator. The MSPE values for the estimates were much larger for the model specification based only on the Census residuals and food stamp usage (CENRES and FS), an indication that this model was less prescient than the model containing the IRS non-filer and poverty rate estimates (IRSNF and IRSPR). With respect to the standard loss functions, the generalized best predictors displayed some MSPEs lower than that

of the EBLUP, and lower than the OBP in some cases, as well. While the GBP based on $w_i = 1/\psi_i$ had the lowest weighted MSPE under the corresponding weight w_i^*, the GBP based on $w_i = \psi_i$, however, did not have the lowest weighted MSPE with respect to its corresponding loss functions weighted by $w_i^* = \psi_i$. In this case our results are somewhat mixed. Finally, We observe that the estimated parameter coefficients were not greatly affected by the increased model variances. Under the standard loss, the MSPE of the weighted best predictor $w_i = 1/\psi_i$ was lower than that of the MLE.

Discussion. Generally, the selection of the weights w_i should be tailored toward a specific weighted loss function. There is some evidence that the weights $w_i = 1/\psi_i$ could be effective as a weighted best predictive estimator in the sense of weighted least squares. This estimator displayed some MSPE scenarios lower than the EBLUP.

Simulation Case 2. β **known**. We now consider the case when the parameter coefficients A is known to be 16 and let β be unknown. For the generalized OBP, we elect to have the weight defined as a function of the sampling variance as the weight: either $w_i = \psi_i$ or $w_i = 1/\psi_i$, and consider corresponding weighted loss functions: $w_i^* = \psi_i$ and $w_i^* = 1/\psi_i$. These loss functions are applied to both, the small area means and the norm of the estimated parameter coefficient vector.

Results. Under the appropriately specified model, the MSPE values for all models were very close across all loss functions, except for that of GBP($w_i = \psi_i$) weighted by the variance, which uniformly higher. We also note the weighted BPE for A based on the inverse weights $1/\psi_i$ has the lowest MSPE across all loss functions. Under the misspecified model, that version of the WBPE

and GBP had the lowest (or equivalent) MSPE values.

Table 2.32. (β known). Simulation results with correct specification for model variance, R = 50 iterations.

Simulation MSPE by Model Variance and Loss Type Correct Model Variance									
Loss Functions	Standard Loss			Weighted Loss			Weighted Loss		
	$w_i^* \equiv 1$			$w_i^* = \psi_i$			$w_i^* = 1/\psi_i$		
$MSPE(\cdot)$	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**	**A=4**	**A=8**	**A=16**
$\hat{\theta}^{EBLUP}$	3.1	4.7	6.6	39.1	63.6	93.7	0.33	0.48	0.63
$\tilde{\theta}^{OBP}$	3.1	4.7	6.6	39.5	64.0	94.3	0.33	0.48	0.64
$\tilde{\theta}^{GBP}$ $(w_i = \psi_i)$	3.4	5.0	6.9	43.0	67.5	96.9	0.37	0.51	0.66
$\tilde{\theta}^{GBP}$ $(w_i = 1/\psi_i)$	3.1	4.7	6.6	39.3	63.5	93.5	0.33	0.48	0.63
A^{LSE}	5.8	11.6	27.7	68.4	136.2	323.4	0.73	1.45	3.44
A^{BPE}	6.4	12.2	30.0	75.0	142.2	350.9	0.80	1.51	3.73
A^{WBPE} $(w_i = \psi_i)$	25.1	49.4	119.0	293.1	578.2	1392	3.11	6.14	14.79
A^{WBPE} $(w_i = 1/\psi_i)$	4.3	8.4	19.1	50.4	98.6	222.7	0.54	1.05	2.37

Table 2.33. (β known). Simulation results with incorrect specification for model variance, $R = 50$ iterations.

Simulation MSPE by Model Variance and Loss Type Misspecified Model Variance			
	Loss Functions		
$MSPE(\cdot)$	$w_i^* \equiv 1$	$w_i^* = \psi_i$	$w_i^* = 1/\psi_i$
$\hat{\theta}^{EBLUP}$	5.77	78.0	0.57
$\tilde{\theta}^{OBP}$	5.79	78.1	0.57
$\tilde{\theta}^{GBP}$ $(w_i = \psi_i)$	6.05	81.5	0.60
$\tilde{\theta}^{GBP}$ $(w_i = 1/\psi_i)$	5.74	77.3	0.57
A^{LSE}	35.8	436	4.4
A^{BPE}	36.0	419	4.5
A^{WBPE} $(w_i = \psi_i)$	79.7	929	9.9
A^{WBPE} $(w_i = 1/\psi_i)$	28.8	336	3.6

2.7 Concluding Remarks

We reviewed the observed best predictor as applied to the Fay-Herriot model. We also revisited the celebrated James-Stein estimator in the context of the OBP and EBP, and investigated some related properties. The concepts of the weighted best predictor and the generalized observed best predictor were introduced. In particular, the selection of appropriate importance weights was considered. To summarize, these could reflect subjective criteria placed on each specific domain, or they could be a function of the sampling variance and model variance to produce new estimators with desirable properties. A number of different GBP's with varying weights were compared to both the EBLUP and OBP using the SAIPE dataset from the U.S. Census Bureau. A simulation study was also conducted, where the true population parameters were known. This was conducted

for both the parameter coefficients β and the model variance.

The simulation studies showed some mixed results. As expected, the EBLUP performed well under properly specified models, as did the OBP to a lesser extent. Under misspecified models, the OBP outperformed the EBLUP. The results of the GBP were dependent on the choice of weights. The GBP inverse variance weights performed well, very close to the optimal EBLUP and OBP in different scenarios, raising the possibility of finding optimal weights in future studies. While the GBP associated with the different risk functions fared well, results were sometimes sensitive to the level of model variance.

Chapter 3

Empirical Bayes Parametric Bootstrap Model Selection and Diagnostics in the Fay-Herriot Model

3.1 Introduction

There are a multitude of possibilities for modeling data sets, and the analyst must confront decisions between competing models. Once a model type is selected, we need to evaluate which dependent variables should be included into the model. This is a problem that goes back to linear regression modeling, and a number of solutions have propagated, including stepwise methods and lasso regression. In machine learning approaches, it is commonplace to build models on a portion of the data (the so-called *training* dataset) and then assess model suitability on the remaining portion of data, coined as the *analysis* dataset. In a similar validation context, the [posterior] predictive criterion has been developed to perform cross-validation. We first review the more traditional model validation approaches and then develop a cross-validation approach for the Fay-Herriot model. In general, this methodology could be used to investigate the normality assumption, assess the soundness of homoskedasticity, diagnostics, for variable selection or even deciding between fixed and random effects.

We will first survey a plethora of available metrics still currently used for model selection: including the coefficient of determination, log-likelihood, AIC, BIC, DIC, and WAIC. These

metrics permit analysts to draw comparisons between different models, but they appear to have various drawbacks which have been brought to light by both applied scientists and theoreticians. For the last metric, the Watanabe-Akaike Information Criterion has the drawback of requiring greater amounts of processing time. On the contrary, with the advent of high-speed computing and analytics, we are more willing to pay this price for greater accuracy and precision in our model building approach. In particular, the parametric bootstrap affords us an opportunity to rely on machine computing in the presence of complicated distributionals. We specifically develop the Empirical Bayes Parametric Bootstrap (EBPB) approach for model selection and diagnostics, and apply the methods to the SAIPE Census Bureau data.

Outline of the Chapter

(3.1) Introduction.

(3.2) Standard Metrics for Model Building. A review of mainstream methods currently in use for model assessment.

(3.3) Empirical Bayes Parametric Bootstrap Cross-Validation. Using the ideas proposed in Lahiri (2020), we develop and evaluate the empirical Bayes leave-one-out cross-validation approach using the parametric bootstrap.

(3.4) Empirical Bayes LOO-CV Model Selection with Application to SAIPE data.

(3.5) Empirical Bayes Parametric Bootstrap Model Diagnostics. We develop a residual analysis based on an empirical Bayes approach.

(3.6) Concluding Remarks.

3.2 Standard Metrics for Model Building

Coefficient of Determination

Suppose that we were only interested in a national estimate and that all of the area-level random effects were null. Then under such a standard linear regression, the coefficient of determination, denoted by R^2, is equivalent to

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}.$$

A typical interpretation R^2 (pronounced 'r-squared') is that it represents the amount of variation explained by the model. It is related to the Pearson correlation coefficient ρ under simple linear regression as $R^2 = \rho^2$. Opinions on the acceptable values of R^2, differ, but they should also be considered in the context of significance tests for the covariates and the global omnibus F-tests. *Partial* coefficients of determination have been developed for reduced models, which can be used as the basis for stepwise variable selection procedures. Since R^2 is sensitive to the number of covariates, an *adjusted* version is used for such purposes, too. The coefficient of determination has been extended for random effects models and time-series models where the independence assumption is violated (e.g. see Buse (1973)). The following *generalized* coefficient of determination is used by default in the SAS Software Panel procedure (SAS V9 (2020)),

$$R^2_{generalized} = 1 - \frac{\hat{\boldsymbol{e}}'\boldsymbol{V}^{-1}\hat{\boldsymbol{e}}}{(\boldsymbol{Y} - \bar{Y})\boldsymbol{V}^{-1}(\boldsymbol{Y} - \bar{Y})},$$

where e = $(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$ and $\bar{Y} = \dfrac{j_m' \boldsymbol{V}^{-1} \boldsymbol{Y}}{j_m' \boldsymbol{V}^{-1} j_m'}$ and j_m is the column vector of dimension (m $\times$ 1) of all ones.

Mean Squared Error

The mean squared error (MSE) is computed as $\frac{1}{m} \sum_{i=1}^{m} (y_i - E(y_i|\theta_i)^2)$. Sometimes it is convenient to calculate the MSE inversely weighted by the variance: $\frac{1}{m} \sum_{i=1}^{m} \left(y_i - E(y_i|\theta_i)^2 / var(y_i|\theta_i)\right)$. These can be difficult to estimate, however, and they have the disadvantage of being less suitable for non-normal models, see Gelman, Hwang, and Vehtari (2013).

Log predictive density (log-likelihood)

The log predictive density, or log-likelihood, log(p($y_i|\theta_i$)) is somewhat more general than the mean-squared error and plays an important role in statistics due its prominence in the Kullback-Liebler information. This value is concordant with the posterior probability and is therefore sensible as a goodness of fit statistic. The prior distribution is typically ignored when attempting to find a model that matches with the existing data. In fact, the limiting distribution of the posterior normal, and the likelihood approaches the same asymptotic distribution, see DeGroot (1970). A convenient goodness of fit statistic is the log-predictive density of new sample points y_0. The out-of-sample prediction is,

$$\log p_{post}(y_0) = log E_{post}(p(y_0|\theta)) = log \int p(y_0|\theta) p_{post}(\theta) d\theta,$$

where $p_{post}(y_0)$ denotes the predictive density for y_0 from the posterior distribution p_{post}. Similarly, $E_{post}(y_0)$ will define the expectation of θ averaged over the posterior distribution. Since we will use all available data in the estimation, the new data point y_0 is unknown and we proceed to condition

upon all possible new values of y_i. We use the acronym **elpd** to denote the expected log predictive density,

$$\text{elpd} = E_f(log p_{post}(y_0)) = \int (log p_{post}(y_0) f(y_0) dy_0.$$

Akaike Information Criterion (AIC)

Despite recent inroads in the prominence of Bayesian methods and the use of posterior distributions to describe unknown parameters, frequentist methods have dominated the statistical literature. Most notably, the MLE is typically taken for $\hat{\theta}$, and $\log p(y_i|\hat{\theta}_i)$. For a model estimating k parameters, the AIC was developed by Aikake (1973) and take the following form based on the elpd:

$$\text{AIC} = \text{-2}\log p(y_i|\hat{\theta}_i) + \text{2k}.$$

The Bayesian information criterion was suggested by Schwartz (1978) is closely related to AIC and is defined by $\text{-2}\log p(y_i|\hat{\theta}_i) + \text{k log(n)}$. Note the sensitivity of the measure on the overall sample size. While the coefficient of determination and the mean squared error can objectively inform on the overall model fit, the AIC and elpd are more useful when comparing competing models. Gelman et al. (2013) describes the deviance information criteria as a function of $\log p(y_i|\tilde{\theta}_i)$, where $\tilde{\theta}_i$ is the posterior mean and using the *effective* number of parameters instead of k.

Watanabe-Akaike Information Criterion (WAIC)

Gelman, Hwang, and Vehtari (2013) describe a couple variations of the WAIC, first introduced by Watanabe (2010). This technique is more faithful to Bayesian philosophy and is based on posterior distributions:

$$WAIC_1 = \sum_{i=1}^{m} log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta_i^s)\right) \text{-}\ 2\sum_{i=1}^{n}\left(log(E_{post}p(y_i|\theta_i) - E_{post}(log p(y_i|\theta_i))\right),$$

where S draws are taken from the posterior distribution resulting in θ_i^s for s=1, 2, . . . , S.

Bayesian Cross Validation.

Cross-validation is the procedure of splitting the dataset into two portions: the training dataset and the analysis dataset. The training data are used for fitting the model, and the goodness of that model fit is evaluated on the remaining data. Our evaluation shall rely upon the use of the *posterior predictive distribution*. If y_{obs} denotes observed samples $y_1, y_2, \ldots, y_m$, then let y_{new} correspond to k new samples beyond what has already been observed: $y_{new} = y_{m+1}, y_{m+2}, \ldots, y_{m+k}$. Letting $f(\cdot)$ denote the probability density function of its argument, then the posterior predictive distribution is given by:

$$f(y_{new}|y_{obs}) = \frac{f(y)}{f(y_{obs})} = \int f(y_{new}|\theta)p(\theta|y_{obs})d\theta,$$

where $p(\cdot|\cdot)$ is the posterior distribution of the first argument conditional on the second. This can be compared with the *prior* predictive distribution (or marginal distribution), $f(y_{new}) = \int f(y_{new}|\theta)\pi(\theta)d\theta$, where $\pi(\cdot)$ is the prior distribution. Unlike the marginal distribution, $f(y_{new}|y_{obs})$ is based on the data at hand. Note that this metric does not suffer any bias from performing model evaluation on any portion of the training data (for example as with $f(y_i|y)$). The predictive distribution can tell us how well we are able to predict new observations.

Leave One Out Cross Validation.

In practice, we are unwilling to sacrifice any such data points to be consigned purely for validation

purposes. One computationally intensive variation of cross-validation is to leave out a single observation, fit the model on the remaining data, and apply that model to the data point left out. This form of cross-validation is referred to as "Leave One Out", LOO, or even LOO-CV. The goodness of the fit metric is averaged over all m evaluations. Let $y_{(-i)}$ denote the vector of y values *without* the i^{th} value y_i, that is, $y_{(-i)} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)'$. Then the predictive distribution is given by

$$f(y_i|\boldsymbol{y}_{(-i)}) = \frac{f(y)}{f(\boldsymbol{y}_{(-i)})} = \int f(y_i|\theta)p(\theta|\boldsymbol{y}_{(-i)})d\theta.$$

Carlin & Louis (2009) refer to the specific values of $f(y_i|y_{(-i)})$ as the CPO's, or conditional predictive ordinates. It should be pointed out that this is a very data-driven technique, that requires knowledge of the posterior distribution. This versatile model diagnostic is built into the Stan platform (https://mc-stan.org/), which is a collection of statistical modeling and calculation libraries that can be used with a variety of statistical software packages.

For dealing with samples with joint probability, it is customary to consider the log transformation. The goodness of fit metric used is the expected log pointwise predictive density, elpd:

$$\text{elpd}_{loo} = \sum_{i=1}^{m} \log f(y_i|\boldsymbol{y}_{(-i)}) = \sum_{i=1}^{m} \log \int f(y_i|\theta)p(\theta|\boldsymbol{y}_{(-i)})d\theta,$$

where $f(\cdot|\cdot)$ is the conditional probability density function of its arguments, and $p(\theta_i|\boldsymbol{y}_i)$ is the posterior density function of θ_i conditional on the data $\boldsymbol{y}_{(-i)}$. While computationally expensive, LOO-CV is more flexible than other information criteria. Under the maximum likelihood estima-

tion, it has been shown that it is asymptotically equivalent to AIC (Stone (1977)). Both DIC and WAIC have also been shown to be asymptotically equal to the LOO-CV under certain conditions (see Shibata (1989) and Watanabe (2010)). We will now discuss the LOO-CV more in depth and apply it to the Fay-Herriot model.

Bayesian Residual Analysis and Diagnostics.

Once a model has been decided upon, we are interested in evaluating the suitability of model fit to the data, without any abstract concepts. To this end, Carlin & Louis (2009) define a *Bayesian* residual in the context of LOO-CV as

$$r_i = y_i - E(Y_i|y_{(-i)}).$$

Plotting these residuals against fitted values can be used to verify the plausibility of normality and homoskedasticity. Time plots will reveal non-constant variation in time or other time dependencies. As with the CPO's defined above, these residuals can be plotted and used to identify outliers which the model does not predict well. The summed absolute values or squares might be used to develop a competing goodness of fit measure. For these latter two diagnostics, it is customary to follow standard linear regression diagnostics and consider a standardized residual:

$$d_i = \frac{y_i - E(Y_i|\boldsymbol{y}_{(-i)})}{\sqrt{Var(Y_i|\boldsymbol{y}_{(-i)})}}.$$

In the presence of a covariate vector x_i, Gelman et al. (2004) define Bayesian residuals as r_i

= y_i - $E(y_i|x_i, \theta_i)$. These are sometimes called "realized" residuals because they are based on a random draw of θ. In contrast, traditional design-based residuals can be thought of as being based on point estimates $\hat{\theta}_i$, resulting in residuals defined by y_i - $E(y_i|x_i, \hat{\theta}_i)$.

3.3 Empirical Bayes Parametric Bootstrap Cross-Validation Model Selection

Recently, Lahiri (2020) proposed a small area model selection methodology, which may be viewed as an empirical Bayes cross-validation implemented through a parametric bootstrap. We develop and evaluate these ideas for the well-known Fay-Herriot model.

Let y_i given θ_i, $i = 1, \cdots, m$, be independent where both y_i's and θ_i are scalars. This is a general model that could cover both the Fay-Herriot model and Rao-Yu model as special cases. Let $y_{(-i)}$ be a vector of y except the ith area, $i = 1, \cdots, m$. We propose the following parametric bootstrap LOO:

Definition 3.1. Parametric Bootstrap Leave-One-Out Cross Validation.

$$\text{LOO}_{\text{boot}} = \log \left\{ \prod_{i=1}^{m} f_{boot}(y_i|y_{(-i)}) \right\} = \sum_{i=1}^{m} \log f_{boot}(y_i|y_{(-i)}), \tag{3.1}$$

where

$$f_{boot}(y_i|y_{(-i)}) = \int f(y_i|\theta^*) p(\theta^*|y_{(-i)}) d\theta^*, \tag{3.2}$$

is the leave-one-out predictive parametric bootstrap density given the data without the i^{th} data

point. In the above expression, $p(\theta^*|y_{(-i)})$ is the parametric bootstrap distribution of θ^* based on $y_{(-i)}$. We can compute (3.1) for a set of competing models and select the one for which LOO_{boot} is the maximum. Model selection can address normality, homoskedasticity in Level 2, variable selection, and more. To produce some formulas for the LOO method, let us assume normality and homoskedasticity and focus on variable selection. First consider the easy case when all the hyperparameters, i.e., β and A are known. Since under the FH model all y_i's are unconditionally independent, $f_{boot}(y_i|y_{(-i)})$ is the $N(x_i'\beta, \psi_i + A)$ density, $i = 1, \cdots, m$. That is,

$$f_{boot}(y_i|y_{(-i)}) = \frac{1}{\sqrt{2\pi}\sqrt{\psi_i + A}} \exp\left\{-\frac{1}{2(\psi_i + A)}(y_i - x_i'\beta)^2\right\}.$$

Thus, when the parameters are known there is no additional information housed in $y_{(-i)}$. Now let's consider the case when hyperparameters are unknown and estimated using some classical method like ML or REML or adjusted ML. Suppose that A is known but β is unknown. For this case,

$$f_{boot}(y_i|y_{(-i)}) = \frac{1}{\sqrt{2\pi}\sqrt{\psi_i + A}} E_{*(-i)}\left[\exp\left\{-\frac{1}{2(\psi_i + A)}(y_i - x_i'\hat{\beta}^*_{(-i)})^2\right\}\right], \tag{3.3}$$

where E_* is expectation with respect to the bootstrap distribution,

$$\hat{\beta}^*_{(-i)} \sim N_p\left(\hat{\beta}_{(-i)},\ \Sigma_i\right),$$

where $\hat{\beta}^*_{(-i)}$ is the weighted least squares estimator of β without the y_i, or x_i, and $Var(\hat{\beta}_{(-i)}) = \Sigma_i = \left(\sum_{j\neq i} \frac{1}{\psi_j + A} x_j x_j'\right)^{-1}$. For completeness, we will obtain some closed form results. Our computations will require the following lemmas:

Lemma 3.2. Sherman-Morrison Formula.

Let M be an invertible matrix of size n, and v is a column vector of dimension n. Then,

$$(M - vv')^{-1} = M^{-1} + \frac{M^{-1}vv'M^{-1}}{1 - v'M^{-1}v}.$$

Theorem 3.3. Fay-Herriot Bootstrap Distribution, A known and β unknown.

When A is known but β is not known, then the bootstrap distribution (3.3) of y_i, conditional upon the data excluding y_i, is given by:

$$f_{boot}(y_i|y_{(-i)}) = \frac{\sqrt{1 - \frac{x_i'\left[\sum_{j=1}^{m} \frac{x_j x_j'}{\psi_j + A}\right]^{-1} x_i}{\psi_i + A}}}{\sqrt{2\pi}\sqrt{\psi_i + A}} \exp\left\{ -\frac{1 - \frac{x_i'\left[\sum_{j=1}^{m} \frac{x_j x_j'}{\psi_j + A}\right]^{-1} x_i}{\psi_i + A}}{2(\psi_i + A)} (y_i - x_i'\hat{\beta}_{(-i)})^2 \right\} \tag{3.4}$$

Proof. Evaluating the distribution in (3.16) analytically, treating all but $\hat{\beta}^*_{(-i)}$ fixed, leads to:

$$
\begin{aligned}
&E_{*(-i)}\left[\exp\left\{-\frac{1}{2(\psi_i+A)}(y_i - x_i'\hat{\beta}^*_{(-i)})^2\right\}\right] \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_i|^{1/2}}\int\left[\exp\left\{-\frac{1}{2(\psi_i+A)}(y_i - x_i'\hat{\beta}^*_{(-i)})^2\right\}\right. \\
&\qquad \left.\times\exp\left\{-\frac{1}{2}(\hat{\beta}^*_{(-i)} - \hat{\beta}_{(-i)})'\Sigma_i^{-1}(\hat{\beta}^*_{(-i)} - \hat{\beta}_{(-i)})\right\}d\hat{\beta}^*_{(-i)}\right. \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2(\psi_i+A)}(y_i^2) - \frac{1}{2}\hat{\beta}'_{(-i)}\boldsymbol{\Sigma}_i^{-1}\hat{\beta}_{(-i)}\right\}\int\exp\left\{\frac{y_i x_i'\hat{\beta}^*_{(-i)}}{\psi_i+A} - \frac{1}{2}\frac{\hat{\beta}^*_{(-i)}{}' x_i x_i'\hat{\beta}^*_{(-i)}}{\psi_i+A}\right\} \\
&\qquad \times\exp\left\{\hat{\beta}'_{(-i)}\Sigma_i^{-1}\hat{\beta}^*_{(-i)} - \frac{1}{2}\hat{\beta}^*_{(-i)}{}'\Sigma_i^{-1}\hat{\beta}^*_{(-i)}\right\}d\hat{\beta}^*_{(-i)} \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2(\psi_i+A)}(y_i^2) - \frac{1}{2}\hat{\beta}'_{(-i)}\boldsymbol{\Sigma}_i^{-1}\hat{\beta}_{(-i)}\right\}\int\exp\left\{\left(\frac{y_i x_i'}{\psi_i+A} + \hat{\beta}'_{(-i)}\Sigma_i^{-1}\right)\hat{\beta}^*_{(-i)}\right\} \\
&\qquad \times\exp\left\{-\frac{1}{2}\hat{\beta}^*_{(-i)}{}'\left(\frac{x_i x_i'}{\psi_i+A} + \Sigma_i^{-1}\right)\hat{\beta}^*_{(-i)}\right\}d\hat{\beta}^*_{(-i)}
\end{aligned}
\tag{3.5}
$$

$$
\begin{aligned}
&= \frac{|\frac{x_i x_i'}{\psi_i+A} + \Sigma_i^{-1}|^{-1/2}}{|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2(\psi_i+A)}(y_i^2) - \frac{1}{2}\hat{\beta}'_{(-i)}\boldsymbol{\Sigma}_i^{-1}\hat{\beta}_{(-i)}\right\} \\
&\qquad \times\exp\left\{\left(\frac{1}{2}\frac{y_i x_i'}{\psi_i+A} + \hat{\beta}'_{(-i)}\Sigma_i^{-1}\right)\left[\frac{x_i x_i'}{\psi_i+A} + \Sigma_i^{-1}\right]^{-1}\left(\frac{y_i x_i}{\psi_i+A} + \Sigma_i^{-1}\hat{\beta}_{(-i)}\right)\right\}.
\end{aligned}
\tag{3.6}
$$

where we recognize the integral in the last line of (3.5) as that of the moment generating function of a multivariate normal random variable $\hat{\beta}^*$ with mean zero and covariance matrix $\left(\frac{x_i x_i'}{\psi_i+A}\right)^{-1} + \Sigma_i^{-1}$, along

with vector $\boldsymbol{t}' = (\frac{y_i x_i'}{\psi_i + A} + \hat{\beta}'_{(-i)}\Sigma_i^{-1})$. Then substituting $\Sigma^{-1} = \frac{x_i x_i'}{\psi_i + A} + \Sigma_i^{-1} = \frac{x_i x_i'}{\psi_i + A} + \sum_{j \neq i}^{m} \frac{x_i x_i'}{\psi_j + A^*_{(-i)}} = \sum_{i=1}^{m} \frac{x_i x_i'}{\psi_i + A}$, and then plugging this value into the covariance matrix of (3.6) yields

$$E_{*(-i)}\left[\exp\left\{-\tfrac{1}{2(\psi_i+A)}(y_i - x_i'\hat{\beta}^*_{(-i)})^2\right\}\right]$$

$$= \tfrac{|\Sigma|^{1/2}}{|\Sigma_i|^{1/2}} \exp\left\{-\tfrac{1}{2(\psi_i+A)}(y_i^2) - \tfrac{1}{2}\hat{\beta}'_{(-i)}\boldsymbol{\Sigma}_i^{-1}\hat{\beta}_{(-i)}\right\} \times \exp\left\{(\tfrac{1}{2}\tfrac{y_i x_i'}{\psi_i + A} + \hat{\beta}'_{(-i)}\Sigma_i^{-1})\Sigma(\tfrac{y_i x_i}{\psi_i + A} + \Sigma_i^{-1}\hat{\beta}_{(-i)})\right\}.$$

Now substitute $\sigma^2 = \frac{\psi_i + A}{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}}$. Note that this will represent a covariance matrix of higher magnitude than $(\sum_{j \neq i}^{m} \frac{1}{\psi_j + A} x_j x_j')^{-1})$ which is the original variance of y_i given all of the remaining data: $y_i|\boldsymbol{y}_{(-i)}$, with $\boldsymbol{y}_{(-i)} = \{y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m\}'$. This makes sense, since there is slightly less data available for prediction when the i^{th} observation is not used. By the Sherman-Morrison Formula, we see that $\Sigma_i = (\Sigma^{-1} - \frac{x_i x_i'}{\psi_i + A})^{-1} = \Sigma + \frac{\Sigma \frac{x_i x_i'}{\psi_i + A}\Sigma}{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}}$, so that $\Sigma_i\Sigma^{-1} = 1 + \frac{\frac{x_i x_i'}{\psi_i + A}\Sigma}{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}} = \frac{1}{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}}$. This leads to the representation

$$E_{*(-i)}\left[\exp\left\{-\tfrac{1}{2(\psi_i+A)}(y_i - x_i'\hat{\beta}^*_{(-i)})^2\right\}\right]$$

$$= \sqrt{1 - \tfrac{x_i'\Sigma x_i}{\psi_i + A}} \exp\left\{-\tfrac{1}{2}\hat{\beta}'_{(-i)}\boldsymbol{\Sigma}_i^{-1}\hat{\beta}_{(-i)}\right\} \times \exp -\tfrac{1}{2}\left\{(\tfrac{y_i^2}{\sigma^2} - \tfrac{2y_i}{\sigma^2}\tfrac{x_i'\Sigma\Sigma_i^{-1}\hat{\beta}_{(-i)}}{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}} - \hat{\beta}'_{(-i)}\Sigma_i^{-1}\Sigma\,\Sigma_i^{-1}\hat{\beta}_{(-i)}\right\}.$$

Recalling that $\Sigma_i^{-1} = \Sigma^{-1} - \frac{x_i x_i'}{\psi_i + A}$, and observing that $\Sigma\Sigma_i^{-1} = \Sigma(\Sigma^{-1} - \frac{x_i x_i'}{\psi_i + A}) = (1 - \frac{\Sigma x_i x_i'}{\psi_i + A})$ which leads us to,

$$E_{*(-i)}\left[\exp\left\{-\tfrac{1}{2(\psi_i+A)}(y_i - x_i'\hat{\beta}^*_{(-i)})^2\right\}\right] =$$

$$= \sqrt{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}} \exp\left\{-\frac{1}{2}\hat{\beta}'_{(-i)}\boldsymbol{\Sigma}_i^{-1}\hat{\beta}_{(-i)}\right\}$$
$$\times \exp -\frac{1}{2}\left\{(\frac{y_i^2}{\sigma^2} - \frac{2y_i}{\sigma^2}x_i'\hat{\beta}_{(-i)} - \hat{\beta}'_{(-i)}[\Sigma^{-1} - \frac{x_i x_i'}{\psi_i + A}]\Sigma\Sigma_i^{-1}\hat{\beta}_{(-i)}\right\} \tag{3.7}$$

$$= \sqrt{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}} \exp\left\{-\frac{1}{2}\hat{\beta}'_{(-i)}\boldsymbol{\Sigma}_i^{-1}\hat{\beta}_{(-i)}\right\}$$
$$\times \exp -\frac{1}{2}\left\{(\frac{y_i^2}{\sigma^2} - \frac{2y_i}{\sigma^2}\frac{(x_i' - \frac{x_i'\Sigma x_i x_i'}{\psi_i + A})\hat{\beta}_{(-i)}}{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}} - \hat{\beta}_{(-i)}\left[(I - \frac{x_i x_i'\Sigma}{\psi_i + A})\Sigma_i^{-1}\right]\hat{\beta}_{(-i)}\right\}$$

$$= \sqrt{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}} \exp\left\{-\frac{1}{2}\hat{\beta}'_{(-i)}\boldsymbol{\Sigma}_i^{-1}\hat{\beta}_{(-i)}\right\}$$
$$\times \exp -\frac{1}{2}\left\{(\frac{y_i^2}{\sigma^2} - \frac{2y_i}{\sigma^2}x_i'\hat{\beta}_{(-i)} - \hat{\beta}'_{(-i)}\left[\Sigma_i^{-1} - \frac{x_i x_i'}{\psi_i + A}(1 - \frac{x_i'\Sigma x_i}{\psi_i + A})\right]\hat{\beta}_{(-i)}\right\} \tag{3.8}$$

$$= \sqrt{1 - \frac{x_i'\Sigma x_i}{\psi_i + A}} \exp -\frac{1}{2}\left\{(\frac{y_i^2}{\sigma^2} - \frac{2y_i}{\sigma^2}x_i'\hat{\beta}_{(-i)} + \frac{1}{\sigma^2}\hat{\beta}'_{(-i)}x_i x_i'\hat{\beta}_{(-i)}\right\}.$$

Thus, we see that $f(y_i|y_{(-i)}) = \frac{1}{\sqrt{2\pi(\psi_i+A)}}\sqrt{1 - \frac{x_i'\Sigma x_i}{\psi_i+A}} \exp -\frac{1}{2}\left\{(\frac{y_i^2}{\sigma^2} - \frac{2y_i}{\sigma^2}x_i'\hat{\beta}_{(-i)} + \frac{1}{\sigma^2}\hat{\beta}'_{(-i)}x_i x_i'\hat{\beta}_{(-i)}\right\}$
$\implies (y_i|y_{(-i)}) \sim N\left(x_i'\hat{\beta}_{(-i)}, \frac{\psi_i+A}{1-\frac{x_i'\Sigma x_i}{\psi_i+A}}\right)$. □

It is instructive to verify this result. Clearly $Ey_i = E\{E(y_i|y_{(-i)})\} = E\{x_i\hat{\beta}_{(-i)}\} = x_i\beta$, which is consistent with the Fay-Herriot model. We also have that $Var(y_i) = E\{Var(y_i|y_{(-i)})\} + Var\{E(y_i|y_{(-i)})\} = E\left\{\frac{\psi_i+A}{1-\frac{x_i'\Sigma x_i}{\psi_i+A}}\right\} + Var(x_i'\hat{\beta}_{(-i)}) = \frac{\psi_i+A}{1-\frac{x_i'\Sigma x_i}{\psi_i+A}} + x'\left[\Sigma + \frac{\Sigma x_i x_i'\Sigma/(\psi_i+A)}{1-\frac{x_i'\Sigma s_i}{\psi_i+A}}\right]x_i = \psi_i + A$, where we have used Lemma 3.2 again.

Example 3.4. Fay-Herriot Parametric Bootstrap Distribution, A unknown and β unknown.

For the balanced case $\psi_i = \psi,\ i = 1, \cdots, m$, we can proceed to obtain the exact distribution of $(\hat{\beta}^*_{(-i)}, \hat{A}^*_{(-i)})$. Now suppose that $\hat{A}^*_{(-i)}$ corresponded to the maximum likelihood estimator obtained from the dataset missing the ith data point. The asymptotic variance $\bar{V}$ for the Fay-Herriot model as $m \to \infty$ for the ML and REML are equal, and is given in Rao and Molina (2015) by:

$$\bar{V}(\hat{A}_{ML}) = \bar{V}(\hat{A}_{REML}) = 2\left[\sum_{i=1}^{m}(A + \psi_i)^{-1}\right]^{-1}.$$

Therefore the "plug-in" bootstrap distribution $\hat{A}^*_{(-i)}$ could plausibly be given by:

$$\hat{A}^*_{(-i)} \dot{\sim} N\left(\hat{A}_{(-i)},\ 2(\sum_{j\neq i}^{m}\frac{1}{\psi_j + \hat{A}_{(-i)}})^{-1}\right),$$

and with A unknown, the bootstrap distribution of $\hat{\beta}^*_{(-i)}$ is now expressed as:

$$\hat{\beta}^*_{(-i)} \mid \hat{A}_{(-i)} \dot{\sim} N_p\left(\hat{\beta}_{(-i)},\ (\sum_{j\neq i}^{m}\frac{1}{\psi_j + \hat{A}_{(-i)}}x_j x_j')^{-1}\right). \tag{3.9}$$

When both β and A are unknown, we have using Theorem 3.3,

$$
\begin{aligned}
f_{boot}(y_i|y_{(-i)}) &= E_{*(-i)}\left[\frac{1}{\sqrt{2\pi}\sqrt{\psi_i+\hat{A}^*_{(-i)}}}\exp\left\{-\frac{1}{2(\psi_i+\hat{A}^*_{(-i)})}(y_i-x_i'\hat{\beta}^*_{(-i)})^2\right\}\right] \\
&= E_{*(-i)}\left\{E\left[\frac{1}{\sqrt{2\pi}\sqrt{\psi_i+\hat{A}^*_{(-i)}}}\exp\left(-\frac{1}{2(\psi_i+\hat{A}^*_{(-i)})}(y_i-x_i'\hat{\beta}^*_{(-i)})^2\right)\middle| A=\hat{A}^*_{(-i)}\right]\right\} \\
&= E_{*(-i)}\left\{\frac{1}{\sqrt{2\pi(\psi_i+\hat{A}^*_{(-i)})}}\sqrt{1-\frac{x_i'\Sigma x_i}{\psi_i+\hat{A}^*_{(-i)}}}\exp\left(-\frac{1}{2\sigma^2(\hat{A}^*_{(-i)})}\left\{(y_i-x_i'\hat{\beta}_{(-i)})^2\right\}\right)\right\} \\
&= \int\left\{\frac{1}{\sqrt{2\pi(\psi_i+\hat{A}^*_{(-i)})}}\sqrt{1-\frac{x_i'\Sigma x_i}{\psi_i+\hat{A}^*_{(-i)}}}\exp\left(-\frac{1}{2\sigma^2(\hat{A}^*_{(-i)})}(y_i-x_i'\hat{\beta}_{(-i)})^2\right)\right\}dF(\hat{A}^*_{(-i)})
\end{aligned}
\tag{3.10}
$$

$$
\begin{aligned}
&= \int\frac{1}{\sqrt{2\pi(\psi_i+\hat{A}^*_{(-i)})}}\sqrt{1-\frac{x_i'\Sigma x_i}{\psi_i+\hat{A}^*_{(-i)}}}\exp\left(-\frac{1}{2\sigma^2(\hat{A}^*_{(-i)})}(y_i-x_i'\hat{\beta}_{(-i)})^2\right) \\
&\times\frac{1}{\sqrt{2\pi}}\Big(\sum_{j\neq i}(\psi_i+\hat{A}_{(-i)})^{-1}\Big)^{1/2}\exp\left\{-\frac{1}{2}\left(\sum_{j\neq i}^{m}\frac{1}{\psi_j+\hat{A}_{(-i)}}\right)(\hat{A}^*_{(-i)}-\hat{A}_{(-i)})^2\right\}d\hat{A}^*_{(-i)}
\end{aligned}
\tag{3.11}
$$

where $\sigma^2=\sigma^2(\hat{A}^*_{(-i)})=\frac{\psi_i+\hat{A}^*_{(-i)}}{1-\frac{x_i'\Sigma x_i}{\psi_i+\hat{A}^*_{(-i)}}}$. Unfortunately, the integral in equation (3.11) is somewhat intractable. Because of the non-negativity of variance parameters, many analysts will use a non-normal distribution for when placing prior distributions on variance components in fully Bayesian models. In fact, the inverted gamma distribution $IG(\alpha,\delta)$ with parameters α and δ is a popular prior

distribution. We will now consider this distribution as an alternative to the Gaussian distribution for the parametric bootstrap. Recall that the probability density function for a $IG(\alpha, \delta)$ random variable u is given by $f(u;\alpha,\delta) = \frac{\delta^\alpha}{\Gamma(\alpha)} u^{-(\alpha+1)} \exp(-\frac{\delta}{u})$ with mean and variance $\frac{\delta}{\alpha-1}$ and $\frac{\delta^2}{(\alpha+1)^2(\alpha+2)}$, respectively, for $\alpha > 2$. Hence, if we wanted to select the parameters α and δ so that a generated $IG(\alpha, \delta)$ random variable had a distribution matching the mean and variance of $\hat{A}_{(-i)}$, then we find there are two equations and two solutions, yielding:

$$\begin{aligned} \alpha &= \hat{A}_{(-i)}\Big(\sum_{j=1}^{m} \frac{1}{\psi_i + \hat{A}_{(-i)}}\Big) + 2, \\ \delta &= (\hat{A}_{(-i)})^2\left(\Big(\sum_{j=1}^{m} \frac{1}{\psi_i + \hat{A}_{(-i)}}\Big) + 1\right). \end{aligned} \tag{3.12}$$

We now consider the following bootstrap distribution for $\hat{A}^*_{(-i)}$:

$$\hat{A}^*_{(-i)} \sim IG\,(\alpha;\delta) = IG\left(\hat{A}_{(-i)}\Big(\sum_{j=1}^{m} \frac{1}{\psi_i + \hat{A}_{(-i)}}\Big); (\hat{A}_{(-i)})^2\left(\Big(\sum_{j=1}^{m} \frac{1}{\psi_i + \hat{A}_{(-i)}}\Big) + 1\right)\right).$$

Example 3.5. Fay-Herriot Parametric Bootstrap Distribution, A unknown and β unknown.

Under the inverse-gamma bootstrap distribution for $\hat{A}_{(-i)}$, and the bootstrap distribution for

$\hat{\beta}^*_{(-i)}$ displayed in (4.1), then Example 3.4 asserts:

$$f_{boot}(y_i|y_{(-i)}) = E_{*(-i)}\left[\frac{1}{\sqrt{2\pi}\sqrt{\psi_i + \hat{A}^*_{(-i)}}}\exp\left\{-\frac{1}{2(\psi_i + \hat{A}^*_{(-i)})}(y_i - x_i'\hat{\beta}^*_{(-i)})^2\right\}\right]$$

$$= \int \frac{1}{\sqrt{2\pi(\psi_i + \hat{A}^*_{(-i)})}}\sqrt{1 - \frac{x_i'\Sigma x_i}{\psi_i + \hat{A}^*_{(-i)}}}\exp\left(-\frac{1}{2\sigma^2(\hat{A}^*_{(-i)})}(y_i - x_i'\hat{\beta}_{(-i)})^2\right)d(\hat{A}^*_{(-i)}) \tag{3.13}$$

$$= \int_{-\psi_i}^{\infty} \frac{1}{\sqrt{2\pi(\psi_i + \hat{A}^*_{(-i)})}}\sqrt{1 - \frac{x_i'\Sigma x_i}{\psi_i + \hat{A}^*_{(-i)}}}\exp\left(-\frac{1}{2\sigma^2(\hat{A}^*_{(-i)})}(y_i - x_i'\hat{\beta}_{(-i)})^2\right)$$

$$\times \frac{\delta^\alpha}{\Gamma(\alpha)}\left(\hat{A}^*_{(-i)} + \psi_i\right)^{-(\alpha+1)}\exp\left(-\frac{\delta}{\hat{A}^*_{(-i)} + \psi_i}\right)d\hat{A}^*_{(-i)}, \tag{3.14}$$

where $\sigma^2 = \sigma^2(\hat{A}^*_{(-i)}) = \frac{\psi_i + \hat{A}^*_{(-i)}}{1 - \frac{x_i'\Sigma x_i}{\psi_i + A^*_{(-i)}}}$. and α and δ are defined as in (4.8). Again, the integral in (4.2) is intractable, particularly because of $\sigma^2(\hat{A}^*_{(-i)})$ and the use of the matrix $\Sigma = \Sigma(A) = \sum_{j=1}^{m}\frac{x_j x_j'}{\psi_i + A}$. We now appeal to the balance assumption, and so that $\psi_0 = \psi_j \equiv \psi_i \forall j \neq i$. In this case, $\sigma^2 = \sigma^2(\hat{A}^*_{(-i)}) = \psi_i + \hat{A}^*_{(-i)} * \left(1 - x_i'\left(\sum_{j=1}^{m} x_j x_j'\right)^{-1} x_i\right)^{-1}$. This assumption lets us proceed to get a closed form expression for the leave-one-out parametric bootstrap distribution,

$$
\begin{aligned}
f_{boot}(y_i|y_{(-i)}) &= E_{*(-i)}\left[\frac{1}{\sqrt{2\pi}\sqrt{\psi_0+\hat{A}^*_{(-i)}}}\exp\left\{-\frac{1}{2(\psi_0+\hat{A}^*_{(-i)})}(y_i-x_i'\hat{\beta}^*_{(-i)})^2\right\}\right] \\
&= \int_{-\psi_0}^{\infty}\frac{1}{\sqrt{2\pi(\psi_0+\hat{A}^*_{(-i)})}}\sqrt{1-x_i'\left(\sum_{j=1}^m x_jx_j'\right)^{-1}x_i} \\
&\times\exp\left\{-\frac{1-x_i'\left(\sum_{j=1}^m x_jx_j'\right)^{-1}x_i}{2(\psi_0+\hat{A}^*_{(-i)})}(y_i-x_i'\hat{\beta}_{(-i)})^2\right\} \\
&\times\frac{\delta^\alpha}{\Gamma(\alpha)}\left(\psi_0+\hat{A}^*_{(-i)}\right)^{-(\alpha+1)}\exp\left(-\frac{\delta}{\psi_0+\hat{A}^*_{(-i)}}\right)d\hat{A}^*_{(-i)} \\
&= \frac{\delta^\alpha}{\Gamma(\alpha)}\sqrt{1-x_i'\left(\sum_{j=1}^m x_jx_j'\right)^{-1}x_i}\int_{-\psi_0}^{\infty}\frac{1}{\sqrt{2\pi}}\left(\psi_0+\hat{A}^*_{(-i)}\right)^{-(\alpha+\frac{3}{2})} \qquad (3.15) \\
&\times\exp\left\{-\frac{1-x_i'\left(\sum_{j=1}^m x_jx_j'\right)^{-1}x_i}{2(\psi_0+\hat{A}^*_{(-i)})}(y_i-x_i'\hat{\beta}_{(-i)})^2\right\}\times\exp\left(-\frac{\delta}{\psi_0+\hat{A}^*_{(-i)}}\right)d\hat{A}^*_{(-i)} \\
&= E_{*(-i)}\left[\frac{1}{\sqrt{2\pi}\sqrt{\psi_0+\hat{A}^*_{(-i)}}}\exp\left\{-\frac{1}{2(\psi_0+\hat{A}^*_{(-i)})}(y_i-x_i'\hat{\beta}^*_{(-i)})^2\right\}\right] \\
&= \frac{1}{\sqrt{2\pi}}\frac{\delta^\alpha}{\Gamma(\alpha)}\sqrt{1-x_i'\left(\sum_{j=1}^m x_jx_j'\right)^{-1}x_i}\int_{-\psi_0}^{\infty}\left(\psi_0+\hat{A}^*_{(-i)}\right)^{-(\alpha+\frac{3}{2})} \\
&\times\exp\left\{-\frac{1}{2}\frac{1}{\psi_i+\hat{A}^*_{(-i)}}\left(2\delta+\left(1-x_i'\left(\sum_{j=1}^m x_jx_j'\right)^{-1}x_i\right)(y_i-x_i'\hat{\beta}_{(-i)})^2\right)\right\}d\hat{A}^*_{(-i)}.
\end{aligned}
$$

Let C = $2\delta + \left(1 - x_i'\left(\sum_{j=1}^{m} x_j x_j'\right)^{-1} x_i\right)(y_i - x_i'\hat{\beta}_{(-i)})^2\Big)$. Then,

$$f_{boot}(y_i|y_{(-i)}) = \frac{1}{\sqrt{2\pi}}\frac{\delta^\alpha}{\Gamma(\alpha)}\sqrt{1 - x_i'\left(\sum_{j=1}^{m} x_j x_j'\right)^{-1} x_i} \times \Gamma(\alpha + 3/2)/(\frac{C}{2})^{\alpha+3/2}.$$

Under the assumption of the balanced model, $\alpha = \frac{m\hat{A}_{(-i)}}{\psi_0+\hat{A}_{(-i)}}$ and $\delta = \hat{A}^2_{(-i)}\big(\frac{m}{\psi_0+\hat{A}_{(-i)}} + 1\big)$. Let us use the following notation for simplicity, $c_i = 1 - x_i'\left(\sum_{j=1}^{m} x_j x_j'\right)^{-1} x_i$. Plugging in all relevant values yields:

$$f_{boot}(y_i|y_{(-i)}) = \sqrt{c_i}\frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha)\Gamma(\frac{1}{2})}\left\{\frac{\hat{A}^2_{(-i)}(\frac{m}{\psi_0+\hat{A}_{(-i)}} + 1)}{\hat{A}^2_{(-i)}(\frac{m}{\psi_0+\hat{A}_{(-i)}} + 1) + c_i(y_i - x_i'\hat{\beta}_{(-i)})^2/2}\right\}^\alpha$$

$$\times\left(\hat{A}^2_{(-i)}(\frac{m}{\psi_0 + \hat{A}_{(-i)}} + 1) + c_i(y_i - x_i'\beta)^2/2\right)^{-3/2}.$$

3.4 Empirical Bayes Leave-One-out Cross-Validation Model Selection with Application to Small Area Income and Poverty Estimates

The parametric bootstrap can be used to assess the goodness of fit between competing models. As with other goodness of fit statistics (e.g., coefficient of determination, BIC), the LOO_{boot} could be used in stepwise variable selection (backward or forward) strategies, or best subsets approaches where the number of candidate variables is not extensive.

It is often the case that standard metrics used for variable selection will give fairly similar results. Although computationally expensive, the LOO-CV provides the highest amount of predictive

power. We propose the following methodology for variable selection:

Variable Selection Procedure using Leave One Out Cross Validation for Small Area Models

Consider the Fay-Herriot Area-Level Model for $i = 1, \cdots, m$,

1. Remove the i^{th} area from the dataset.

2. Compute $\hat{\beta}_{(-i)}$ and $\hat{A}_{(-i)}$, the estimates for the unknown model parameters based on the dataset without the i^{th} observation. The SAE package in R utilizes a default residual maximum likelihood estimator, but the standard maximum likelihood or any other favored estimator will suffice.

3. Generate $y^*_{(-i)} = (y^*_{1(-i)}, \cdots, y^*_{i-1(-i)}, y^*_{i+1(-i)}, \cdots, y^*_{m(-i)})'$, a copy of the dataset by sampling from a normal distribution for each state $j \neq i$, with mean $x'_j\hat{\beta}_{(-i)}$ and variance $\psi_j + \hat{A}_{(-i)}$. Repeat this process K times to obtain $\boldsymbol{y}^{*(1)}_{(-i)}, \boldsymbol{y}^{*(2)}_{(-i)}, \boldsymbol{y}^{*(K)}_{(-i)}$.

4. For each parametric bootstrap replicate k for the fifty remaining states, re-estimate the β and A to obtain $\hat{\beta}^{*(k)}_{(-i)}$ and $\hat{A}^{*k}_{(-i)}$ for $k = 1, \cdots, K$.. This should be the same estimator used in step 2 to obtain $\hat{\beta}_{(-i)}$ and $\hat{A}_{(-i)}$. Essentially, this requires using the SAE package repeatedly for different bootstrap replicates.

5. Plug $\hat{\beta}^*_{(-i)}$ and $\hat{A}^*_{(-i)}$ in to the likelihood function for y_i, which is calculated also using y_i, x_i, and ψ_i,

$$f_{boot}(y_i|y_{(-i)})$$
$$\approx \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{\sqrt{2\pi}\sqrt{\psi_i + \hat{A}^{*(k)}_{(-i)}}}\left[\exp\left\{-\frac{1}{2(\psi_i + \hat{A}^{*(k)}_{(-i)})}\left(y_i - x_i'\hat{\beta}^{*(k)}_{(-i)}\right)^2\right\}\right]\right), \tag{3.16}$$

6. Repeat steps 1-5 for all states to obtain

$$\text{LOO}_{boot} = \log\left\{\prod_{i=1}^{m} f_{boot}(y_i|y_{(-i)})\right\} = \sum_{i=1}^{m}\log f_{boot}(y_i|y_{(-i)}), \tag{3.17}$$

Application to SAIPE Dataset

The U.S. Census Bureau's Small Area Income and Poverty Estimates program (SAIPE) produces estimates at the county and state level to measure the total number of people (and children) in poverty along with median household income. The reader is asked to refer to Chapter 2 of this dissertation for details on the dataset. We briefly review the variable information, nevertheless. The following variables were included from SAIPE 1993 comprising poverty estimates for children between 5 and 17 for fifty states and the District of Columbia.

1. **CPS** - direct estimate from the Current Population for the state poverty rate of children ages 5-17

2. **IRSPR** - Poverty rate based on IRS tax data, defined as # Child tax exemptions from impoverished households / Total child tax exemptions from all households.

3. **IRSNF** - The tax non-filer rate based on IRS tax data, defined as [Population - # Tax Exemptions under Age 65] / Estimated Population of Persons 65+.

4. **FS** - Food Stamp participation rate as measured by the Supplemental Nutritional Assistance Program.

5. **GVFSE** - GVF estimates of sampling standard errors from the CPS. These are based on an iterative procedures developed by Otto and Bell (1995) that switches between ML estimates of model parameters vs. the estimation of sampling standard errors.

6. **CENRES** - Residuals from the 1990 census where the full model (with the same covariates) was fit to the same outcome of the poverty rate for children between ages 5 and 17.

The concept follows the approach to the model building described in Bell et al. (2016) and Erciulescu et al. (2020). For the parametric bootstrap, each state we left out and estimated using 200 bootstrap samples.

Table 3.6. Variable Selection Using EBPB LOO-CV.

Mode l Selection using Parametric Bootstrap, K=200			
Model Group	Model Variance	AIC	LOO_{boot}
CENRES FS IRSFR IRSPR	1.703	281.601	42.936
FS IRSFR IRSPR	3.061	290.188	49.162
CENRES IRSFR IRSPR	4.418	300.373	59.194
CENRES FS IRSPR	4.850	289.749	44.156
CENRES FS IRSFR	3.257	285.264	43.533
CENRES FS	5.468	289.568	42.671
CENRES IRSFR	24.310	339.345	76.160
CENRES IRSPR	9.741	311.074	66.894
FS IRSFR	3.449	290.033	49.532
FS IRSPR	6.051	295.332	49.732
IRSFR IRSPR	7.972	308.810	67.144
CENRES	30.631	345.437	85.132
FS	6.049	294.007	48.309
IRSFR	25.307	341.525	75.175
IRSPR	12.803	316.638	72.628

Results. The posterior predictive distribution based on the leave-one-out cross validation, represented by LOO_{boot} is mostly consistent with the model variance and AIC metrics in the table. Variable groups with lower predictive power have higher scores across all three metrics, and vice versa, better models have lower scores. This shows the approach is both reasonable and viable. There are some differences, however. The best model selected by LOO_{boot} was very close to that of AIC and the model variance. The comparison is somewhat clouded by possible multicollinearity. Nevertheless, there is evidence that the LOO_{boot} would produce distinct results from other methods.

3.5 Empirical Bayes Parametric Bootstrap Model Diagnostics

Once the best model is selected from among the class of viable models, we wish to further evaluate the model's suitability against the data. This is necessary because the model selection process really only compares viable models and procures the best one, but all of those models may still fit the underlying data rather poorly. As part of the diagnostic process, it is customary to perform a residual analysis to visually detect outliers and departures from model assumptions. The parametric bootstrap we have developed has a natural application to residual analysis. building on the concepts proposed in Lahiri (2020), we further examine empirical Bayes parametric model diagnostics for the Fay-Herriot model.

Definition 3.7. Empirical Bayes Parametric Bootstrap LOO Cross-Validatory Residual.

We define the EBPB LOO-CV cross-validatory residuals as:

$$r_i = y_i - E_*[y_i^*|y_{(-i)}], \tag{3.18}$$

where y_i is the observed value of the outcome variable for area i and the parametric bootstrap expectation E_* can be approximated as

$$E_*[y_i^*|y_{(-i)}] \approx \frac{1}{K}\sum_{k=1}^{K} y_i^{*(k)}, \tag{3.19}$$

where $y_i^{*(k)}$ are independently generated from $N\left(x_i'\hat{\beta}_{(-i)}, \hat{A}_{(-i)} + \psi_i\right)$.

Note that for the simple FH model, one may analytically obtain $E_*[y_i^*|y_{(-i)}]$ as $x_i'\hat{\beta}_{(-i)}$. But the real advantage of the proposed parametric bootstrap will be for complex models such as the Rao-Yu model. In this case, the parametric bootstrap distribution is analytically obtained as $N\left(x_i'\hat{\beta}_{(-i)}, \hat{A}_{(-i)} + \psi_i\right)$. For the Rao-Yu model, it would be difficult to obtain this bootstrap expectation. But independent samples can be drawn in different levels and could be approximated using bootstrap samples.

For example, consider the Fay-Herriot model. We could have generated $y_i^{*(k)}$ by first generating $\theta_i^{*(k)}$ from $N\left(x_i'\hat{\beta}_{(-i)}, \hat{A}_{(-i)}\right)$ and then $y_i^{*(k)}$ from $N(\theta_i^{*(k)}, \psi_i)$ The same comments apply to the computation of $f_{boot}(y_i|y_{(-i)})$ for complex models – samples can be drawn in a hierarchical fashion.

Definition 3.8. Standardized Empirical Bayes Parametric Bootstrap LOO Cross-Validatory residual:

$$\tilde{r}_i = \frac{E_*[y_i^*|y_{(-i)}]}{\sqrt{V_*(y_i^*|y_{(-i)})}} = \frac{r_i}{\sqrt{V_*(y_i^*|y_{(-i)})}}, \tag{3.20}$$

where,

$$V_*(y_i^*|y_{(-i)}) = \frac{1}{K}\sum_{k=1}^{K}\left[y_i^{*(k)}\right]^2 - \left(E_*[y_i^*|y_{(-i)}]\right)^2 \tag{3.21}$$

We now proceed to illustrate these residuals to produce a visual inspection of the adequacy of the assigned FH model in the previous section. In particular, we will plot $f_{boot}(y_i|y_{(-i)})$ against i for outlier diagnostics. We shall do this for the SAIPE model which uses all four available covariates

from the parametric bootstrap model.

Example 3.9. Empirical Bayes Parametric Bootstrap Residual Analysis SAIPE.

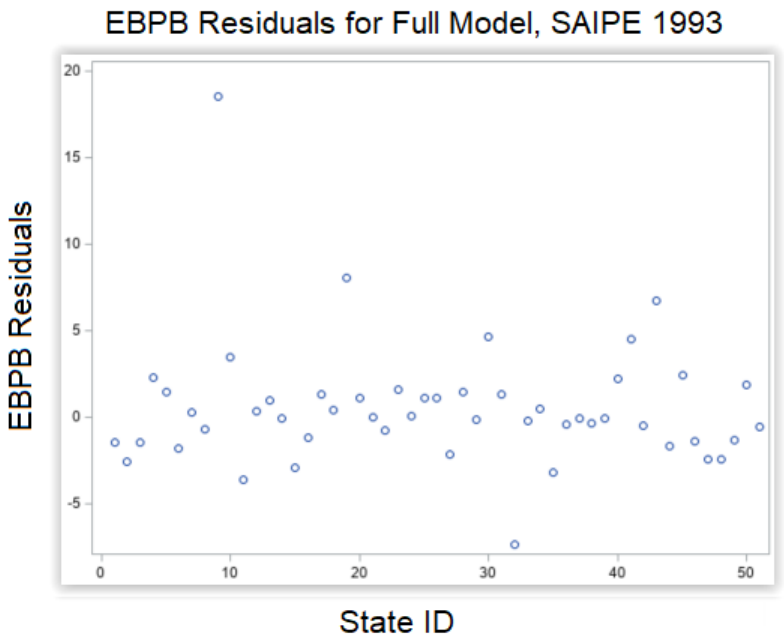


Figure 3.1: Empirical Bayes Parametric Bootstrap Residual Analysis (SAIPE 1993)

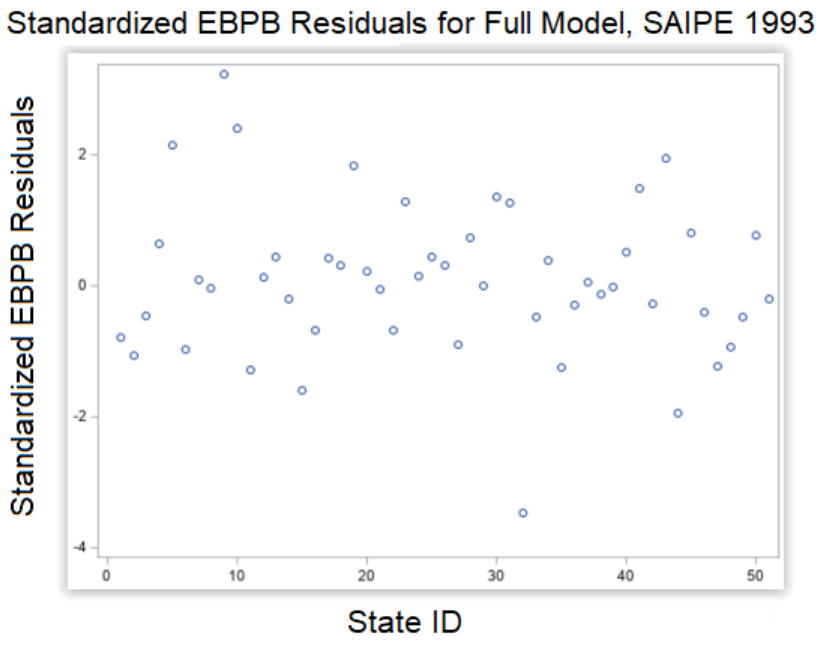


Figure 3.2: Empirical Bayes Bootstrap Residuals, Standardized. (SAIPE 1993)

3.6 Concluding Remarks

We reviewed some of the common likelihood-based, Bayesian, and other approaches to model evaluation. Following the concepts in Lahiri (2020) and the “leave one out approach" in Gelfand, Day, & Chang (1992), we constructed an empirical Bayes alternative to the Bayesian expected log predictive density (elpd) which is known to provide unstable results. This is especially true in the context of improper prior distributions. Moreover, this may require intense calculations using Markov Chain Monte Carlo (MCMC) monitoring, while bootstrap procedures are a lot easier to implement. Some closed form expressions were even provided for empirical Bayes bootstrap distributions under certain conditions. The LOO_{boot} has some profound potential for many aspects of model building, and can also be applied more readily to complex probability distributions.

The LOO_{boot} was applied to the SAIPE data and was found to be consistent with a mainstream variable selection method using AIC. Yet there was some distinction between scores, possibly due to differences in the way multi-collinearity influences each metric. Finally, EBPB LOO-CV residuals were defined and implemented with respect to one of the recommended models in the variable selection procedure. While there seemed to evidence of outliers in the non-standardized residual plot, the standardized plot seemed to indicated the goodness of fit of the full model. The points of that plot seemed to be evenly distributed across states within the range (-2 to 2).

Chapter 4

Generalized Observed Best Prediction for General Linear Mixed Models

4.1 Introduction

We now extend the usage of weighted best predictive estimators and empirical Bayes parametric bootstrap leave-one-out cross-validation to the estimation and validation of general linear mixed models, which consider the Fay-Herriot model as a special case. In particular, the general linear mixed model can be used to effectively address small area estimation problems in the context of time series applications. We will now consider the data y_{it}, whence y_i now represents the area-specific data data vector over time t=1 to T : $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$' and $\mathrm{y} = (y_1, y_2, \ldots, y_m)$'.

Chapter Outline

(4.1) Introduction

(4.2) Review of General Linear Mixed Models and Cross-Sectional Time Series Models. We will introduce some important time series from the small area literature and express them as special cases of the general linear mixed model.

(4.3) Best Predictive Estimation applied to the Rao-Yu Model. We apply the BPE to the Rao-Yu model as precursor to defining the GBP for general linear mixed model.

(4.4) Best Predictive Estimation for Variance Components. We continue the usage of observed

best prediction, now in the context of variance components and an autoregressive parameter.

(4.5) An Analytical Method for Estimating the Autogressive Parameter in the Rao-Yu Model. We derive a closed-form expression as part of our focus on model building.

(4.6) Generalized Observed Best Prediction for the General Linear Mixed Model. We extend the GBP to the general linear mixed model.

(4.7) Empirical Bayes Parametric Bootstrap Stationarity Model Selection. We demonstrate how the EBPB methodology can be applied to more sophisticated models, which in turn have greater complexity with regard to model building and diagnostics.

(4.8) Concluding Remarks.

4.2 Review of General Linear Mixed Models and Cross-Sectional Time Series Models

The General Linear Mixed Model

The *general linear mixed model* corresponds to the data obeying the following relationship:

$$y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}v + e, \tag{4.1}$$

- y is the $n \times 1$ vector of the response data.
- $\boldsymbol{X}$ is a $n \times p$ full rank matrix composed of p known characteristics for each sample point.

- β is a vector of p parameter coefficients.
- $\boldsymbol{Z}$ is an arbitrary $n \times h$ full rank matrix.
- $\boldsymbol{v}$ is a vector of random effects, distributed as $N_h(0, \boldsymbol{G})$, for some covariance matrix $\boldsymbol{G}$.
- $\boldsymbol{e}$ denotes the vector of error terms, distributed as $N_n(0, \boldsymbol{\Sigma})$, for some covariance matrix $\boldsymbol{\Sigma}$.
- e and v are independent.

The objective is to estimate a linear combination of the β and v vectors.

$$\theta = l'\beta + m'v,$$

where l and m are known vectors. Henderson (1950) showed that when the variance components $\boldsymbol{\Sigma}$, and $\boldsymbol{G}$ are known, then the best linear unbiased predictor (BLUP) of θ is expressed by:

$$\hat{\theta}^{BLUP} = l'\hat{\beta} + m'\hat{v} = l'\hat{\beta} + m'\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(y - \hat{\beta}\boldsymbol{X}), \tag{4.2}$$

where $\boldsymbol{V} = V(y) = \boldsymbol{\Sigma} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$ and $\hat{\beta} = \hat{\beta}^{BLUE} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}y$.

Time Series Models

Cross-sectional time-series models for application to small area estimation problems may be examined using the general linear mixed model (4.1). We will first present the Rao-Yu model in more intuitive terms, and then describe how it may be expressed under as a general linear mixed model. It is generally the goal to predict the area means from the most recent time point T, but we will merely write the predictor as y_{it} instead of y_{iT}.

Rao-Yu Model (Rao & Yu (1994))

For $i = 1, 2, \ldots, m$ and $t = 1, 2, \ldots, T$,

$$\begin{aligned} &\text{Level 1: } y_{it} = \theta_{it} + e_{it} \\ &\text{Level 2: } \theta_{it} = \boldsymbol{z}_{it}'\beta + \nu_i + u_{it} \\ &\text{Level 3: } u_{it} = \rho u_{i,t-1} + \epsilon_{it} \end{aligned} \tag{4.3}$$

- y_{it} - direct survey estimate in area i at time t
- θ_{it} - true area mean in area i at time t, unknown
- e_{it} - Normally distributed error term, with $\boldsymbol{e} = [e_{11}, \ldots, e_{mT}]' \sim N(0, \Psi)$ where Ψ is a block diagonal matrix consisting of m square matrix blocks Ψ_i of dimension $T \times T$.
- $\boldsymbol{z}_{it}$ - a vector of time-varying covariates. Not all entries are necessarily time-varying.
- β - vector of parameter coefficients, note the constancy across time and space.
- ϵ_{it} are independent and identically distributed as $N(0, \sigma^2)$.
- The autoregressive parameter, ρ satisfies $|\rho| < 1$.
- ν_i - independent and identically distributed Normal random variables with mean 0 and variance A.
- ϵ_{it}, u_{it}, and ν_i are pairwise and mutually independent across all areas and all times

Note that $Var(u_{it}) = Var(\sum_{t=1}^{T} \rho^{t-1}\epsilon_t) = \sum_{t=1}^{T} \rho^{2(t-1)} Var(\epsilon_t) = \frac{\sigma^2}{1-\rho^2}$ Putting this together yields the following marginal distribution of y_{it}:

$$\boxed{y_{it} \sim N\left(\boldsymbol{z}_{it}'\beta, \psi_{it} + \frac{\sigma^2}{1-\rho^2}\right)}$$

The model on θ_{it} may also be expressed as a distributed lag model:

$$\boxed{\theta_{it} = \rho\theta_{i,t-1} + (\boldsymbol{z}_{it} - \rho\boldsymbol{z}_{i,t-1})'\beta + (1-\rho)\nu_i + \epsilon_{it}}. \tag{4.4}$$

The above formulation is useful in that it removes dependency on the area-level random effects. These no longer need to be estimated, and focus may be placed on the estimation of the other variance parameters σ^2 and ρ. Note that the removal of one of the parameters is purchased at the cost of one time sample point per area (that is, there are only $T-1$ data elements instead of T).

Besides the intuitive approach above, the Rao-Yu model can be shown to be equivalent to the general linear mixed model using the same notation as follows:

$$\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iT}]',$$

$$\boldsymbol{X}_i = [\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2}, \ldots, \boldsymbol{z}_{iT}]',$$

$$\boldsymbol{Z}_i = [\boldsymbol{1}_T, \boldsymbol{I}_T],$$

$$\boldsymbol{v}_i' = [\boldsymbol{\nu}_i, \boldsymbol{u}_i'],$$

$$\boldsymbol{e}_i = [e_{i1}, e_{i2}, \ldots, e_{iT}]',$$

where $\boldsymbol{1}_T$ denotes the column vector of all ones with length T, and $\boldsymbol{I}_T$ is the T $\times$ T identity matrix. We also have that $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ are block diagonal matrices with $\boldsymbol{\Sigma} = \text{diag}(\Psi_1, \Psi_2, \ldots, \Psi_m)$ and similarly

that $\boldsymbol{G}$ = diag(G_1, $G_2, \ldots, G_m$), where the block G_i denotes the matrix of variance components defined below:

$$\boldsymbol{G}_i = \begin{bmatrix} A & \mathbf{0}_T' \\ \mathbf{0}_T & \sigma^2 \boldsymbol{\Lambda}_i \end{bmatrix},$$

with $\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_T = (\lambda)_{t,s}$ as the covariance matrix of the vector $\boldsymbol{u}_i = [u_{i1}, u_{i2}, \ldots, u_{iT}]$'. For the version of the Rao-Yu model with autoregressive errors, which is the subject of the current treatise, the entries for $\boldsymbol{\Lambda}_i$ are given by $(\lambda)_{t,s} = \rho^{|t-s|}/(1-\rho^2)$. The covariance matrix of $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iT}]$' is given by

$$\boldsymbol{V}_i = \Psi_i + \sigma^2 \boldsymbol{\Lambda}_i + A\boldsymbol{J}_T,$$

where $\boldsymbol{J}_T = \mathbf{1}_T \mathbf{1}_T'$ is the T × T matrix consisting of all ones. Finally, the small area parameter of interest, the most recent time period estimate in area i, θ_{iT} is given by $\theta_{iT} = l'\beta + r'v_i$ where $l = \boldsymbol{z}_{iT}$ and r = [1,0,. . . ,0,1]'.

Best Linear Unbiased Predictor

The **BLUP** estimator (with each A, σ^2, and ρ known) for the i^{th} area mean during the most recent time period T is given by:

$$\boxed{\tilde{\theta}_{iT}^{BLUP} = \boldsymbol{z}_{iT}\hat{\beta} + (A\mathbf{1}_T + \sigma^2 \boldsymbol{\lambda}_T)' \boldsymbol{V}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{X}_i \hat{\beta})}, \tag{4.5}$$

where $\boldsymbol{\lambda}_T$ is the T^{th} row of $\boldsymbol{\Lambda}_T$ and

$$\hat{\beta} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i' \boldsymbol{V}^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i' \boldsymbol{V}^{-1} y_i.$$

The BLUP estimator minimizes the mean-squared error among the class of unbiased linear estimators for θ. When the variance parameters are unknown, they must be first estimated using some method and then plugged into the BLUP to obtain the *empirical* best linear unbiased predictor, or EBLUP.

The **EBLUP** estimator (A, σ^2, and ρ all unknown) for the i^{th} area mean during the most recent time period T is given by:

$$\boxed{\tilde{\theta}_{iT}^{EBLUP} = \boldsymbol{z}_{iT}\hat{\beta} + (\hat{A}\mathbf{1}_T + \hat{\sigma}^2\boldsymbol{\lambda}_T)'\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{X}_i\hat{\beta})} \tag{4.6}$$

Estimation of Covariance Parameters

The empirical best linear unbiased predictor in the Rao–Yu model requires the following parameters, which must be estimated: σ^2, A, and ρ. Rao and Yu (1992) provided estimators for A and σ^2 for when the auto-regressive parameter ρ was both known and unknown. When ρ is known we can obtain somewhat simpler consistent (albeit biased) estimators for σ^2 and A. Rao and Yu (1994)

proposed the following estimator for the unknown ρ:

$$\hat{\rho}_{RY} = \frac{\sum_{i=1}^{m}\sum_{t=1}^{T-2}\left(\hat{a}_{it}(\hat{a}_{i,t+1} - \hat{a}_{i,t+2}) - (\psi_{t,t}^{(i)} - \psi_{t,t+1}^{(i)})\right)}{\sum_{i=1}^{m}\sum_{t=1}^{T-2}\left(\hat{a}_{it}(\hat{a}_{i,t} - \hat{a}_{i,t+1}) - (\psi_{t,t+1}^{(i)} - \psi_{t,t+2}^{(i)})\right)}, \tag{4.7}$$

where $\psi_{t,t}^{(i)}$ denotes $Var(e_{it})$, and $\psi_{t,t+1}^{(i)}$ represents $\text{Cov}(e_{it}, e_{i,t+1})$, and $\hat{a}_{i,t}$ is the $(it)^{th}$ least squares residual: $y_{it} - x_{it}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y_{it}$. This estimator for ρ is consistent, and its use as a plug in estimator does not affect the unbiasedness of $\hat{\theta}_{iT} = \hat{\theta}_{iT}(\hat{\rho})$. However, simulations performed by Rao and Yu observed some instability in the estimates of ρ, as values appeared to be outside of the domain $[-1, 1]$ of ρ. Recall that the condition that $|\rho| < 1$ guarantees the stationarity of the series u_{it} and leads to the following variance expression:

$$V(u_{it}) = \frac{\sigma^2}{1-\rho^2}$$

With this formula, the variance of u_{it} diverges as $\rho \to 1$ from the left.

Random Walk Model (Datta, Lahiri, Maiti (2002))

For i=1,2,. . . ,m and t=1,2,. . . ,T,

$$\begin{aligned} &\text{Level 1: } y_{it} = \theta_{it} + e_{it} \\ &\text{Level 2: } \theta_{it} = z_{it}'\beta + \nu_i + u_{it} \\ &\text{Level 3: } u_{it} = u_{i,t-1} + \epsilon_{it} \end{aligned} \tag{4.8}$$

The random walk model shares the same representation under the general linear mixed model

as the Rao-Yu model, with the important caveat that with $\mathbf{\Lambda}_i = \mathbf{\Lambda}_T = (\lambda)_{t,s} = \min(t, s)$. Rao and Yu (1994) observed some instability with regards to the method of moment estimators when ρ was close to zero. Datta et al. (2002) also encountered this same phenomenon of $\rho \approx 1$, and they reacted by simply forcing the value of ρ to be equal to one, instead. This leads to the random walk model. While conveniently avoiding the issue of estimating ρ, the random walk model is *non-stationary*, and results in a non-finite variance for the u_{it}. From a practical standpoint, the random walk model is best applied to finite series (which is the realistically the only type of problem in applied statistics).

Diallo (2014) explored maximum likelihood estimators for ρ. In particular, these estimators may still perform poorly when ρ is close to 1. It seems sensible to perform a preliminary test to determine if ρ is significantly different from unity. If the null hypothesis $\rho = 1$ is not rejected, then we assume the Random Walk model. The Dicky-Fuller test provides econometricians (who typically ignore the uncertainty in the first level of the hierarchy) with a means of deducing whether the parameter ρ is equal to one in autoregressive models. Levin, Lin, and Chu (2002) developed a unit root test for panel data under very general conditions, but as with most econometric models, they did not incorporate uncertainty with regard to sampling errors.

4.3 Best Predictive Estimation applied to the Rao-Yu Model

We now provide an explicit calculation of the best predictive estimator for the Rao-Yu model. For the parameter of interest, the vector of mixed effects is defined as $\theta = \boldsymbol{F}'\mu + \boldsymbol{R}'v$, where $\boldsymbol{F}$ and $\boldsymbol{R}$ are known matrices, and μ is the *true* mean of y, $Ey = \mu \neq \boldsymbol{X}\beta$. Then the best linear predictor of θ is given by

$$\begin{aligned} E_M(\theta|y) = \boldsymbol{F}'\mu + \boldsymbol{R}'E_M(v|y) &= \boldsymbol{F}'\mu + \boldsymbol{R}'\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(y - \boldsymbol{X}\beta) \\ &= \boldsymbol{F}'y - \boldsymbol{\Gamma}(y - \boldsymbol{X}\beta), \end{aligned} \tag{4.9}$$

where $\Gamma = \boldsymbol{F}' - \boldsymbol{B}$, with $\boldsymbol{B} = \boldsymbol{R}'\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}$. The best predictive estimator for the general linear mixed model (4.1) is given by $\text{BPE}(\beta) = \tilde{\beta}_{BPE} = (\boldsymbol{X}'\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{y}$. Recall that $\boldsymbol{V} = \boldsymbol{\Sigma} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$ and that $\boldsymbol{F} = \boldsymbol{R} = \boldsymbol{I}_{mT}$ for the Rao-Yu model. Because of the block diagonal structure associated with this model, it is sufficient to consider $\boldsymbol{F}_i = \boldsymbol{R}_i = \boldsymbol{I}_T$, $\Gamma_i = \boldsymbol{I}_T - \boldsymbol{B}_i$, and $B_i =' \boldsymbol{G}_i\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}$. Plugging in the appropriate values for $\boldsymbol{G}_i$, $\boldsymbol{Z}_i$, and $\boldsymbol{V}_i$ corresponding to the Rao-Yu model yields the following

expression for the best predictive estimator:

$$\begin{aligned}
\boldsymbol{\Gamma}_i'\boldsymbol{\Gamma}_i &= \left\{ \boldsymbol{I}_T - \begin{bmatrix} \sigma_\nu^2 & \mathbf{0}' \\ \mathbf{0} & \sigma^2\boldsymbol{\Lambda}_i \end{bmatrix} [1_T, \boldsymbol{I}_T][\boldsymbol{\Psi}_i + \sigma^2\boldsymbol{\Lambda}_i + A\boldsymbol{J}_T]^{-1} \right\}' \\
&\quad \times \left\{ \boldsymbol{I}_T - \begin{bmatrix} \sigma_\nu^2 & \mathbf{0}' \\ \mathbf{0} & \sigma^2\boldsymbol{\Lambda}_i \end{bmatrix} [1_T, \boldsymbol{I}_T][\boldsymbol{\Psi}_i + \sigma^2\boldsymbol{\Lambda}_i + A\boldsymbol{J}_T]^{-1} \right\} \\
&= \left\{ \boldsymbol{I}_T - [\boldsymbol{\Psi}_i' + \sigma^2\boldsymbol{\Lambda}_i' + A\boldsymbol{J}_T]^{-1} \begin{bmatrix} 1_T' \\ \boldsymbol{I}_T \end{bmatrix} \begin{bmatrix} \sigma_\nu^2 & \mathbf{0}' \\ \mathbf{0} & \sigma^2\boldsymbol{\Lambda}_i' \end{bmatrix} \right\} \\
&\quad \times \left\{ \boldsymbol{I}_T - \begin{bmatrix} \sigma_\nu^2 & \mathbf{0}' \\ \mathbf{0} & \sigma^2\boldsymbol{\Lambda}_i \end{bmatrix} [1_T, \boldsymbol{I}_T][\boldsymbol{\Psi}_i + \sigma^2\boldsymbol{\Lambda}_i + A\boldsymbol{J}_T]^{-1} \right\}
\end{aligned} \tag{4.10}$$

$$\begin{aligned}
&= \boldsymbol{I}_T + [\boldsymbol{\Psi}_i' + \sigma^2\boldsymbol{\Lambda}_i' + A\boldsymbol{J}_T]^{-1} \begin{bmatrix} 1_T' \\ \boldsymbol{I}_T \end{bmatrix} \begin{bmatrix} \sigma_\nu^4 & \mathbf{0}' \\ \mathbf{0} & \sigma^4\boldsymbol{\Lambda}_i'\boldsymbol{\Lambda}_i \end{bmatrix} [1_T, \boldsymbol{I}_T][\boldsymbol{\Psi}_i + \sigma^2\boldsymbol{\Lambda}_i + A\boldsymbol{J}_T]^{-1} \\
&\quad - 2 \begin{bmatrix} \sigma_\nu^2 & \mathbf{0}' \\ \mathbf{0} & \sigma^2\boldsymbol{\Lambda}_i \end{bmatrix} [1_T, \boldsymbol{I}_T][\boldsymbol{\Psi}_i + \sigma^2\boldsymbol{\Lambda}_i + A\boldsymbol{J}_T]^{-1} \\
&= \boldsymbol{I}_T + (\boldsymbol{V}_i')^{-1} \begin{bmatrix} 1_T' \\ \boldsymbol{I}_T \end{bmatrix} \begin{bmatrix} \sigma_\nu^4 & \mathbf{0}' \\ \mathbf{0} & \sigma^4\boldsymbol{\Lambda}_i'\boldsymbol{\Lambda}_i \end{bmatrix} [1_T, \boldsymbol{I}_T]\boldsymbol{V}_i^{-1} - 2(\boldsymbol{V}_i')^{-1} \begin{bmatrix} 1_T' \\ \boldsymbol{I}_T \end{bmatrix} \begin{bmatrix} \sigma_\nu^2 & \mathbf{0}' \\ \mathbf{0} & \sigma^2\boldsymbol{\Lambda}_i'\boldsymbol{\Lambda}_i \end{bmatrix} \\
&= \boldsymbol{I}_T + \boldsymbol{V}_i'\boldsymbol{Z}_i\boldsymbol{G}_i'\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}_i^{-1} - 2\boldsymbol{V}_i'\boldsymbol{Z}_i\boldsymbol{G}_i'.
\end{aligned} \tag{4.11}$$

Finally, we have that $\boldsymbol{\Gamma}'\boldsymbol{\Gamma} = diag(\boldsymbol{\Gamma}_1'\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_m'\boldsymbol{\Gamma})$. Equivalently, we can express the best predictive estimator for β in the Rao-Yu model in terms of the overall diagonal matrices:

$$\boxed{\tilde{\beta}_{RY}^{BPE} = \left\{\boldsymbol{X}'(\boldsymbol{I}_{mT} + \boldsymbol{V}'\boldsymbol{Z}\boldsymbol{G}'\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1} - 2\boldsymbol{V}'\boldsymbol{Z}\boldsymbol{G}')\boldsymbol{X}\right\}^{-1} \boldsymbol{X}'(\boldsymbol{I}_{mT} + \boldsymbol{V}'\boldsymbol{Z}\boldsymbol{G}'\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1} - 2\boldsymbol{V}'\boldsymbol{Z}\boldsymbol{G}')y}.$$

Compare this result with the maximum likelihood estimator for β, that is,

$$\boxed{\hat{\beta}_{RY}^{MLE} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}\left(\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{y}_i\right)}.$$

Computational Example

The Current Population Survey (CPS) collects unemployment data on a monthly basis. There are 60,000 respondents each month drawn among the non-institutionalized civilian population, persons 15 and older. The data were obtained from the Bureau of Labor Statistics, see Tiller (2001) or Tiller (2005) for more information. The data contained 56 geographical domains (mostly states) that spanned the years 1990 through 2013.

We now proceed to compare the standard Rao-Yu model, which utilizes β as the weighted least squares estimator $\hat{\beta} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}\left(\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{y}_i\right)$, versus the Rao-Yu model with utilizes the BPE. We again run the same model as before, now utilizing the R package SAE2 (Fay & Diallo (2015)) which calculates the Rao-Yu model given the number of areas (D=m), the number of time periods (T), a list of variance-covariance matrices corresponding to each area. We consider T=10 annual time points corresponding to April months from 2004 through 2013 for 10 state domains: AK, AL, VA, MD, MI, SD, SC, ID, MA, and NJ. The 1-year, 2-year, and 3-year correlations between unemployment estimates within the same domain were 0.11, 0.04, and 0.2,

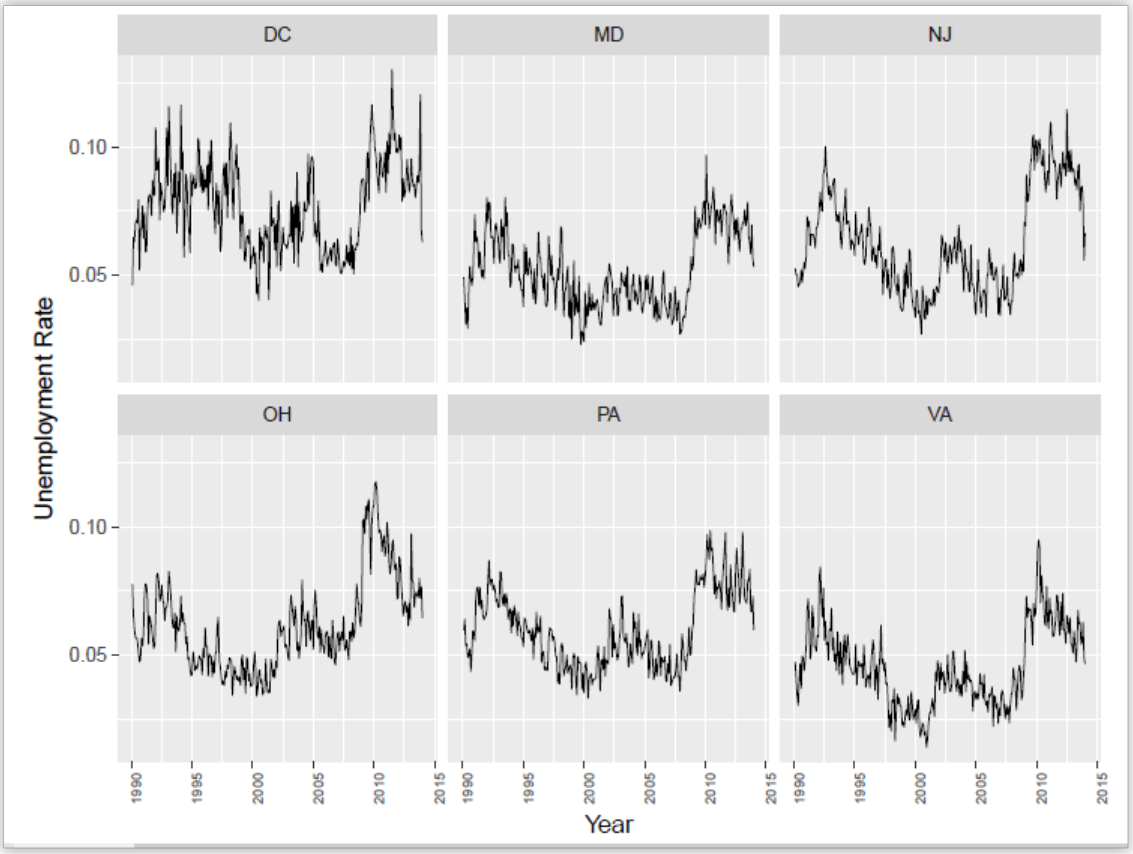


Figure 4.1: Direct Estimates of Monthly Unemployment Rate 1990-2015
Source: Current Population Survey

respectively. Estimates between 4 and 10 years were deemed to be uncorrelated. The response variable was the unemployment rate, with monthly covariates of CPS employment rate, state payroll records, and the number of state unemployment insurance claims.

Model: CPS Unemployment = CPS Employment + Payroll + Unemployment Insurance Claims

resultT < − eblupRY(CPSun CPSem + Cntwoer + CesEm, D, T, vardir = vc, data=state, ids=state$ST)

Results: The model parameter estimates converge after 26 iterations, and uses Residual Maximum Likelihood in order to estimate the covariance parameters. All four of the covariates are found to be significant to the model, including that of the intercept. A sizable estimate of the correlation coefficient, ρ is given by the model (0.48).

```
$fit$model
[1] "T: Rao-Yu, REML"

$fit$convergence
[1] TRUE

$fit$estcoef
                  beta  std.error    tvalue       pvalue
(Intercept) 23.2464143 2.36831656  9.815586 0.000000e+00
emRate      -0.1018979 0.05305433 -1.920632 5.477809e-02
Cntwoer      1.1071381 0.18482693  5.990134 2.096679e-09
CesEm       -0.2463037 0.05936216 -4.149171 3.336821e-05

$fit$estvarcomp
        estimate  std.error
sig2_u 0.1616483 0.05937534
sig2_v 1.3849655 0.71688286
rho    0.4785927 0.27526331

$fit$iterations
num.iter
      26

$fit$goodness
          loglike restrictedloglike
        -113.0929         -107.1357


$parm
          rho         sig2_u         sig2_v  loglikelihood constrained.ll
     0.4785927      0.1616483      1.3849655   -113.0928786   -107.1357387
      num.iter
    26.0000000
```

Figure 4.2: Results for Rao-Yu Model

Discussion: We can observe that both sets of parameter estimates were intuitive and followed the same sign: larger values for the CPS employment estimate or larger state payroll records have a negative effect on the unemployment rate, while the number of employment claims had a positive

Comparison of Parameter Estimates: BPE vs. Weighted LSE		
Covariate	**WLS**	**BPE**
Intercept	23.246	6.552
CPS Emp.	-0.102	-0.016
Unemp. Ins	1.107	1.174
Payroll	-0.246	-0.056

Comparison of MSE Estimates: WLS vs BPE	
Estimator for β	Estimated MSE($\boldsymbol{\theta}$).
WLS $\hat{\beta}$	40.5
BPE $\tilde{\beta}$	26.6

effect. Interestingly, the intercept from the BPE parameter estimates is around $6\frac{1}{2}\%$, which is near the average unemployment rate. On the other hand, the intercept for the WLS is at 26% - astronomically higher than typical unemployment rates observed during the time period. This seems to indicate a greater stability in the rate prediction. Indeed, the Rao-Yu model infused with the BPE $\tilde{\beta}^{BPE}$ is shown to have a much lower MSE approximation (facilitated under the presumed model) than that of the Rao-Yu model using the standard WLS estimator for the parameter coefficients, β^{WLS}. This is somewhat surprising, as the BPE is not supposed to be as efficient than the EBLUP when the model is correctly specified. The predictors are all highly prescient, and exhibited very high correlations with the response variable CPS Unemployment estimates. This could be an indication of misspecification. In most applications, there are many unknown and immeasurable factors that may have undue influence on the dependent variable in the model. The robustness provided by the usage of the BPE should always be compared to the of the EBLUP. Any disparity between the two measures should be investigated prior the final reporting of small area estimates. Additional metrics may be introduced, however, including the MSE approximation given by Datta for correlated data,

for examples see Rao and Molina (2015).

4.4 Best Predictive Estimation for Variance Components applied to the Rao-Yu Model

We now proceed to the interesting case when the variance parameters A, σ^2, ρ are all unknown. As with the model variance A in the Fay-Herriot model, estimators for σ^2 can often be zero or negative, and estimators for ρ can be negative or even greater than 1. The best predictive estimator affords a new approach to obtaining good approximations for the variance components, in both the Rao-Yu and Random Walk models. For the Random Walk model, note that the BPE can be obtained in a similar fashion by changing $\mathbf{\Lambda}_i = (\lambda_{t,s}) = \rho^{|t-s|}/(1-\rho^2)$ to the random walk matrix $\mathbf{\Lambda}_i = (\lambda_{t,s}) = min(t, s)$.

Now suppose that the true underlying distribution of $\boldsymbol{y}$ was not (4.1) but was instead given by the following:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z}v + \boldsymbol{e}, \tag{4.12}$$

where $\boldsymbol{\mu} = E(\boldsymbol{y})$. Now reverting to estimation of a vector $\boldsymbol{\theta} = \boldsymbol{F}'\boldsymbol{\mu} + \boldsymbol{R}'\boldsymbol{v}$ of unknown parameters (fixed effects + variance parameters). Jiang, Nguyen, & Rao (2011) noted the mean-

square prediction error of some generic estimator $\hat{\boldsymbol{\theta}}$, MSPE($\hat{\boldsymbol{\theta}}$), may be expressed as:

$$\begin{aligned} MSPE(\hat{\boldsymbol{\theta}}) = tr[\boldsymbol{R}'(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G})r] + \boldsymbol{\mu}'[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}]' \\ \times \boldsymbol{\Gamma}'\boldsymbol{\Gamma}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}] + tr[(\boldsymbol{L} - \boldsymbol{B})\boldsymbol{V}(\boldsymbol{L} - \boldsymbol{B})'], \end{aligned} \tag{4.13}$$

where $\boldsymbol{L} = [\boldsymbol{F}' - \boldsymbol{\Gamma}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega})$. The weighting matrix $\boldsymbol{\Omega}$ determines whether the estimator utilizes the MLE ($\boldsymbol{\Omega} = \boldsymbol{V}^{-1}$), BPE ($\boldsymbol{\Omega} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma}$), or some other estimator. Note that there are three terms in the MSPE expression above, the first term does not include Ω and so is independent of the choice of estimator for β. This, therefore, pertains to the general MSE when all parameters are known. The second term corresponds to the additional uncertainty incurred from not knowing β, and the final term is due to uncertainty from estimating the variance parameters. Under this general expression for the mean squared prediction error for general linear mixed model models, we are able to obtain an estimate for the variance component vector:

$$BPE : (\tilde{A}, \tilde{\sigma}^2, \tilde{\rho}) = \underset{A,\sigma^2,\rho}{\arg\max}\ MSPE(\tilde{\theta}),$$

where $A \in (0, \infty), \sigma^2 \in (0, \infty)$, and $\rho \in (-1, 1)$. As before with the Fay-Herriot Model, we consider a vector optimization where the objective function is that of the MSE, and the parameters to be optimized are the σ_v^2, σ, and ρ. Unfortunately, the surface generated by the MSPE as a function of (A, σ^2, ρ) was not suitable for optimization. Convergence could not be achieved using different starting points, or using multiple SAS software non-linear programming optimization algorithms (i.e., Newton Raphson, conjugate gradient, Nelder-Mead simplex, Trust region, etc.).

Optimization Start			
Active Constraints	0	Objective Function	5.5783776179
Max Abs Gradient Element	0.5871828		

Iteration		Restarts	Function Calls	Active Constraints	Objective Function	Objective Function Change	Max Abs Gradient Element	Step Size	Slope of Search Direction
1	*	0	3	1	5.75455	0.1762	0.5304	0.0996	-94.509
2		0	4	1	6.11957	0.3650	0.3684	0.440	-1.002
3		0	5	1	6.76819	0.6486	0.1927	0.912	-1.000
4		0	6	1	7.46273	0.6945	0.0963	1.000	-1.002
5		0	7	1	8.15716	0.6944	0.0482	1.000	-1.002
6		0	8	1	9.16395	1.0068	0.0176	1.735	-1.001

Optimization Results			
Iterations	6	Function Calls	39
Hessian Calls	7	Active Constraints	1
Objective Function	-1.79769E308	Max Abs Gradient Element	0.0176161665
Slope of Search Direction	-1.000833791	Ridge	0

ERROR: The function value of the objective function cannot be computed at the starting point.

pleted.

Page Break

Optimization Results				
Parameter Estimates				
N	Parameter	Estimate	Gradient Objective Function	Active Bound Constraint
1	X1	56.824005	0.017616	
2	X2	0.010000	-0.215352	Lower BC
3	X3	-0.264599	-0.000578	

Value of Objective Function = -1.79769E308

xrc		
56.824005	0.010000	-0.264599

Figure 4.3: Numerical Results for Rao-Yu Model Optimization

4.5 An Analytical Method for Estimating the Autogressive Parameter in the Rao-Yu Model

We briefly investigate an alternative method for the estimation of the autoregressive parameter ρ. Recall that the distributed lag model becomes:

$$\boxed{y_{it} - y_{i,t-1} = (\boldsymbol{z}_{it} - \boldsymbol{z}_{i,t-1})'\beta + e_{it} - e_{i,t-1} + (\rho - 1)u_{i,t-1} + \epsilon_{it}}$$

When β and σ^2 are known, it is straightforward to obtain the MLE of ρ:

$$logL(\boldsymbol{y}, \rho) = -\tfrac{1}{2}\sum_{i=1}^{m}\sum_{t=1}^{T} log\left[D_{it} + D_{i,t-1} + (1 + \tfrac{(1-\rho)(1-\rho^T)}{1+\rho})\sigma^2\right]$$

$$\times -\tfrac{1}{2}\sum_{i=1}^{m}\sum_{t=1}^{T}\left\{\frac{(y_{it}-y_{i,t-1}-z_{it}\beta+z_{i,t-1}\beta)^2}{D_{it}+D_{i,t-1}+(1+\frac{(1-\rho)(1-\rho^T)}{1+\rho})\sigma^2}\right\}$$

$$\implies \tfrac{d}{d\rho}logL(\boldsymbol{y}, \rho) = -\tfrac{1}{2}\sum_{i=1}^{m}\sum_{t=1}^{T}\left[D_{it} + D_{i,t-1} + (1 + \tfrac{(1-\rho)(1-\rho^T)}{1+\rho})\sigma^2\right]$$

$$- \tfrac{1}{2}\sum_{i=1}^{m}\sum_{t=1}^{T}(y_{it} - y_{i,t-1} - z_{it}\beta + z_{i,t-1}\beta)^2 = 0$$

$$\implies (1 + \tfrac{1+(1-\rho)(1-\rho^T)}{1+\rho}) = \left\{\sum_{i=1}^{m}\sum_{t=1}^{T}(y_{it} - y_{i,t-1} - z_{it}\beta + z_{i,t-1}\beta)^2 - \sum_{i=1}^{m}\sum_{t=1}^{T}(D_{it} + D_{i,t-1})\right\} \Big/ mT\sigma^2.$$

For large T, we can let $1 - \rho^T \to 1$, resulting in,

$$\implies \hat{\rho}_{MLE} = \frac{2}{S} - 1$$

where $S = \frac{\sum_{i=1}^{m}\sum_{t=1}^{T}(y_{it}-y_{i,t-1}-z_{it}\beta+z_{i,t-1}\beta)^2-\sum_{i=1}^{m}\sum_{t=1}^{T}(D_{it}+D_{i,t-1})}{mT\sigma^2}$. We can see directly that $\hat{\rho}_{MLE}$ is consistent as m, T $\rightarrow \infty$, but it also applies to all MLE estimators.

Observe that the distributed lag function only depends on ρ through the function f(ρ) = $\frac{1}{1+\rho}$. The estimator S is in fact an unbiased predictor for $\frac{1}{1+\rho}$. It may be more advantageous to use the distributed lag model in order to first estimate the ρ and σ^2 before attempting to predict the small area means in the general model. Although ρ is consistent, it is possible that adequate sample sizes may not exist to obtain a precise estimator for ρ, but simulations showed that estimates for $\frac{1}{1+\rho}$ were viable.

Simulation Study

We considered the balanced case and generated a population of 500 time series observations y_{it}, so t=1, . . . , 500. Since the random effects are lost in these successive differences within a single area i, there is no loss of information. N=50 iterations were executed for variances from 1, 5, and 10 for both the sampling error, Ψ, and the time series errors σ^2. Statistical Analysis Software (SAS) was used to generate random samples from a normal distribution using the RANNOR sampling function with explicit starting seeds. We used two main metrics to evaluate estimators:

$$Bias^* = \frac{1}{500}\sum_{t=1}^{500}(\hat{\rho}-\rho)$$

$$MSE^* = \frac{1}{500}\sum_{t=1}^{500}(\hat{\rho}-\rho)^2$$

The simulation results for the examined MLE estimator of ρ are provided in the following tables,

followed by comments.

Simulation Results: N=50, m=1, T=500, D=1, σ^2=1		
Value of ρ	Bias*	MSE*
0.1	-0.034	0.039
0.5	-0.055	0.108
0.9	-0.087	0.248

We can see in the first table that the estimator can perform well under different extreme values for ρ. It is also noteworthy, and perhaps expected, that the bias and mean-squared error are constant whenever D=σ^2. The result for when D=5 and σ^2=1 appears curious, but the phenomenon does not seem to persist. Within the final table, we see that the estimator for ρ requires a larger sample when the sampling variances are larger than the time series variance component, σ^2. However, the simulation rate of convergence was significantly lower for the estimation of $\frac{1}{1+\rho}$. There does seem to some issue with the fact that all bias estimates are negative, although we do notice that this bias is anti-correlated with sample size and approaches zero as datasets are mage larger.

Simulation Results: N=50, m=1, T=500, ρ = 0.5			
Sample Variance Ψ	Value of σ^2	Bias*	MSE*
1	1	-0.055	0.108
1	5	-0.010	0.020
1	10	-0.008	0.015
5	1	-0.919	308.603
5	5	-0.055	0.108
5	10	-0.021	0.042
10	1	-0.050	21.853
10	5	-0.210	0.604
10	10	-0.055	0.108

Simulation Results: N=50, m=1, D=1, σ^2=1, ρ = 0.5		
Number of Samples Area	Bias*	MSE*
50	-0.682	3.418
100	-0.377	1.169
250	-0.164	0.438
500	-0.055	0.108

4.6 Generalized Observed Best Prediction for the General Linear Mixed Model

We now extend our results to the general model. In this context the loss function we seek to minimize is,

$$E\left[(\hat{\theta}^B - \theta)'\boldsymbol{W}(\hat{\theta}^B - \theta)\right]$$

with respect to the hyperparameters, where E is expectation with respect to the true underlying model; θ is a $m \times 1$ vector of small area means, $\hat{\theta}^B$ is a $m \times 1$ vector of BP of small area means, and W is a positive semi-definite matrix, which may not necessarily be a diagonal matrix. This may be useful for practitioners when the conditional distribution of $\hat{\theta}_i^B$'s are correlated. We follow the approach in Jiang, Nguyen, & Rao (2011) to obtain a general formula.

The assumed linear mixed model is $Y = \boldsymbol{X}\beta + \boldsymbol{Z}v + \boldsymbol{e}$, where $e \sim N(0, \boldsymbol{R}), v \sim N(0, \boldsymbol{G})$ are independent vectors with covariance matrices which are non-singular and symmetric positive-definite. It is also helpful to use the following representation, which may or may not be equivalent to the true model: $Y = \mu + \boldsymbol{Z}v + \boldsymbol{e}$. In general, we are interested in the prediction of: $\theta = \boldsymbol{F}'\boldsymbol{\mu}$ + $\boldsymbol{R}'\boldsymbol{v}$, where $\boldsymbol{F}$ and $\boldsymbol{R}$ are known matrices. The Fay-Herriot model, the Rao-Yu model and the Random-Walk may each fit into this paradigm without any information loss. The Fay-Herriot model,

in particular, corresponds to the case where $\boldsymbol{F} \equiv \boldsymbol{R} \equiv \boldsymbol{I}_m$, the $m \times m$ identity matrix.

Case when $\boldsymbol{\Sigma}$ and $\boldsymbol{G}$ are known. When $\boldsymbol{\Sigma}$ and $\boldsymbol{G}$ are known, then the best predictor (i.e., again the one that minimizes the mean squared error), is given by the following expectation under the assumed model:

$$\begin{aligned} E_M(\theta|y) &= \boldsymbol{F}'\mu + \boldsymbol{R}'E_M(v|y) \\ &= \boldsymbol{F}'\boldsymbol{X}\beta + \boldsymbol{R}'\boldsymbol{G}\boldsymbol{Z}'(\boldsymbol{\Sigma} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}')^{-1}(y - \boldsymbol{X}\beta) \\ &= \boldsymbol{F}'\boldsymbol{X}\beta + \boldsymbol{B}'\boldsymbol{V}^{-1}(y - \boldsymbol{X}\beta) \\ &= \boldsymbol{F}'y - \boldsymbol{\Gamma}(y - \boldsymbol{X}\beta), \end{aligned} \tag{4.14}$$

where we have made the substitutions $\boldsymbol{B} = \boldsymbol{R}'\boldsymbol{G}\boldsymbol{Z}'$, $\boldsymbol{V} = \boldsymbol{\Sigma} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$, and $\boldsymbol{\Gamma} = \boldsymbol{F}' - \boldsymbol{B}$. It could be argued that the appropriate starting point for considering different loss functions would be to start with obtaining a new version of the best predictor which is a function main EBLUP theory to the extent in which the estimation of the variance parameters. Under the weighted mean squared prediction error, $MSPE(\hat{\theta}^{BP}) = E\left[(\hat{\theta}^{B} - \theta)'W(\hat{\theta}^{B} - \theta)\right]$, we first examine the expression inside of

the integral.

$$
\begin{aligned}
(\hat{\theta}^B - \theta)' \boldsymbol{W} (\hat{\theta}^B - \theta) &= \left\{ \boldsymbol{F}' \boldsymbol{X} \beta + \boldsymbol{R}' \boldsymbol{G} \boldsymbol{Z}' \boldsymbol{V}^{-1} (y - \boldsymbol{X}\beta) - \boldsymbol{F}' \boldsymbol{\mu} + \boldsymbol{R}' \boldsymbol{v} \right\}' \boldsymbol{W} \\
&\times \left\{ \boldsymbol{F}' \boldsymbol{X} \beta + \boldsymbol{R}' \boldsymbol{G} \boldsymbol{Z}' \boldsymbol{V}^{-1} (y - \boldsymbol{X}\beta) - \boldsymbol{F}' \boldsymbol{\mu} + \boldsymbol{R}' \boldsymbol{v} \right\} \\
&= \{ \boldsymbol{F}' y - \boldsymbol{\Gamma} (y - \boldsymbol{X}\beta) - \boldsymbol{F}' \boldsymbol{\mu} + \boldsymbol{R}' \boldsymbol{v} \}' \boldsymbol{W} \\
&\times \{ \boldsymbol{F}' y - \boldsymbol{\Gamma} (y - \boldsymbol{X}\beta) - \boldsymbol{F}' \boldsymbol{\mu} + \boldsymbol{R}' \boldsymbol{v} \}' \\
&= \{ \boldsymbol{H}' v + \boldsymbol{F}' e - \boldsymbol{\Gamma} (y - \boldsymbol{X}\beta) \}' \boldsymbol{W} \boldsymbol{H}' v + \boldsymbol{F}' e - \boldsymbol{\Gamma} (y - \boldsymbol{X}\beta) \\
&= (e' \boldsymbol{F} + v' \boldsymbol{H}) \boldsymbol{W} (\boldsymbol{H}' v + \boldsymbol{F}' e) - 2 (e' \boldsymbol{F} + v' \boldsymbol{H}) \boldsymbol{W} \boldsymbol{\Gamma} (y - \boldsymbol{X}\beta) \\
&+ (\beta' \boldsymbol{X}' - y') \boldsymbol{\Gamma}' \boldsymbol{W} \boldsymbol{\Gamma} (y - \boldsymbol{X}\beta),
\end{aligned}
\tag{4.15}
$$

where $\boldsymbol{H} = \boldsymbol{Z}' \boldsymbol{F} - \boldsymbol{R}$. We can readily observe that the first term of the last identify above is independent of β. As for the second term, we can see that

$$
\begin{aligned}
E \left\{ (e' \boldsymbol{F} + v' \boldsymbol{H}) \boldsymbol{W} \boldsymbol{\Gamma} (y - \boldsymbol{X}\beta) \right\} &= E \left\{ (e' \boldsymbol{F} + v' \boldsymbol{H}) \boldsymbol{W} \boldsymbol{\Gamma} (y - \mu) + (e' \boldsymbol{F} + v' \boldsymbol{H}) \boldsymbol{W} \boldsymbol{\Gamma} (\mu - \boldsymbol{X}\beta) \right\} \\
&= E \left\{ (e' \boldsymbol{F} + v' \boldsymbol{H}) \boldsymbol{W} \boldsymbol{\Gamma} (\boldsymbol{Z} v + e) \right\},
\end{aligned}
\tag{4.16}
$$

and therefore, the second term is also independent of β. Denoting the first and second terms as I_1 and I_2, respectively. We conclude that $MSPE(\hat{\theta}^{BP}) = E \left[(\hat{\theta}^B - \theta)' W (\hat{\theta}^B - \theta) \right] = EQ(\beta)$, where,

$$
Q(\beta) = I_1 - 2I_2 + (\beta' \boldsymbol{X}' - y') \boldsymbol{\Gamma}' \boldsymbol{W} \boldsymbol{\Gamma} (y - \boldsymbol{X}\beta).
$$

Since only the third term is a function of β, this is the only one that must be minimized. From the theory of generalized least squares we can observe that that the minimizer of β is given by:

$$\tilde{\beta}_W^{BPE} = (\boldsymbol{X'\Gamma'W\Gamma X})^{-1}\boldsymbol{X\Gamma'W\Gamma}y, \tag{4.17}$$

provided that $\boldsymbol{X}$ is of full rank and $\boldsymbol{\Gamma'\Gamma}$ is non-singular. Plugging in $\tilde{\beta}^{BPE}$ for β into the best predictor leads to the following expression for the weighted observed best predictor, $\tilde{\theta}_W^{OBP}$,

$$\tilde{\theta}_W^{OBP} = \boldsymbol{F'}y - \boldsymbol{\Gamma}(y - \boldsymbol{X}\tilde{\beta}_W^{BPE}).$$

Example: Rao-Yu Model

The unweighted best predictive estimator for the Rao-Yu model when the variance components are known is given by,

$$\tilde{\beta}_{RY}^{BPE} = \left\{\boldsymbol{X'}(\boldsymbol{I}_{mT} + \boldsymbol{V'ZG'GZ'V}^{-1} - 2\boldsymbol{V'ZG'})\boldsymbol{X}\right\}^{-1}$$

$$\times\boldsymbol{X'}(\boldsymbol{I}_{mT} + \boldsymbol{V'ZG'GZ'V}^{-1} - 2\boldsymbol{V'ZG'})y,$$

$$\text{where } \boldsymbol{G} = diag(\boldsymbol{G_1, \ldots, G_m}), \text{ with } \boldsymbol{G_i} = \begin{bmatrix} A & \mathbf{0'} \\ \mathbf{0} & \sigma^2\boldsymbol{\Lambda}_i \end{bmatrix},$$

$$\boldsymbol{Z} = diag(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m), \text{ with } \boldsymbol{Z}_i = [1_T, \boldsymbol{I}_T],$$

$$\boldsymbol{V} = diag(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m), \text{ with,}$$

$$\boldsymbol{V}_i = [\boldsymbol{\Psi}_i + \sigma^2\boldsymbol{\Lambda}_i + A\boldsymbol{J}_T]^{-1}, \text{ with } \boldsymbol{\Lambda}_i = (\lambda_{t,s}), \text{ where } \lambda_{t,s} = \rho^{|t-s|}/(1-\rho^2).$$

The weighted version of β under the Rao-Yu model, denoted by $\hat{\beta}_{W,RY}^{BPE}$ is

$$\tilde{\beta}_{W,RY}^{BPE} = \left\{\boldsymbol{X'}(\boldsymbol{W} + \boldsymbol{V'ZG'WGZ'V}^{-1} - 2\boldsymbol{V'ZG'W})\boldsymbol{X}\right\}^{-1}$$

$$\times\ \boldsymbol{X'}(\boldsymbol{W} + \boldsymbol{V'ZG'WGZ'V}^{-1} - 2\boldsymbol{V'ZG'W})y.$$

The expression of the best predictive estimator for the Random-Walk model is identical, except with $\mathbf{\Lambda}_i = (\lambda)_{s,t} = min(t, s)$ for $1 \leq s, t \leq T$.

4.7 Empirical Bayes Parametric Bootstrap Stationarity Model Selection

When fitting cross-sectional time series models to real applications, there is no widely-accepted way to choose between the Rao-Yu and Random Walk models. The specification of the auto-regressive component ($\rho = 1$, $|\rho| < 1$) in less complex time series models may apply the Dickey Fuller Unit Root test (Dickey & Fuller (1979)) to evaluate the better model fitness between the unit root model (i.e., random walk) or whether there should be a drift component (i.e., Rao-Yu). Widely used in econometrics, these models are lacking the variance component from the sampling stage (i.e., cross-sectional) in Level 1 of the hierarchical Bayesian models considered in this treatise. We will again use the concept of the EB parametric bootstrap.

What makes this approach novel is that we are able to use the power of the bootstrap for model selection and obtain good approximations to the predictive density for very complex distributions quite easily. Naturally, the covariates should remain fixed between comparisons. We point out that the same technique could be used again for variable selection, either before or after the autoregressive aspect is decided. Once again, the log predictive distribution avoids confounding differences in likelihood functions and produces comparisons between different classes of model which are more viable.

We now present a parametric bootstrap approach for model selection and diagnostics for cross-sectional time series model applied to small area estimation. Two important models in this regard are the Rao-Yu model and the Random Walk model. Let $i = 1, 2, \ldots, m$ denote the m areas within a survey, and let $t = 1, 2, \ldots T$ denote the time periods in which the data y_{it} are collected. For the FH model, it is straightforward to compute $f_{boot}(y_i|y_{(-i)})$. For a complex time series cross-sectional models, direct computation of $f_{boot}(y_i|y_{(-i)})$ could be quite cumbersome.

There are multiple ways to partition the areas and time periods for cross-validation. For example, we could delete data from one entire time point for all areas, or remove data from one area for all time points, or just omit a single data point. The latter would be the most time-consuming with the heaviest amount of computation. It would be more reasonable and informative to first run the predictive distributions based on area and time omissions only. We summarize these choices below:

Predictive Datasets for Cross-Sectional Time Series

- *Area-Based* - $y_{(-i)}$. Remove data from all T time points from the i^{th} area, thus delete y_{i1}, y_{i2}, $\ldots$, y_{iT}.
- *Time-Based* - $y_{(-t)}$. For all m areas, remove data from time point t, so delete points from the i^{th} area, thus delete y_{1t}, y_{2t}, $\ldots$, y_{mt}.
- *Singleton* - $y_{(-it)}$. In the i^{th} area, remove the entry with time point t, thus delete y_{it} only.

Let $y_{(-d)}$ generically denote the data excluding arbitrary d data points and y_d denote the rest of the

data. Note that

$$f(y_d|y_{(-d)}) = \int f(y_d|\theta)f(\theta|y_{(-d)})d\theta,$$

where θ (could be vector) is an appropriately chosen random effects of the model.

Area Based Predictive Dataset

We now proceed to construct an algorithm for computing the bootstrap distribution of y_{it} to approximate the [posterior] predictive distribution from using data partitions omitting observations from domain i. We define the vector $\boldsymbol{y}_i$ as the dataset omitting all observations from the i^{th} area. So all time points $y_{i1}, y_{i2}, \ldots, y_{iT}$ corresponding to area i are excluded. The component estimators $\hat{\beta}_{(-i)}$, $\hat{A}_{(-i)}$, $\hat{\sigma}_{(-i)}$, and $\hat{\rho}_{(-i)}$ correspond to those estimates computed on the reduced dataset that lacks all time observations from the i^{th} area.

Our approach mimics elpd loo hyperprior Bayesian calculations as follows, but the method is parametric bootstrap, which is empirical Bayes. If we delete data from an entire area for time series cross-sectional model, i.e., if $d = i$ then LOO_{boot} would follow

$$\begin{aligned} f(y_{it}|y_{(-i)}) &= \int f(y_{it}|y_{(-i)}, \theta_{it})f(\theta_{it}|y_{(-i)}, \phi)f(\phi|y_{(-i)})d\theta_{it}d\phi \\ &= \int f(y_{it}|\theta_{it})f(\theta_{it}|\phi)f(\phi|y_{(-i)})d\theta_{it}d\phi \end{aligned} \tag{4.18}$$

In the above, we have used the independence assumptions from the time series cross-sectional model. The hyperparameters ϕ are all unknown model parameters such as regression coefficients, variance components or autocorrelation. This motivates us to the following delete-d parametric

bootstrap model selection criterion, the leave one area out:

$$E_* \left[f(y_{it}|\theta^*_{it}) f(\theta^*_{it}|\hat{\phi}^*_{(-i)}) \right] = \frac{1}{K} \sum_{k=1}^{K} \left[f(y_{it}|\theta^{(k)}_{it}) f(\theta^{(k)}_{it}|\hat{\phi}^{(k)}_{(-i)}) \right]$$

The above measures the strength of auxiliary variables in model selection, particularly if the covariates z_{it} are mostly area-dependent rather than time-dependent.

Time Based Predictive Dataset

In the time series cross-sectional case, we are also interested in the strength of the time component. In that case, we delete a time point entirely from all areas, i.e., $d = t$. Note that in this case the Bayesian calculation is given by

$$\begin{aligned} f(y_t|y_{(-t)}) &= \int f(y_t|y_{(-t)}, \theta_t) f(\theta_t|y_{(-t)}, \theta_{(-t)}, \phi) f(\theta_{(-t)}|y_{(-t)}, \phi) f(\phi|y_{(-t)}) d\theta d\phi \\ &= \int f(y_t|\theta_t) f(\theta_t|\theta_{(-t)}, \phi) f(\theta_{(-t)}|y_{(-t)}, \phi) f(\phi|y_{(-t)}) d\theta d\phi \end{aligned}$$

Here again we have used independence assumptions from the model. Note that $f(y_t|\theta_t)$ can be directly obtained from Level 1 and is a multivariate normal. Since θ=($\theta_{11}, \theta_{12}, \ldots, \theta_{mT}$)' is multivariate normal with parameters (β,ϕ), where ϕ contains all variance components, we can obtain $f(\theta_t|\theta_{(-t)}, \phi)$ using multivariate normal properties (the conditional distribution is another multivariate normal). Hence, $f(\theta_{(-t)}|y_{(-t)}, \phi)$ is the probability density function used to obtain the best predictor (the full sample version of it). Indeed, the mean of this distribution for the full sample is the best predictor. This motivates the following delete-d parametric bootstrap model selection:

$$E_* \left[f(y_t|\theta_t^*) f(\theta_t^*|\theta_{(-t)}^*, \hat{\phi}_{(-t)}^*) f(\theta_{(-t)}^*|y_{(-t)}, \phi_{(-t)}^*) \right]$$

$$= \tfrac{1}{K} \textstyle\sum_{k=1}^{K} \left[f(y_t|\theta_t^{(k)}) f(\theta_t^{(k)}|\theta_{(-t)}^{(k)}, \hat{\phi}_{(-i)}^{(k)}) f(\theta_{(-t)}^{(k)}|y_{(-t)}, \phi_{(-t)}^{(k)}) \right].$$

Let us first consider the case when the underlying distribution of the observed data behave according to the Rao-Yu model (4.7).

Example. Rao-Yu Model.

We exemplify the area-based cross-validation procedure using the Rao-Yu model.

$$\begin{aligned} f_{boot}(y_{it}|y_{(-i)}) = E_* \Bigg[& (2\pi)^{-m/2} \det(\boldsymbol{V}_i(\hat{A}_{(-i)}, \hat{\sigma}^2_{(-i)}, \hat{\rho}_{(-i)}))^{-m/2} \\ & \times \exp\left\{ -\frac{1}{2} (y_{it} - z'_{it}\hat{\beta}_{(-i)})' \boldsymbol{V}_i^{-1}(\hat{A}_{(-i)}, \hat{\sigma}^2_{(-i)}, \hat{\rho}_{(-i)}) (y_{it} - z'_{it}\hat{\beta}_{(-i)}) \right\} \Bigg], \end{aligned} \tag{4.19}$$

where $\boldsymbol{V}_i(\cdot) = \Psi_i + \sigma^2 \boldsymbol{\Lambda}_i + A\boldsymbol{J}_T$, with the following definitions for $\boldsymbol{\Lambda}_i$

- $(\lambda)_{t,s} = \rho^{|t-s|}/(1-\rho^2)$ for the Rao-Yu model ($|\rho| < 1$),
- $(\lambda)_{t,s} = \min(t,s)$ for the Random Walk model ($\rho = 1$).

Model Selection for Stationarity

Now that we have expressions for the posterior predictive bootstrap distributions, we are now in a positional to calculate LOO_{boot}. We restate Definition 3.1 from chapter 3.

Definition 3.1. Parametric Bootstrap Leave-One-Out Cross Validation.

$$\text{LOO}_{\text{boot}} = \log\left\{\prod_{i=1}^{m} f_{boot}(y_i|y_{(-i)})\right\} = \sum_{i=1}^{m} \log f_{boot}(y_i|y_{(-i)}), \tag{4.20}$$

1. Select a group of covariates which will be fixed for both the RY and RW models.

2. Compute LOO_{Boot}(RW) and LOO_{Boot}(RY), the empirical Bayes parametric bootstrap leave-one-out predictive distributions.

3. Select RW vs. RY based on the higher value of LOO_{Boot}

Empirical Bayes Parametric Bootstrap for Cross-sectional Time Series

1. Remove the i^{th} area from the dataset. This includes all T time measures from that area, leaving $(m-1)\times T$ observations in the reduced dataset.

2. Compute $\hat{\beta}_{(-i)}$ and the unknown nuisance parameters $\hat{A}_{(-i)}$, $\hat{\sigma}^2_{(-i)}$, and $\hat{\rho}_{(-i)}$ based on the dataset without the i^{th} observation.

3. Generate $y^*_{(-i)} = (y^*_{1(-i)}, \ldots, y^*_{i-1(-i)}, y^*_{i+1(-i)}, \ldots, y^*_{m(-i)})'$, a copy of the dataset by sampling from a multivariate normal distribution for each state $j \neq i$, with mean $x'_j\hat{\beta}_{(-i)}$ and covariance matrix $\boldsymbol{V}_i = \Psi_i + \hat{\sigma}^2_{(-i)}\boldsymbol{\Lambda}_i + \hat{A}_{(-i)}\boldsymbol{J}_T$. Repeat this process K times to obtain $\boldsymbol{y}^{*(1)}_{(-i)}$, $\boldsymbol{y}^{*(2)}_{(-i)}, \ldots, \boldsymbol{y}^{*(K)}_{(-i)}$.

4. For each parametric bootstrap replicate k for the fifty remaining states, re-estimate all parameters to obtain $\hat{\beta}^{*(k)}_{(-i)}$, $\hat{A}^{*k}_{(-i)}$, $\hat{\sigma}^{2*}_{(-i)}$, and $\hat{\rho}^*_{(-i)}$ for $k = 1, \ldots, K$.. This should be the same

estimator used in step 2 to obtain $\hat{\beta}_{(-i)}$, $\hat{A}_{(-i)}$, $\hat{\sigma}^2_{(-i)}$, and $\hat{\rho}_{(-i)}$. This requires using the SAE2 package repeatedly for different bootstrap replicates.

5. Plug $\hat{\beta}^{*(k)}_{(-i)}$, $\hat{A}^{*(k)}_{(-i)}$, $\hat{\sigma}^{2*(k)}_{(-i)}$, and $\hat{\rho}^{*(k)}_{(-i)}$ in to the likelihood function for y_i, which is calculated also using y_i, x_i, and ψ_i. This is the predictive distribution for the k^{th} iteration: $f^{(k)}_{boot}(y_i|y_{(-i)})$. Approximate the predictive distribution for $y_i|\boldsymbol{y}_{(-i)}$ as

$$f_{boot}(y_i|y_{(-i)}) \approx \frac{1}{K}\sum_{k=1}^{K} f^{(k)}_{boot}(y_i|y_{(-i)}).$$

6. Using all of the m predictive distributions in (5), calculate LPP_{boot} as Definition 3.1 and (4.20).

4.8 Concluding Remarks

We reviewed the general linear mixed models, with a couple of time series models that have been investigated as part of the small area literature. We illustrated the usage of the BPE using unemployment estimates and correlates obtained from BLS. We attempted to find the BPE of the variance components and autoregressive parameter in the Rao-Yu model, but were unable to obtain viable estimates due to lack of convergence. Therefore, there is evidence that that the REML/ML methods can provide a more convex surface for vector optimization when multiple parameters are under consideration. We explored an alternative method for deriving the autoregressive parameter using the distributed lag function. We defined the GBP and WBPE for the general linear mixed model scenario, thus extending the results in Chapter 2. Finally, we defined multiple strategies

for time-series leave-one-out cross-validation (Area-based, Time-based, Singleton). A procedure was detailed for using the EBPB LOO-CV methodology for time-series model building, including deciding whether time series data are better fit by the Rao-Yu model (Rao & Yu (1994)) vs. the random walk model of Datta, Lahiri, and Maiti (2002).

Chapter 5

Future Research

5.1 Future Research by Chapter

Chap. 2 Investigate constraints on the weighting matrix $\boldsymbol{W}$ that will improve optimization procedures in the Fay-Herriot model. Identity aspects that could impede convergence of best predictive estimators for variance components. Determine whether the GBP could be leveraged to reduce the propensity of zero estimates for the model variance in the Fay-Herriot model. Find an asymptotic approximation to the variance of the best predictive estimator $\tilde{A}^{BPE}$ of the variance component A in the FH model. Obtain more insight into the performance of the GBP against the EBLUP under general loss functions.

Chap. 3 Conduct a simulation study to observe the performance of the EBPB LOO_{boot} for variable selection, compared to a wider array of model comparison metrics (e.g., AIC, R^2, BIC, etc.). Compare the empirical Bayes predictive distribution LOO_{boot} to the pure Bayesian scenario using Markov Chain Monte Carlo methods.

Chap. 4 Conceive of a methodology to obtain the best predictive estimators for the variance components and autoregressive parameters in for the Rao-Yu model, and demonstrate this approach using the BLS data on unemployment. Use the LOO_{boot} to determine whether the BLS data is better suited for the Rao-Yu or Random Walk cross-sectional time series models. Conduct

simulations to assess the type I and type II errors of the approach. Define an expression for the weighted G matrix within the general linear mixed model. Find properties of weighting matrix $\boldsymbol{W}$ that will provide better stability for obtain the BPE of the variance components within mixed models.

Bibliography

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csáki, F. (eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: *Akadémiai Kiadó*, 267–281. Republished in Kotz, S.; Johnson, N. L., eds. (1992), *Breakthroughs in Statistics*, I, Springer-Verlag, Berlin · Heidelberg · New York.

Akaike, H. (1974), "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, 19 (6), 716–723.

Balabay, O. (2016), em Time Series Modeling in Repeatedly Conducted Sample Surveys, Doctoral Dissertation, Administered and Published by Maastricht University, Printed in the Netherlands by Datawyse Maastricht.

Bandyopadhyay, R. (2017), "Benchmarking the Observed Best Predictor", Doctoral dissertation, University of California, Davis, ProQuest Dissertations Publishing, 2017. 10624146.

Battese, G.E., Harter, R.M., Fuller, W.A. (1988), "An error-components model for prediction of county crop areas using survey and satellite data," *Journal of the American Statistical Association*, 83(401), 28-36.

Bell, W.R., Basel, W.W., Maples, J.J. (2016), "An overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates Program," *Analysis of Poverty Data by Small Area Estimation*,

pp. 349-378.

Bell, W. R., and Franco, C. (2017), Small Area Estimation, State Poverty Rate Model Research Data Files. Available at https://www.census.gov/library/working-papers/2017/adrm/rrs2017-05.html [accessed November 13, 2020].

Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis*. 2nd Edition. Springer-Verlag, New York.

Bhattacharya, P.K. (1966), "Estimating the mean of a multivariate normal population with general quadratic loss function." *Annals of Mathematical Statistics*, 37, 1819-1824.

Buse, A. (1973), "Goodness of fit in generalized least squares estimation.", *The American Statistician* 27, 106–108.

Butar, F. B., (1997), “Empirical Bayes methods in survey sampling", Doctoral Dissertation available at ETD collection for University of Nebraska - Lincoln. AAI9736923. https://digitalcommons.unl.edu/dissertations/AAI9736923.

Butar, F.B., Lahiri, P. (2003), "On measures of uncertainty of empirical Bayes small-area estimators", *Journal of Statistical Planning and Inference*, 112, 63– 76.

Carlin, L., Louis, B. (2009), *Bayesian Methods for Data Analysis*. 3rd Edition. Chapman & Hall, Boca Raton, F.L.

Chatterjee, S., Lahiri, P., Li, H. (2008), "Parametric Bootstrap Approximation to the Distribution of EBLUP and Relatied Prediction Intervals in Linear Mixed Models", *Annals of Statistics*, 36, 1221-1245.

Chen, S., Jiang, J., Nguyen, T. (2015), Observed Best Prediction for Small Area Counts, Journal of Survey Statistics and Methodology, 3(2), 136–161.

Cochran, L., (1977), *Survey Sampling*. John Wiley & Sons, New York.

Cressie, N. (1991). Small Area Prediction of Undercount using the General Linear Model. *Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality*, Ottawa, Canada: Statistics Canada, 93-105.

Das, K., Jiang, J., Rao, J. N. K. (2004), "Mean Squared Error of Empirical Predictor," *Annals of Statistics*, 32, 818–840.

Datta, G.S., Fay, R.A., Ghosh, M. (1991), "Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation". in Proceedings of Bureau of the Census 1991 Annual Research Conference, U.S. Census Bureau, Washington, D.C.

Datta, G.S., Ghosh, M. (1991), "Bayesian Prediction in Linear Models: Application to Small Area Estimation", *Annals of Statistics*, 19, 1748-1770.

Datta, G.S., Lahiri, P., Maiti, T., and Lu, K.L. (1999), "Hierarchical Bayes estimation of unemployment rates for the U.S. states", *Journal of the American Statistical Association*, 94, 1074-1082.

Datta, G.S., Lahiri, P., Maiti, T. (2002), "Empirical Bayes Estimation of Median Incomes in Four-Person Families by State Using Times Series and Cross-Sectional Data", *Journal of the Statistical Planning and Inference*, 102, 83-97.

DeGroot, M.H. (1970), *Optimal Statistical Decisions*, Wiley & Sons, Hoboken, NJ.

Deming, W.S., Stephan, F.F. (1940), "On a Least Squares Adjustment of a Sample Frequency Table when the Expected Marginal Totals are Known", *Annals of Statistics*, 11, 427-444.

Diallo, M. (2014), *Small Area Estimation: Skew-Normal Distributions and Time Series.* Unpublished Ph.D. dissertation. Carleton University, Ottowa, Canada.

Dickey, D. A., Fuller, W. A. (1979), "Distribution of the Estimators for Autoregressive Time Series with a Unit Root". *Journal of the American Statistical Association*. 74 (366): 427–431.

Efron, B., Morris, C. (1973), "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach". *Journal of the American Statistical Association*. 68 (341), 117–130.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife", *Annals of Statistics*, 7 (1), 1-21.

Efron, B., Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Egerton, M. F., Laycock, P. J. (1982), "An Explicit Formula for the Risk of James-Stein Estimators", *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 10(3), 199-205.

Erciulescu, A., Franco, C., Lahiri, P. (2020), “Use of Administrative Records in Small Area Estimation.” To appear in Chun, A. Y. and Larsen, M. (Eds.) *Administrative Records for Survey Methodology*, Wiley, New York.

Fay III, R.E., Herriot, R.A. (1979), "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74(366a), 269-277.

Fay III, R.E., (1987), Platek, R., Rao, J.N.K., Särndal, C.E., Singh, M.P. (Editors), "Application of Multivariate Regression to Small Domain Estimation, in R". *Small Area Statistics*. John Wiley & Sons, New York.

Fay III, R.E., Diallo, M. (2015), *Small Area Estimation: Time-series Models*. SAE2 Package in R Statistical Software.

Fuller, W.A. (2009), *Sampling Statistics*. John Wiley & Sons, Hoboken, NJ.

Gelfand, A.E., Dey, D.K. and Chang, H. (1992), “Model determination using predictive distributions with implementation via sampling-based methods". In: *Bayesian Statistics*, 4 (J. Bernardo et. al., eds.). Oxford University Press. Oxford, 147-167.

Gelman, A., Carlin, J.B., Stern, H.A., Rubin, D.B. (2004), *Bayesian Data Analysis*. Chapman & Hall, Boca Raton, F.L.

Gelman, A., Hwang, J., Vehtari, A. (2013), “Understanding predictive information criteria for Bayesian models", *Statistics and Computing*, 24, 997-1016.

Hall, P. (2003), “A Short Prehistory of the Bootstrap”, *Statistical Science* 18 (2), 158-167.

Hall, P., Maiti, T. (2006), "On parametric bootstrap methods for small area prediction," *Journal of the Royal Statistical Society: Series B*, 68, 221-238.

Hansen, M.H., Hurwitz, W.N., Madow, W.G. (1953), *Sample Survey Methods and Theory*. John Wiley & Sons, New York.

Hartley, H.O., Rao, J.N.K. (1967), "Maximum likelihood estimation for the mixed analysis of variance model", *Biometrics*, 34, 233-234.

Henderson, C.R. (1950), "Estimation of Genetic Parameters" (*Abstract*), *Annals of Mathematical Statistics*, 21, 309-310.

Henderson, C.R. (1953), "Estimation of Variance and Variance Components", *Biometrics*, 9, 226-252.

Henderson, C.R. (1975), "Best Linear Unbiased estimation and prediction under a selection model". *Biometrics*, 31, 423-447.

Hosmer, D.W., Lemeshow, S. (2000), *Linear Models in Statistics*, New York: J. Wiley and Sons.

Hsiao, C. (2014), *Analysis of Panel Data.* 3rd ed, Cambridge University Press, New York.

James, W., Stein, C. (1961), "Estimation with quadratic loss", Proc. Fourth Berkeley Symp. Math. Statist. Prob., 1, pp. 361–379.

Jiang, J. and Lahiri, P. (2006), Mixed model prediction and small area estimation. (with discussion), *TEST*, 15, 1–96.

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

Jiang, J., Nguyen, T., Rao, J.S. (2011), "Best Predictive Small Area Estimation." *Journal of the American Statistical Association*, 106, 732-745.

Jiang, J., Nguyen, T., & Rao, J. S. (2015), Observed best prediction via nested-error regression with potentially misspecified mean and variance. Survey Methodology, 41(1), 37-55.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., and Lee, T.C. (1985), *The Theory and Practice of Econometrics* 2nd ed, John Wiley & Sons, New York.

Kish, L. (1965), *Survey Sampling*. John Wiley & Sons, New York.

Korn, E.L., Graubard, B.I. (1999), *Analysis of Health Surveys*, Wiley & Sons, New York.

Laird, N., Louis, T., (1987), “Empirical Bayes Confidence Intervals Based on Bootstrap Samples", *Journal of the American Statistical Association*, 82 (399), 739-750.

Lahiri, P. (2003), "On the impact of the bootstrap in survey sampling and small-area estimation," *Statistical Science*, vol. 18, pp. 199-210.

Lahiri, P. (2020), "Hierarchical Bayesian and Parametric Bootstrap Methods for Small Domains", Research Proposal Submitted to the National Science Foundation.

Levin, A., Lin, C., Chu, C. (2002), “Unit root tests in panel data:asymptotic and finite-sample properties", *Journal of Econometrics* 108, 1–24.

Li, H., Lahiri, P. (2010), "Adjusted maximum method for solving small area estimation problems", *Journal of Multivariate Analysis*, 101, 882-892.

Lohr, S. (2010), *Sampling: Design and Analysis*. 2nd Edition. Brooks-Colee, Boston, MA.

Morris, C.N. (1983), "Parametric Bayes Inference: Theory and Application," *Journal of the American Statistical Association*, 78, 47-55.

National Center for Health Statistics. (2018), "Variance Estimation Guidance, NHIS 2016-2017.pdf". Downloaded on 10/17/2020 from https://www.cdc.gov/nchs/data/nhis/2016var.pdf.

Otto, M.C., Bell, W.R. (1995), “Sampling Error Modelling of Poverty and Income Statistics for States,” Proceedings of the American Statistical Association, Section on Government Statistics, 160-165, Alexandria, VA: American Statistical Association.

Petrucci, A., Salvati, N. (2006), “Small area estimation for spatial correlation in watershed erosion assessment," *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 169-182.

Pfeffermann, D., Barnard, C. (1991), "Some New Estimators for Small Area Means with Applications to the Assessment of Farmland Values", *Journal of Business and Economic Statistics*, 9, 73-84.

Pfeffermann, D., Burck, L. (1990), "Robust Small Area Estimation Combining Time Series and Cross-Sectional Data", *Survey Methodology*, 16, 217-237.

Pfeffermann, D. (2013), "New important developments in small area estimation," *Statistical Science*, vol. 28, pp. 40-68.

Prasad, N.G.N., Rao, J.N.K., (1990), "The Estimation of the Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85, 163-171.

Pratesi, M., Salvati, N. (2008), “Small area estimation: the EBLUP estimator based on spatially correlated random area effects," *Statistical Methods and Applications*, 17, 113–141.

Purcell, N.J. and L. Kish (1979), "Estimation for Small Domains," *Biometrics*, 35, 365-384.

Purcell, N.J., Kish, L. (1980), "Postcensal for Local Areas or (Domains)", *International Statistical Review*, 48, 3-18.

Rao, J.N.K., Molina, I. (2015), *Small Area Estimation*, John Wiley and Sons, Hoboken, NJ.

Rao, J.N.K., Yu, M. (1994), "Small area estimation by combining time series and cross-sectional data," *Canadian Journal of Statistics*, 22, 511-528.

Rao, J.S. (2018), "Observed Best Prediction for Small Area Estimation: A Review", *Statistics and Applications*, {ISSN 2452-7395 (online)}, 16(1), (New Series), 305-314.

Rencher, A.C. (1999), *Linear Models in Statistics* J. Wiley, New York.

Robinson, G.K. (1991), "That BLUP is a Good Thing: The Estimation of Random Effects". *Statistical Science*, 1, 15-51.

Saegusa, T., Sugasawa, S., Lahiri, P. (2020), "Parametric Bootstrap Confidence Intervals for the Multivariate Fay-Herriot Model", to appear in the *Journal of Survey Statistics and Methodology*.

Särndal, C.E., Swensson, B., Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

SAS Software (2020), The [output/code/data analysis] for this paper was generated using SAS software version 9. Copyright © 2020 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Searle, S.R. (1990), "C.R. Henderson the Statistician and his Contributions to Variance Components Estimation", 85th Annual Meeting of the American Dairy Science Association, Raleigh, N.C. June 25-26.

Searle, S.R., Casella, G., McCulloch, C.E. (2006), *Variance Components*, Wiley, New York.

Shibata, R. (1989), “Statistical aspects of model selection", In *From Data to Model*, ed. J. C. Willems, 215–240. Springer-Verlag Berlin · Heidelberg.

Stan Development Team (2016a), The Stan C++ Library, version 2.10.0. http://mc-stan.org/.

Stan Development Team (2016b), RStan: the R interface to Stan, version 2.10.1.

Stapleton, J.H. (1995), *Linear Statistical Models*, J. Wiley, New York.

Stein, C.M., (1956), "Inadmissibility of the usual estimator for the mean of a multivariate distribution", Proc. Third Berkeley Symp. Math. Statist. Prob., 1, 197–206.

Stein, C.M. (1966), "An approach to the recovery of inter-block information in balanced incomplete block designs". *Research Papers in Statistics*: Festschrift for J. Neyman (F.N. David, ed.), Wiley, London, 351-366.

Stone, M. (1977), “An asymptotic equivalence of choice of model cross-validation and Akaike’s criterion", *Journal of the Royal Statistical Society B*, 36, 44-47.

Sugasawa, S., Tamae, H., Kubokawa, T. (2017), "Bayesian estimators for small area models shrinking both means and variances," *Scandinavian Journal of Statistics*, 44(1), 150-167.

Thompson, W.A. (1962),"The problem of negative estimates of variance components". *Annals of Mathematical Statistics*, 33(1), 273-289.

Tiller, R. (2001), “Seasonal Adjustment of CPS Time Series with Large Survey Errors,” paper presented at the Federal Economic Statistics Advisory Committee Meeting, Washington, DC, December 13–14.

Tiller, R. (2005), “Model-based seasonally adjusted estimates and sampling error,” downloaded 11/6/2020 from the Bureau of Labor Statistics website: https://www.bls.gov/opub/mlr/author/tiller-richard-b.htm.

Tiller, R.B., Evans, T.D. (2018), “Seasonal Adjustment Methodology for National Labor Force Statistics from the Current Population Survey (CPS),” viewed 11/12/2020 from the Bureau of Labor Statistics website: https://www.bls.gov/cps/seasonal-adjustment-methodology.htm

Van der Brakel, J., Roels, J. (2010), "Intervention analysis with state-space models to estimate discontinuities dues to a survey design", *Annals of Applied Statistics*, 4, 1105-1138.

Watanabe, S. (2010), “Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory", *Journal of Machine Learning Research* 11, 3571-3594.

Yoshimori, M., Lahiri, P. (2014), "A new adjusted maximum likelihood method for the Fay-Herriot small area model", *Journal of Multivariate Analysis*, 124, 281-294.

You, Y. (2008), “An Integrated Modeling Approach to Unemployment Rate Estimation in Sub-Provincial Areas of Canada", *Survey Methodology*, 34, 19-27.