EXTENDED ENSEMBLE ESTIMATION: A TOOL FOR SENSITIVITY ANALYSIS FOR RESEARCHERS

By

Jordan Tait

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods – Doctor of Philosophy

2022

ABSTRACT

Once research questions are posed, researchers must answer many aPriori questions regarding research design before analysis can be performed and any conclusions can be made, including sample selection criteria, data collection method, model specification, analysis and estimation technique. The choices made by researchers along this forking path of possible specifications, such as model specification and estimation technique, can lead to varying results. On one hand, researchers that are not able to identify the best answers to these questions are faced with a list of plausible specifications, inherent uncertainty in resulting model estimates, and are tasked with how to best balance alternative specifications. On the other hand, researchers suffer from an "embarrassment of riches" in computational capacity, in that they have more computational power than what is reflected in most journal articles and that the amount of alternative analyses researchers could perform have expanded dramatically (Young, 2018).

Although it is common to choose one set of specifications and report the resulting estimates in the absence of other specifications, this research proposes a framework called extended ensemble estimation that utilizes the alternative specifications to quantify and visualize the sensitivity of an estimated treatment effect. This paper also proposes a method to combine the estimated treatment effects across specifications into a single estimated treatment effect, weighted by precision. Along with the proposed methodology, this work contains a best practice guide for users in order to best understand the sensitivity of an estimated treatment effect within the extended ensemble estimation framework, a proposed method to update an estimated treatment effect by utilizing alternative specifications, simulated performance within common covariance structures, and a case study application regarding the effects of kindergarten retention on math and reading performance.

ACKNOWLEDGEMENTS

Though rightfully one of the most challenging and taxing journeys I have endured, I can wholeheartedly say that I do not believe I would be at this point in my life without all of the support I have received from professors, family, friends, and colleagues. It is at this moment that I would like to truly say how thankful I am for having them in my life and how crucial they were throughout my journey, from start to finish

I would like to thank my fiancée, Christine Pacewicz. Your unconditional love and consistent support gave me the confidence to persist through the toughest moments and remined me to celebrate the best moments. You always believed in me and that meant more to me than I can ever accurately express. I could not have accomplished this without you.

I would like give my deepest recognition and thank you to my advisor, Dr. Ken Frank. While professors are not all created equal and few are greater than the sum of their parts, you have been one of the most positively influential people in one of the most challenging times in my life. You routinely showed me how to not only perform good research, but how to be a better person. As I look forward to mentoring many students throughout my career, I strive to leave such a lasting impression on them as you have left on me. Thank you for always believing in me and for showing me what it means to not only be an excellent advisor and role model, but how to be genuine and good to others.

I would like to also thank my committee, Dr. Jeffery Wooldridge, Dr. Kimberly Kelly, and Dr. Kaitlin Knake. Although it is through their expertise that my research has risen to this point, I can safely say they have all contributed more to my journey than just their expertise alone. The many semesters of econometrics with Dr. Wooldridge have been some of the most challenging and rewarding experiences during my time at Michigan State, and experiences that I routinely

reference. The opportunity to learn about machine learning under the guidance of Dr. Kelly proved to be a valuable experience for my dissertation and it is through her support that I garnered the confidence to pursue computational research. With many prior years of teaching experience, I may have experienced the most growth during my time working with Dr. Kaitlin Knake with Teachers in Social Media. It is through the opportunities and experiences Dr. Knake provided me that helped me grow into a well-rounded professional and mentor for my future students. I cannot thank my committee members enough for their contributions.

Finally, I would like to give a special thanks to my friends and family, especially my parents, Allan Tait and Ginger Tait, and my brothers, Jon Tait and Jarrad Tait. You have always provided me with unconditional support and through your actions, always served as a reminder of how big of an influence a loving family can have. I would not have made it this far if it weren't for you all, and for that, I want to thank you for always being there for me.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Motivation

The process of using quantitative methods to answer research questions is multifaceted, requiring due diligence. Once testable hypotheses have been generated, often leveraging existing theory and past research, an encompassing research design must be formulated in order to provide a blueprint that details critical components such as sample selection, measurement instruments, model selection and estimation technique (De Vaus 2001). Together, the various choices that could be made at each step of the research design process create model uncertainty (Young 2018). That is to say that there isn't necessarily a single correct sequence of choices to be made by a researcher for a given study regarding research design. Gelman and Loken 2014 describe this process as a "garden of forking paths", where various sampling methods, models and estimation techniques could result in thousands of possible specifications. In the complete absence of nefarious motives, we may be unable to identify the most appropriate specifications, resulting in a level of uncertainty from the list of seemingly plausible specifications. On the other hand, Broaduer et. al 2016 claims that researchers have incentives to find statistically significant results, thus favoring specifications that result in statistical significance. In either case, researchers a like seem to be aware that the choices they make at the design level may lead to varying results.

In the case of model selection, omitted variable bias is one common potential issue researchers aim to address. In the case of the Ordinary Least Squares (OLS) estimator, omitted variable bias means a violation of the assumptions of OLS. This particular violation prevents the OLS estimator from converging in probability to the true parameter, causing the OLS estimator

to be biased and inconsistent. For example, if an omitted variable is positively correlated to a regressor and dependent variable, the resulting OSL estimate of the aforementioned regressor will be inflated (Clark, 2005; Green, 2003; Wooldridge 2009). If focus was directed towards the estimated effect of a single covariate, e.g. a treatment variable in a randomized control trial, the treatment variable would be the regressor in question regarding omitted variable bias and whose estimate and standard error would come under scrutiny among authors, fellow researchers and peers in the peer review process. Of course, model selection is one choice of many that a researcher must make that may directly change the resulting estimate of a regressor of interest.

Research on the effects of school suspensions by Craigie 2022 discuss this issue directly, stating:

> A pointed evaluation of insubordination/disrespect and student-aimed obscene language infraction does not indicate statistically significant changes in Out of School Suspension (OSS) outcomes in response to the second reform. However, when disorderly conduct is differentiated from the other disruption-specific infractions, the impact of the second reform on OSS duration is now statistically different from zero, increasing the average OSS duration by 0.17 days.

The above transparency of a common issue Craigie 2022 is reflective of conversations between authors, researchers, and peer reviewers. In the above case, two different model specifications resulted in two different values for the estimated effect for the treatment variable of interest, ultimately resulting in two different conclusions. While the issue of deciding which specification is most appropriate may seem apparent, the act of choosing one specification and ultimately disregarding the rest of the pool of plausible specifications poses an immediate

follow-up issue. As pointed out earlier, the pool of plausible specifications can grow rather quickly, exacerbating the issue of choosing the "right" specification.

In the case of published research, one estimate and conclusion are often chosen from the pool of plausible estimates to be reported, in the absence of other estimates. This essentially masks the model uncertainty. That is, once a specification is made and an estimate is reported, it is not known whether the estimated effect varied wildly across model specifications or remained relatively constant. Specifically, the sensitivity of the estimated effect across model specifications is not frequently reported. What if the original inference does not hold in one of the plausible specifications? What if the specification that overturns the original inference was associated with a relatively large standard error? Does the estimated effect of a treatment variable remain constant across plausible specifications? If the estimated effect of interest varies, are there commonalities among specifications that contribute to variability in estimates?

Instead of creating a sharp precipice around a single decision regarding specification, this study proposes a method to utilize the variability of estimates and standard errors, resulting from plausible specifications and estimation techniques, in order to better understand the effect model uncertainty has on the estimated effect of a treatment variable. Visualizing and quantifying the uncertainty across the estimated effects of a treatment variable creates a framework that could stand to promote discourse regarding robustness and sensitivity.

1.2 Extended Ensemble Estimation

In a broader sense, Ensemble Estimations increase transparency between authors and potentially skeptical readers who want to see more than the authors preferred results (Young 2018). Extended Ensemble Estimation does this by expanding what estimates are provided to the reader, bridging the gap between the authors research efforts and a skeptic wondering what

would have happened under a plausible alternative specification, differing from that of the authors.

More narrowly, Extended Ensemble Estimation is a method of analysis that aims to address the issues of model uncertainty during the quantitative research process, specifically regarding the estimated effect of a treatment variable. Extended Ensemble estimation is, in part, the process of visualizing, quantifying, and utilizing the uncertainty of the estimated effects of a treatment variable that may arise due to specifications made during the quantitative research process, such as but not limited to model specification, estimation technique and weighting scheme.

Extended Ensemble Estimation occurs after initial analysis has been executed and an initial estimated effect and corresponding standard error for the treatment variable have been reported, as a result of specifications that have been discussed, vetted and supported through past research, findings and literature. After the original estimated effect has been documented, a pool of plausible, alternative specifications are required to achieve alternate estimated effects and standard errors of the treatment variable. The specifications may include, but are not limited to, alternative covariate selection, change in estimation technique, or choice of standard error calculation. The quality of the pool of alternative, plausible specifications has been well documented to be a main driver in the quality of ensemble estimation in general, while carefully selected specifications can improve results (Saez-Rodriguez et al., 2016). Specifically, as the quality of the alternative models plays a large role in terms of robustness of the model averaging approaches, increasing the amount of alternative reasonable models plays a lesser role (Stumpf, 2021).

1.3 Literature Review

Methods that focus around combining numerous models have a relatively long history in various fields, such as computer science, Bayesian statistics and econometrics (Laan et al., 2017). Statistical learning techniques, such as deep neural networks, have risen in popularity in tandem with the rise in computational power (Su & Chen, 2015) which Young (2018) described as an "embarrassment of riches". Boosting, an ensemble algorithm based in machine learning, shares history with older machine learning techniques that fall under the supervised learning family of algorithms that aim to reduce bias and variance (Breiman, 1996). Bayesian model averaging has roots in theoretical statistics, following the basis of Bayesian statistics in order to achieve similar goals of combining multiple models into a better, single predictive model (Raftery et al., 1997; Raftery et al. 2005; Wasserman, 2000).

1.3.1 Model Averaging in Computer Science

In many computer science-based methods, such as Deep Neural Networks, the main goal is prediction of a dependent variable or multiple dependent variables (Hofman et al., 2021; Murphy, 2012; Wasserman 2000). Thus, measures of performance rely on and center around the prediction of the dependent variable, such as accuracy, precision, recall, MSE, RMSE, True positive rate, False positive rate and F1 score. The pool of parameters used to best predict the dependent variable are derived from minimization/maximization techniques, such as gradient decent. In many computer science-based methods, the dataset in question is split into two subsets; a training set and a test set. Gradient decent is applied on the training set to determine the best subset of parameters, along with any possible combination of interactions and layers, that best predict the dependent variable by minimizing a defined loss function. Once the best parameters, layers and interactions are determined, the model is used to predict values of the

dependent variable within the test set and compared to the actual values of the dependent variable within the test set. The measures of performance are derived directly from these values; the predicted values and the actual values of the dependent variable within the test set. By construction, a model with an arbitrary amount of interaction terms and layers that may be riddled with unknown dependencies, makes inferential statistics regarding the parameters extremely difficult at best, and near impossible at worst. That is, it is often not possible to understand the strength of relationships between parameters, which parameters are statistically significant, or achieving confidence intervals for parameters. Extended Ensemble Estimation is more narrow and carefully constructed to focus on inference for a single parameter of interest. Quantifying the sensitivity or robustness a single treatment effect of interest in order to better serve the conversation around causality is the main objective of Extended Ensemble Estimation.

1.3.1 Bayesian Model Averaging and Model Selection

Wasserman (2000) details Bayesian methods of comparing model performance, as well as averaging predictions from several models. Similar to many methods grounded in computer science, the goal is prediction of a dependent variable and performance is measured at a model level. Wasserman (2000) points out that the many Bayesian methods involve the computation of posterior distributions which are heavily reliant on prior distribution selection. Although Robust Bayesian methods (Berger 1990, Berger and Delempady, 1987) that focus on a set of priors, as opposed to single prior, are in the same spirit as Extended Ensemble Estimation, they do not discuss sensitivity or robustness in terms of estimation technique or model selection.

Raftery et al. (2005) discusses two approaches to Bayesian model averaging in order to account for model uncertainty. The first approach is to apply Occam's window algorithm (Madigan and Raftery, 1994) to linear regression models, where models are selected based on their ability to

predict. Namely, if a model predicts particular data poorly, it is not considered. On the other hand, models with high posterior probabilities are kept for model averaging for the goal of prediction. The second approach is to apply Markov Chain Monte Carlo model composition of Madigan et al. (1995) by considering a pool of models, plus all models with either one covariate fewer or one extra covariate than those in the pool of models. In this approach, as well as the previous approach, uncertainty due to estimation technique and sensitivity in model selection are not quantified or addressed.

1.3.2 Model Averaging and Ensemble Estimation in Economics

Belloni et al (2016) explored the problem of generalized linear models in the presence of a pool of possible controls while examining a single effect of interest, which resulted in a method that allows for the estimation of a single parameter of interest that is robust to model selection mistakes regarding control selection. Their method uses a three-step approach; estimating the part of the regression function associated with the controls via post model selection, estimating an optimal instrument via post model selection, and the combination of these two steps to establish estimating equations that are robust against crude estimation of nuisance functions. One benefit of this technique is the established $\sqrt{n}$-consistency and asymptotic normality of estimators under high level conditions of nuisance parameters. In a sense, Belloni et al (2016) proposed a static solution to the covariate selection problem that achieves desirable properties under certain conditions. Similar to the previous approaches, this approach performs in the absence of uncertainty due to estimation technique and does not describe or quantify sensitivity or robustness of specification choices.

1.4 Ties to Sensitivity Analysis

Uncertainty is inherent within statistical inference. Hypothesis testing and significance testing rely on being able to quantify uncertainty in order to make a decision regarding a null hypothesis. Even in the extreme case of randomized control trials, researchers are not able to make deterministic claims regarding treatment effects due to inherent levels uncertainty. While non-experimental data is often easier to obtain, it is often very difficult or impossible to disentangle correlation from causation, as Holland (1986) points out. On both ends of the spectrum, well controlled experiments and observational studies both suffer from the lack of ability to confidently control for every possible alternative explanation or being able to account for every possible confounding effect.

Instead of using dichotomous decision making, as is the case with statistical significance, sensitivity analysis takes a more general approach to determine how much conditions must change in order to change the statistical inference at hand. If the original inference has been rejected based on the conducted hypothesis test, what alternative specifications could lead to a failure to reject the inference and how similar or different are those alternative specifications from the original specifications. If the same conclusion is made regarding the original inference under alternative specifications, the original inference is said to be robust and may be evidence of a causal relationship. If the original conclusion changes in the presence of an alternative specification, the original inference is said to be sensitive to specification, to the degree to which the alternative specification is similar or different to the original specification. Instead of running into a dead end in the research process by not including particular covariates, the conversation of a potential causal relation can continue by answering the question “How does the estimated effect change under alternative specifications?”.

Frank (2000) developed an index that measures the required impact a potential confounding variable would need in order to change the original inference. This process centers around correlations of independent variables, dependent variables, and a posited confounding variable since hypothesis tests for regression coefficients are equivalent to those of correlations (Cohen, West & Aiken, 2014). The required impact a confounding variable would need in order to change the original inference is given by a simple expression

$$TICV_{r_{xy}} = \frac{r_{xy} - r_{xy}^{\#}}{1 - r_{xy}^{\#}}$$

In this expression for the threshold for the required impact, $TICV_{r_{xy}}$, the correlation of x and y is given by r_{xy} while $r_{xy}^{\#}$ denotes the threshold for statistical significance. Frank (2000) also extends this expression to account for additional covariates, g, with the follow-up expression

$$TICV_{r_{xy}} = \left(\sqrt{(1 - r_{xg}^2)(1 - r_{yg}^2)}\right)\left(\frac{r_{(xy|g)} - r_{xy}^{\#}}{1 - r_{xy}^{\#}}\right)$$

In this expression, r_{xg} and r_{yg} are the multiple correlations between x & g and y &g, respectively, while $r_{(xy|g)}$ is the partial correlation of x and y given g. Frank (2000) argues that quantifying this sensitivity allows the researcher to respond to critiques concerned about lack of control covariates by quantifying how large of an impact the missing covariates must impart in order to change the inference.

Frank (2013) introduces a measure of bias needed in order to change an inference, namely

$$\frac{\hat{\delta} - \delta^{\#}}{\hat{\delta}} = 1 - \frac{\delta^{\#}}{\hat{\delta}}$$

Where $\delta^{\#}$ is the threshold effect for statistical significance and $\hat{\delta}$ is the estimated effect. In this formulation, this represents the proportion of bias necessary to invalidate an inference.

Using a case-replacement framework, this answers the question of how many cases one would need to replace in the data with counterfactual, zero effect cases in order to chance the inference at hand. Frank points out that this method of measuring sensitivity by expressing sensitivity in terms of the units of observation instead of variables is more appealing than other forms of sensitivity analysis. For example, in the context of schools, language in terms of students and schools may be more appealing and may facilitate conversations surrounding causality more effectively than language centered around technical details.

Work by Emily Oster (2019) examines sensitivity analysis through the lens of selection on observables and unobservables in order to quantify changes in estimated treatment effects and R-squared. In order to correct for biased treatment effects, perhaps due to omitted variables, Oster offers a value, δ, that is the relative importance of unobservables compared to observables that would be required to invalidate the inference, called the coefficient of proportionality. Namely,

$$\delta \frac{\sigma_{1x}}{\sigma_1^2} = \frac{\sigma_{2x}}{\sigma_2^2}$$

Where σ_{1x} is the covariance between treatment and observables, σ_{2x} is the covariance between treatment and unobservables, σ_1^2 is the variance of the observables and σ_2^2 is the variance of unobservables. Equal selection where both are equally important would be represented by $\delta = 1$. One working assumption is that the relationship between treatment and unobservables can be recovered from the relationship between treatment and observables. Oster notes that although coefficient stability is related to the coefficient of proportionality, that it is possible for coefficients to be completely unchanged in the presence of large bias.

Although these forms of sensitivity analysis address issues of bias from unobserved sources, such as omitted variables, they do not account for variation in estimation technique or

the precision of the estimated treatment effects across alternative specifications. That is, creating an estimated treatment effect weighted by precision across various alternative specifications, including estimation technique and model selection, remains largely unexplored.

Specifications within the analysis phase fall into one of two categories. The first case is where it is impossible to achieve the desired re-specification. Examples of this type include cases where researchers may not have access to an omitted variables or are not aware of such variables. Even in perfectly controlled experiments that are typically thought of as "gold standards", the search for possible confounding variables never ceases, even when random assignment is possible (Cook 2002). The second category is where it is possible to examine alternative specifications. These scenarios include but are not limited to the ability to change model specification, estimation technique, and to consider various subsamples. While the work by Frank (2000,2013) and Oster (2019) address scenarios in the first case, regarding unobservables and counterfactuals, Extended Ensemble Estimation addresses the second case where changes in specification are possible.

1.5 Goals of Extended Ensemble Estimation

The main objective of Extended Ensemble estimation is to provide a broader picture in terms of the potential causal relationship between a treatment effect and outcome by quantifying the robustness and sensitivity of an estimated treatment effect relative to alternative specifications. In order to achieve this objective, there are three primary goals of Extended Ensemble Estimation.

1) Compare the original estimated treatment effect to the estimated treatment effect under plausible alternative specifications.

2) Form a distribution of estimated treatment effects from the plausible alternative specifications, using the shape, center and spread to quantify robustness or sensitivity.
3) Combine the estimated treatment effects from the original specification and plausible alternative specifications, based on precision, to achieve a single estimated treatment effect, called the Extended Ensemble Estimate.

Accomplishing these goals gives the opportunity to observe how the estimated treatment effect changes in the presence of alternative covariate selectin, alternative estimation methods, subsample selection, or other alternative specifications. By visualizing the changes in estimated treatment effects with an empirical distribution and quantifying the robustness or sensitivity, Extended Ensemble Estimation provides answers to researchers or readers who may question what would happen to the estimated treatment effect under different circumstances than those chosen by the authors.

1.6 Summary of findings

This study proposes a within-study procedure to utilize estimated treatment effects from plausible alternative specifications to better understand the potential causal relationship between a treatment and outcome. The combined and precision-weighted estimate provided by the Extended Ensemble Estimation framework is used in tandem with the empirical distribution of estimates from alternative specifications in order to promote discourse surrounding the potential causality of the estimated treatment effect. While Belloni et al (2016) proposes an estimator that addressed control selection, Extended Ensemble Estimation is able to take other model specifications into account, such as estimation technique. In order to best serve the discourse around the estimated effect of a parameter of interest, the empirical distribution of estimates can be further extended to include other specifications, such as the estimated effect produced by

Belloni et al (2016). By removing potential subjectivity regarding specification, extended ensemble estimation also provides a framework as a safeguard against common statistical pitfalls such as p-hacking and cherry picking.

1.7 Structure of Study

In the next Chapter, I will detail the general methodology for Extended Ensemble Estimation where I will discuss the roles of estimation techniques in general, distribution of estimated treatment effects, measures of central tendency and standard errors relating to the estimated treatment effects. Ordinary Least Squares and Instrumental Variables will be compared and contrasted in terms of their application within Extended Ensemble Estimation. The chapter will end with how to incorporate precision of estimated treatment effects. In Chapter 3, I will discuss a method to update an estimated treatment effect using Extended Ensemble Estimation. I will talk about the general approach to updating an estimated treatment effect, ties to empirical Bayesian methods, and the sensitivity regarding the selection of weighting scheme. Next, Chapter 4 will serve as a guide of best practices during the Extended Ensemble Estimation process. I will discuss how the end user may proceed to best serve the conversation around a potential causal treatment effect while avoiding common statistical pitfalls such as cherry-picking results and p-hacking. Chapter 5 will focus on using simulation to observe how Extended Ensemble Estimation performs under various conditions, including sensitivity regarding sample size, within a Randomized Control Trial Ancova design, in the presence of a strong and weak pre-test, and in the presence of a strong and weak instrumental variable. Finally, Chapter 6 will detail the use of Extended Ensemble Estimation through a case study regarding Kindergarten Retention and work by Hong and Raudenbush (2005). I will use Extended Ensemble Estimation to compare the estimated effects of kindergarten retention on student's scale reading and math

scores by Hong and Raudenbush, with estimated treatment effects under various alternative specifications in order to serve the conversation regarding the potential causal effect of retaining kindergarten students on reading and math scores.

CHAPTER 2

GENERAL METHODOLOGY FOR EXTENDED ENSEMBLE ESTIMATION

2.1 General Framework of Extended Ensemble Estimation

This section will lay out the components involved in using extended ensemble estimation. The components discussed regarding the estimated treatment effects will include estimation techniques in general, the distribution of estimated treatment effects, central tendency measures, as well as standard errors of the estimated treatment effects themselves. This section will also discuss two particular estimation techniques, namely ordinary least squares and instrumental variables. Explanation of these estimation techniques and the role they play in extended ensemble estimation will be followed by how weighting can be utilized, including using the standard errors of the estimated treatment effects to achieve a combined estimated treatment effect that is weighted for precision using a meta-analysis style approach.

2.1.1 Role of Estimation Technique

In order to compare estimates of a treatment effect across various specifications, a choice must be made that will determine how each treatment effect will be estimated, given a specified model. Among the possible choices for estimators are the more common, but not limited to, Least Squares, Maximum Likelihood, Bayes, and Markov Chain Monte Carlo. Particular desirable properties can help researchers determine which estimator to use, such as unbiasedness, minimum variance unbiasedness (MVUE) or best linear unbiased (BLUE). There may be cases that are able to satisfy conditions for some estimators with particular properties, while failing to satisfy the same conditions in other scenarios, thus researchers may pick an estimation technique based on the conditions they can comfortably satisfy or avoid an estimation technique that is not robust in the presence of failed conditions that can be hard to hold or justify, such as general

independence or random sampling. In terms of understanding the sensitivity or robustness of a treatment effect, one may observe different estimated treatment effects based on the chosen estimation technique. In that sense, estimation technique can be taken into account as a specification in the extended ensemble estimation framework by estimating the treatment effect using various estimation techniques in order to observe possible sensitivity or robustness.

2.1.2. Distribution

Once estimation of the treatment effect has been carried out for the various model specifications, visually inspecting the estimated treatment effects as a distribution can reveal any present sensitivity to model specification, or robustness. Namely, a distribution of estimated treatment effects with low variance would be evidence of robustness regarding specification. That is, the estimated treatment effect does not vary far from specification to specification. A distribution of estimated treatment effects with high variance would be evidence of sensitivity regarding specification. That is, the observed estimated treatment effect depends on the particular specification. Multimodal shapes in the distribution can be used to help detect commonalities or differences in the specifications. For example, model specifications that contain a highly predictive covariate in terms of the treatment effect may exhibit similar estimated treatment effects, while model specifications that do not include that covariate may still group together in terms of their estimated treatment effects, but higher or lower than those that included the highly predictive covariate. The commonalities and difference of these two groups of specifications would be visible in the distribution of estimated treatment effects as a bi-modal or multi-modal shape.

2.1.3 Central Tendency Measures

Characterizing the central tendency of estimated treatment effects via the various specifications further assists in determining the level of robustness or sensitivity of the estimated treatment effects as compared to the estimated treatment effect from the chosen specifications. As complimentary measures of center, the mean and median estimated treatment effects can be baseline measures for comparison. The mean estimated treatment effect can stand as an accurate measure of central tendency when the distribution of estimated treatment effects is more symmetric, while the median estimated treatment effect should be considered if the distribution of treatment effects is more skewed since the mean is generally sensitive to outliers.

2.1.4 Standard Errors of the Estimated Treatment Effects

Choices in specification, particularly model specification, can not only result in various estimated treatment effects but also various levels of estimation precision. If a particular model specification results in utilizing a smaller subsample of the data, this can directly impact the standard error of the estimated treatment effect. Accounting for the precision of each estimated effect plays a large role in the extended ensemble estimation framework. A single specification that results in a rejection of any treatment effect may not be reason for concern, but may draw extra attention in the extended ensemble estimation framework once compared to numerous other plausible specifications that resulted in opposite conclusions. A single specification that results in a conclusion counter to that of many alternative specifications could arise by an imprecise estimate of the treatment effect, that is, a larger standard error than those of the alternative specifications. So long as the pool of alternative specifications is rich, the precision of the estimated treatment effects can lead to a deeper, more nuanced conversation regarding the

actual treatment effect, instead of relying on a single specification to make a decision about the treatment at hand.

2.1.5 Discussion

Obtaining the distribution of estimated treatment effects accomplishes the first and second goal of this study. Using visual and quantitative inspections of the shape, center and spread helps researchers and readers understand the robustness or sensitivity of the estimated treatment effect under alternative specifications.

2.2. Estimation Techniques

As stated previously, to more fully understand the sensitivity or robustness of a treatment effect, one may want to consider how estimated treatment effects differ based on estimation technique. That is, a critical piece of Extended Ensemble Estimation is accounting for estimation technique as a specification in the extended ensemble estimation framework by estimating the treatment effect using various estimation techniques. This study will consider two primary estimation techniques; Ordinary Least Squares and Instrumental Variable approach.

2.2.1 Ordinary Least Squares

One of the most common methods of estimation across fields is least squares, particularly ordinary least squares (OLS). This method estimates unknown parameters by minimizing the sum of squared residuals, or differences between observed values of the dependent variable and the predicted value of the dependent variable based on the model specification. Among the many benefits of OLS are a closed form solution for estimates that is quickly and easily produced by most entry level software, having many desirable properties in terms of estimation under certain assumptions, as well as being fairly robust in the case of unsatisfied assumptions. OLS is a consistent estimator in the case of exogenous predictors that form a matrix that has full column

rank. When estimating variance, OLS is consistent when regressors have finite fourth moments. By the Gauss Markov Theorem, OLS is the best linear unbiased estimator in that it achieves the smallest variance among other linear unbiased estimators when the errors are homoscedastic and are serially uncorrelated. OLS is equivalent to the maximum likelihood estimator, another popular estimation technique, when the errors are normally distributed with a mean of zero. In the case of endogenous regressors (regressors that are correlated with the error term), OLS produces biased estimates. When endogeneity is present, other estimation techniques may be more desirable, such as an Instrumental Variables approach.

2.2.2. Instrumental Variables

Instrumental variables is an estimation technique that is often used to estimate causal relationships by addressing potential confounding effects and measurement error. This method is often used when controlled experiments are not feasible, such as observational studies (Angrist & Imbens, 1995). In an observational study, an individual may be more likely to receive treatment than another individual, in turn affecting the resulting outcome. In other words, random assignment does not necessarily hold. The first order condition of Ordinary Least Squares requires the independent variables and the error term to be uncorrelated. If this condition does not hold, Ordinary Least Squares will not provide the causal impact of the independent variable, but instead will produce biased estimates. This first order condition is often known as an exogeneity condition, where the independent variable that satisfies the condition is known an exogenous. In order to handle potential endogeneity, Instrumental Variable estimation hinges on utilizing a variable that is correlated with the endogenous variable, only affecting the outcome indirectly through the endogenous variable, but is not correlated with the error term. A variable that is not correlated with the error term does not suffer the problem of breaking the first order

condition, but also captures the desired effect if it is correlated with the endogenous variable. A variable that satisfies these conditions is known as an instrumental variable and is said to satisfy the exclusion restriction. Instrumental variable estimation requires estimating multiple models in a sequence, known as stages. A common technique using instrumental variables requires two modeling steps, thus is known as two stage least squares. Instrumental variables tends to underperform if variables used as instrumental variables are weak, that is, are poor predictors of the endogenous predictor. Using a weak instrumental variable can result in poor predicted values of the endogenous variable, leading to little variation and a smaller likelihood of predicting the final outcome of interest in the second stage of modeling. Since the endogenous variables and any variable intended to be used as an instrument are all observed, the strength of instruments can be tested directly (Stock et al., 2002). It should be noted that when covariates are exogenous, the desirable small sample properties of Ordinary Least squares can be derived through the moments of the estimator conditional on the covariates. On the other hand, if such properties cannot be easily obtained due to endogenous covariates, inferences using instrumental variables in these scenarios are often based on asymptotic approximations of the sampling distribution of the estimator. A model that is exactly identified produces finite sample estimators with no moments, leading to an estimator that is said to be neither biased or unbiased, where the size of the test statistic may be significantly distorted and could stray far from the value of the parameter of interest (Nelson & Startz, 1988). In terms of precision, instrumental variables tends to produce larger standard errors when compared to ordinary least squares, but remains a consistent estimator in the presence of endogeneity while ordinary least squares is inconsistent. The precision of instrumental variables tends to increase with the strength of the instruments.

2.3 The Extended Ensemble Estimate, Weighting for Precision

Once specifications have been made and estimation techniques have been selected, the next step in extended ensemble estimation is to account for the precision of the estimated treatment effects across specifications and estimation techniques. As stated previously in this chapter, standard errors can rise or shrink for a variety of reasons. The sample size utilized for estimation may shrink due to model specification, leading to larger standard errors in estimation. As pointed out in the earlier section, estimation techniques may also produce various standard errors depending on the relationships between covariates, dependent variables and errors. In this vein, part of extended ensemble estimation is to weight each estimated treatment effect by its precision, thus this will be accomplished through two approaches. The first approach is directly weighing each estimated treatment effect by its associated standard error when creating the distribution of estimated treatment effects. That is, creating a weighted distribution of estimated treatment effects. The second approach is combining the estimated treatment effects into a single effect, weighing each estimate by its associated standard error. This combined estimated treatment effect, weighted by precision, will be called the Extended Ensemble Estimate. Meta-analysis techniques for combining estimated effects across studies are commonly used to estimate effects across experiments or observational studies, accounting for both random and fixed effect models (Hedges and Vevea, 1998). The populations of the studies contained within a meta-analysis need not be constant, as this is one of many strengths of meta-analysis. On the other hand, the models within each study are constant. That is, meta-analysis is not well suited to shed light on the sensitivity or robustness of the model specification within a single study. Extended ensemble estimation is intended to be a within-study tool where the population and sample at hand are constant, while the robustness or sensitivity of the estimates produced by

various model specifications are the focus. In order to gain the capacity to consider all possible model specification within a study, Extended Ensemble Estimation adopts a similar technique, but to combine estimated effects across specifications, within a single experiment or observational study. In this sense, Extended Ensemble Estimation has the capacity to consider all possible specifications, while typical meta-analysis utilize what is already generated, potentially missing important models or alternative specifications.

In the extended ensemble estimation framework, let $i = 1, \dots, k$ denote the k various specifications and y_i denote the observed value of the treatment effect in the i^{th} specification. The meta-analytic approach is a special case of the general linear mixed effect model with heteroscedastic sampling variances, assumed to be known. This type of model can be fitted by a two step approach outlined in Raudenbush (2009). Let θ_i denote the unknown true treatment effect, such that

$$y\_i \,|\theta\_i \sim N(\theta\_i, v\_i)$$

In the random-effects model, we assume that $\theta_i \sim N(\mu, \tau^2)$, namely that the true treatment effects are normally distributed with average treatment effect μ and variance τ^2. This model can be expressed as

$$y_i = \mu + u_i + \varepsilon_i$$

Where $u_i \sim N(0, \tau^2)$ and $\varepsilon_i \sim N(0, v_i)$. With this setup, the Extended Ensemble Estimate is denoted by

$$\hat{\mu}_{\mathrm{EEE}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{w_i y_i}}{\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{w_i}}$$

Where w_i denotes the weighting of each estimated treatment effect from specification i, specifically,

$$w_i = \frac{1}{\hat{\tau}^2 + v_i}$$

Where $\hat{\tau}^2$ denotes an estimate of τ^2, the variance in the true effect across specifications, and v_i denotes the sampling variance for specification i. A special case is the equal-effects model, specifically when $\tau^2 = 0$. In this case, the true treatment effects across specifications are homogenous and can be written as $\theta_1 = \theta_2 = \cdots = \theta_k = \theta$. The model of this special case can be written

$$y_i = \theta + \varepsilon_i$$

Where θ denotes the true treatment effect. In this case, the Extended Ensemble Estimate is denoted by

$$\hat{\theta}_{EEE} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}$$

Where $w_i = \frac{1}{v_i}$. In the both models, v_i is assumed to be known and is the square of the standard errors of the estimated treatment effects. As such, this method of weighing is also known as the inverse-variance method, or variance known, in meta-analysis literature. For reference, the unweighted least squares estimate of the treatment effect (Laird and Mostelle, 1990) can be expressed as

$$\bar{\theta} = \frac{\sum_{i=1}^{k} y_i}{k}$$

The first step in deriving the Extended Ensemble Estimate is to estimate τ^2 using one of many estimators, including the Hunter-Schmidt estimator (Hunter and Schmidt, 2004), the Hedges estimator (Hedges and Olkin, 1985; Raudenbush, 2009), the DerSimonian-Lair estimator (DerSimonian and Laird 1986' Raudenbush 2009), the Sidik-Jonkman estimator (Sidik and Johnkman, 2005a,b), the maximum likelihood or restricted maximum likelihood estimator

(Viechtbauer 2005; Ruadenbush 2009), or the empirical Bayes estimator (Morris, 1983; Berkey et al., 1995). The second step is to use weighted least squares to estimate the weights w_i. Once the weights w_i are known, $\hat{\theta}_{EEE}$ can be calculated directly in order to achieve the third goal.

CHAPTER 3

UPDATING REGRESSION COEFFICIENTS USING EXTENDED ENSEMBLE ESTIMATION

3.1 General Framework for Updating

Once the Extended Ensemble Estimate (β_{EEE}) has been attained, it can be used to update the original estimated treatment effect from the original specification, $\beta_{original}$. One way to achieve an updated treatment effect $\beta_{updated}$, is to form the weighted average of $\beta_{original}$ and β_{EEE} as follows

$$\beta_{updated} = \left[\pi\beta_{original} + (1-\pi)\beta_{EEE}\right]$$

Where π is used to weight each estimated treatment effect. This can be thought of as updating the original estimated treatment effect by the extended ensemble estimate. The choice of π determines how much to weight the original estimated treatment effect as opposed to the extended ensemble estimate. Possible choices of π and potential consequences will be discussed in section 3.3. In the sense of using empirical data to update an estimate, this has similarities to empirical Bayesian methods that will be discussed in the following section.

3.2 Ties to Empirical Bayes Methodology

Bayesian statistical inferences refer to the techniques of modeling a parameter of interest, say θ, with a distribution of potential values instead of assuming it is fixed, as in a frequentist approach. The distribution of the parameter of interest, θ, allows for the ability to account for any prior beliefs regarding θ, thus is often referred to as the prior distribution (Jackman, 2009; Lynch, 2007). Observed data is used to update the prior distribution of θ by scaling the prior distribution by the likelihood of the observed data, producing a new distribution referred to as the posterior distribution of θ. The posterior distribution of θ is, by definition, conditional on the

observed data while the prior distribution of θ is fixed before any data are observed. Once the posterior distribution is known, the unknown parameter θ is estimated using a single measure of the posterior distribution, often the mean or median, known as the bayes estimate. Empirical Bayes methods are a subset of methods within this general framework that estimate the prior distribution of θ using observed data (Casella, 1985; Lynch, 2007; Robbins, 1992).

Extended Ensemble Estimation takes a meta-analysis approach of combining estimated effects in order to produce a single estimate of the treatment effect by weighting each estimated treatment effect by its precision, called the Extended Ensemble Estimate. In meta-analysis, Bayesian methods have a few distinct advantages in this application. The Extended Ensemble Estimate depends on the variance of the true treatment effect across specifications, τ^2. The Bayesian framework allows for the ability to directly model any uncertainty in the estimation of τ^2. Bayesian methods produce full posterior distributions for both μ, the average treatment effect, and τ^2 (Chung et al., 2013; McNeish, 2016). Thus, in general, Bayesian methods allow us to account for any prior knowledge or assumptions we want to incorporate. There are many existing estimators for τ^2, previously discussed in Chapter 2 within the general methodology, including the empirical bayes estimator (Morris, 1983; Berkey et al., 1995). The derivation of this estimator in Berkey et al. (1995) assumes that $y_i \,|a, D \sim N(X_i\, a, D + s_i^2)$, where y_i is the observed treatment effect in specification i, X_i is a row vector that contains values of the covariates of study i, "a" is a column vector of regression coefficients, "D" is the between specification variance (τ^2 in the previous notation) and s_i^2 is the estimated error variance. The estimate of "a" is given by

$$\hat{a} = (X^T\, VX)^{-1}\, X^T\, VY$$

Where $V = diag(W_1, \ldots, W_k)$, the diagonal matrix of weights

$$W_i = \frac{1}{(\widehat{D} - s_i^2)}$$

and $\widehat{D}$ is an approximately unbiased estimator of D.

3.3 Choosing a Weighting Scheme for Updating Regression Coefficients

In the context of Extended Ensemble Estimation, the goal in this chapter is to provide an updated estimated treatment effect using the original estimated treatment effect and the extended ensemble estimate. Frank and Min (2007) adapted a Bayesian methodology for updating indices of robustness in the context of observed and unobserved samples in order to form an ideal estimate. The authors defined the likelihood in terms of observables and the prior in terms of the sample from the potentially unobservables. In this sense, they are able to achieve a posterior estimate from updating the prior via the likelihood. Since the significance test for correlations and partial correlations is equivalent to regression coefficients (Cohen and Cohen 1983; Fisher 1924), the authors worked in terms of correlations and partial correlations. Using the Fisher z transformation (Lee, 1989), sample correlations are normally distributed with variance $\frac{1}{n}$ and are an unbiased estimate of the Fisher z transformation of the corresponding population correlation. Denoting $z(r)$ as the Fisher z transformation of a sample correlation r, the estimated posterior mean can be expressed as

$$z\left(r_{xy}^{ideal}\right) = Var\left(r_{xy}^{ideal}\right)\left[\frac{z\left(r_{xy}^{ob}\right)}{Var\left(r_{xy}^{ob}\right)} + \frac{z\left(r_{xy}^{un}\right)}{Var\left(r_{xy}^{un}\right)}\right]$$

Where r_{xy}^{ob} is the statistically significant sample correlation for the observed cases, r_{xy}^{un} is the sample correlation coefficient for the unobserved cases and r_{xy}^{ideal} is the correlation coefficient for the ideal based on a combination of observed and unobserved cases. Since $Var\left(r_{xy}^{obs}\right) = \frac{1}{n^{obs}}$, $Var\left(r_{xy}^{unob}\right) = \frac{1}{n^{unob}}$, and $Var\left(r_{xy}^{ideal}\right) = \frac{1}{n^{obs}+n^{unob}}$,

$$z\left(r_{xy}^{ideal}\right) = \frac{1}{n^{obs} + n^{unob}}\left[n^{obs} z\left(r_{xy}^{obs}\right) + n^{unob} z\left(r_{xy}^{unob}\right)\right]$$

Letting $\pi = \frac{n^{obs}}{n^{obs}+n^{unob}}$, the authors provide the final estimated posterior mean as

$$z\left(r_{xy}^{ideal}\right) = \left[\pi z\left(r_{xy}^{obs}\right) + (1-\pi) z\left(r_{xy}^{unob}\right)\right]$$

Through the lens of Empirical Bayesian methodology, the posterior distribution for ρ_{xy} is $N\left(z\left(r_{xy}^{ideal}\right), \frac{1}{n^{obs}+n^{unob}}\right)$. Thus, by using the mean of the posterior, the Empirical Bayes estimate for ρ_{xy} is $z\left(r_{xy}^{ideal}\right)$. The variance can be used to quantify robustness by considering what values of r^{un} would be necessary for r_{xy}^{ideal} fall within a 95 percent highest posterior density (HPD) interval (Frank & Min, 2007). In the case of the extended ensemble estimate, one can use the standard errors of the estimated treatment effects to define π in a similar way to update the original estimated treatment effect with the extended ensemble estimate. That is,

$$\pi = \frac{(SE_{EEE})^2}{(SE_{EEE})^2 + \left(SE_{Original}\right)^2}$$

could be used to weight the original estimated treatment effect and the extended ensemble estimate, where SE_{EEE} and $SE_{Original}$ are the standard errors of the extended ensemble estimate and the original estimated treatment effect, respectively. Using the standard errors in this weighting scheme accounts for estimation efficiency, which is directly related to sample size.

3.3.1 The Effects of Weighting Scheme

The philosophical choice to represent the unknown parameter of interest, θ, with a distribution rather than by a fixed value is a key difference between Bayesian and frequentist methods. This captures our typical view that progress in science generally is derived from learning from past findings, incorporating information from these findings and realizing that no study is conducted in the absence of previous research. Bayesian inferences require that the prior

knowledge of θ be stated explicitly via the prior distribution, a non-conditional distribution representing our prior knowledge regarding θ (Kaplan, 2014). Since the posterior distribution of θ is derived from the prior distribution, the posterior hinges directly on the choice of a prior. In some scenarios, we may not have strong prior knowledge of θ. When prior knowledge is completely lacking, one would select a prior that models this directly by selecting a prior distribution of possible values of θ that are no more or less likely than each other. A uniform distribution is a common choice in this case. This case is an extreme example of priors known as non-informative priors. In other cases where we have prior information that we wish to incorporate, we can select a prior distribution of θ such that we believe some potential values are more or less likely than others. As stated previously, the inferential statistics based on the posterior distribution may change depending on the choice of prior, thus one must be careful and deliberate when deciding on whether to select an informative or non-informative prior (Kaplan, 2014).

In the case of Extended Ensemble Estimation, the prior knowledge can be thought of as the original estimated treatment effect and the associated variability. In order to achieve an updated estimated treatment effect, one can use the estimated treatment effect that is weighted for precision from the various alternative specifications, namely, the extended ensemble estimate and its associated variability. Choosing values for π that utilize the associated standard errors of the estimated treatment effects would allow researchers to weight the original estimated treatment effect and extended ensemble estimate based on the level of uncertainty of each estimated treatment effect. As significance statements rely directly on standard error of estimates, standard errors are often focal points of discussion (Deaton & Cartwright, 2018) and

may serve as an initial choice of π. Formulated in terms of efficiency, one example of a possible weighting scheme could be formulated as given on page 28:

$$\pi = \frac{(SE_{EEE})^2}{(SE_{EEE})^2 + \left(SE_{Original}\right)^2}$$

It is worth nothing that other values of π could be selected in order to weight the original estimated treatment effect and the extended ensemble estimate. That is also to say that the resulting updated estimated treatment effect, $\beta_{updated}$, hinges on the choice of π. Taking a closer look at the formulation of $\beta_{updated}$ below, one can map out the consequences of various choices of π.

$$\beta_{updated} = \left[\pi\beta_{original} + (1-\pi)\beta_{EEE}\right]$$

Choosing $\pi = 1$ would result in the original estimated treatment effect, $\beta_{original}$, while choosing $\pi = 0$ would result in the extended ensemble estimate, β_{EEE}. Weighting both estimated treatment effects equally would be achieved by choosing $\pi = 0.5$.

As this is a choice that impacts the updated treatment effect, the choice of π should be made carefully and intentionally. For example, using the standard errors of both estimated treatment effects for a weighting scheme that reflects the efficiency of each estimated treatment effect.

CHAPTER 4

USER GUIDE FOR BEST PRACTICES WHEN IMPLEMENTING EXTENDED ENSEMBLE ESTIMATION

4.1 Introduction to Best Practices

This chapter is intended to guide end users in using extended ensemble estimation in a way to achieve the best performance possible, avoiding traditional statistical pitfalls, and how to best promote the conversation of potential causality of a treatment effect. The process of Extended Ensemble Estimation, including estimated treatment effect weighted by precision, is suited to serve the conversation around the potential causal relationship between a treatment variable and outcome by quantifying the robustness and sensitivity of the estimated treatment effect across alternative specifications.

In order to maximize the effectiveness of Extended Ensemble Estimation, the end user should strive for a pool of alternative specifications that are strongly supported by existing theory, supporting empirical evidence and past research. These specifications increase the quality of the pool of specifications, while poor specifications that are not vetted can hinder the performance of ensemble estimation. Once alternative specifications have resulted in estimated treatment effects, they can be used to find the extended ensemble estimate, weighted for precision. The distribution, in tandem with the estimate weighted for precision, can be used in comparison to the original proposed specification in order to quantify any sensitivity or robustness, and to inform the conversation around a potential causal treatment effect. As the pool of alternative specifications grows, the distribution of estimated treatment effects will tend to be more smooth than discrete, assisting in the ability to decipher shape, center and spread. As further sections will discuss, ensemble estimation techniques can suffer in the presence of poor pools of

specifications, so it is generally best to grow the pool of alternative specifications while holding the quality of specifications as high as possible.

Like many quantitative methodologies, the effectiveness of extended ensemble estimation can be impaired by outside influences. The following sections will discuss how the end user can mitigate or even eliminate these potential weaknesses.

4.2 Alternative Specifications

Chapter 1 discusses weaknesses of ensemble methods in general, specifically how ensemble techniques in general are traditionally susceptible to poor performance in the presence of poor alternative specifications. Extended Ensemble techniques utilize alternative specifications, thus poor alternative specifications may hinder ultimate performance (Saez-Rodriguez et al., 2016). A poor alternative specification may show up in the form of large standard errors, perhaps due to a shrinkage in the utilized sample imparted by the model specification. In such a case, this would be reflected in the weighted distribution of estimated treatment effects. The extended ensemble estimate, weighted for precision, would also reflect this by weighting estimated treatment effects with large standard errors less than those with smaller standard errors. In the case of instrumental variables, the strength of instruments used within a specification may provide a more appropriate weight than standard errors alone since particular estimation techniques may suffer from larger standard errors. For instrumental variables, the strength of an instrument provides the user with a sense of confidence regarding the associated standard error. At this point, the user may decide whether to weight estimated treatment effects by standard errors, strength of instruments used, or a combination of both if applicable and appropriate.

In order to keep the quality of the pool of alternative specifications as high as possible, alternative specifications should be derived, at least in part, by past research, evidence or

empirical evidence. In that sense, performance of extended ensemble estimation could be hindered by an inflated pool of alternative specifications and it is in the end user's best interest to keep the quality of the pool of alterative specifications as high as possible. A potential weakness of extended ensemble estimation is weak pool of alternative specifications. For example, specifications that omit an important covariate may under or overestimate the treatment effect, thus hindering the ability of the user to assess the sensitivity or robustness of the estimated treatment effect across specification.

In the case of a rich pool of alternative specifications, a strength of Extended Ensemble Estimation is that it leaves no room for spurious results to hide. In that sense, it is the goal of the user to provide a pool of high quality alternative specifications.

4.3 Estimation Techniques

Discussed in Chapter 2, estimation techniques play a crucial role in extended ensemble estimation as they produce the estimated treatment effects. The observed estimated treatment effects, given a set of data, may differ slightly or largely in part due to the choice to use OLS or Instrumental Variables. While OLS contains many desirable properties as an estimator, the Instrumental Variables estimation approach strengths can compliment potential weaknesses of OLS. Although Instrumental Variable estimation tends to produce larger standard errors when compared to OLS, it remains a consistent estimator in the presence of endogeneity, a phenomena that results in inconstancy in the OLS estimator. The Extended Ensemble Estimate, weighted for precision, may naturally weigh estimates produced by OLS more favorably as compared to Instrumental Variables due to the tendency of smaller standard errors, all else being equal. In order to display any sensitivity regarding estimation technique, the end user may follow the

extended procedure twice; achieving separate estimated treatment effects weighted by precision for each estimation technique.

4.4 P-hacking and Cherry Picking

Extended Ensemble Estimation is a statistical tool, in that its effectiveness can be minimized or maximized by the end user. This methodology, by construction, has the ability to help researchers decipher between spurious estimated treatment effects and potentially causal relationships through the robustness of the estimated treatment effects. So long as the pool of alternative specifications is rich, it also has the ability to display outlying estimated treatment effects, in the case of end users acting in good faith, and the ability to detect potentially cherry picked results, in the case of end users attempting to support particular positions.

Cherry picking refers to the act of selecting individual cases or data that confirm a particular result while ignoring cases that contradict that result, intentionally or unintentionally (Klass, 2008). If the original estimated treatment effect had been cherry picked, that is, a result that significantly differs from the majority of alternative plausible models, distribution of estimates produced by the extended ensemble estimation process would be highly variable, suggesting a level of sensitivity of the original result. This assumes that the user presents other specifications and resulting estimates from which they selected from. On the other hand, if the original result fell within a reasonable range of the alternative estimates, this would suggest a level of robustness of the original result and would also be supporting evidence against the notion of cherry picking.

P-hacking, another common statistical pitfall, is when a researcher attempts several statistical analyses or model specifications, then selectively reports those that produce significant results (Brodeur et al., 2016; Simmons et al., 2016; Gadbury & Allison, 2012; L.K. et al., 2012). While

end users are able to provide specifications of their choosing, extended ensemble estimation requires a pool of specifications for comparison. It is the responsibility of the user to provide plausible alternative specifications in order to help quantify robustness or sensitivity of an estimated treatment effect. During the peer review process, reviewers may propose alternative specifications in an attempt to identify potential p-hacking. In this case, extended ensemble estimation would provide the framework to compare the estimated treatment effects of the authors specifications to the, potentially many, estimated treatment effects of the reviewers' alternative specifications. Extended ensemble estimation does not allow the end user to select single specifications that result in desirable results by requiring alternative model specifications, ultimately providing a level of transparency to the specification phase and the impact it has on the resulting estimated treatment effect.

CHAPTER 5

SIMULATIONS

5.1 How To Use Simulation to Inform Extended Ensemble Estimation

It is not feasible to account for every possible situation one may encounter during the research process, specifically regarding the estimation phase. To build on the guiding principles for the user to keep in mind in order to best implement the extended ensemble estimation technique in the previous chapter, this chapter is meant to serve as a compliment by exploring the performance of extended ensemble estimation across various common, possible scenarios using simulation. I will start by laying out the different scenarios to be explored, the details of the simulation that will be used and finally, I will discuss the performance of extended ensemble estimation in each scenario.

5.2 Pre-treatment and Confounding Variables

When considering the estimated effect of a treatment variable, much time and effort is often spent on trying to account for potential confounding variables that may cloud a researcher's ability to make confident, clear conclusions. Even if one does everything in their power to rule out alternative explanations for estimated effects of a treatment variable, it is extremely difficult to feasibly rule out all alternative explanations. In other words, accounting for every possible confounding variable is not only a massive hurdle, but often impossible. One method to help overcome unknown variables that may confound estimates of a treatment variable is to include a pre-treatment variable that would be present during which a potential confounder would also be present, thus negating the need to include the confounding variable. This variable may not be randomly assigned but should be strongly correlated with the outcome variable. An example of such a variable could be an academic pre-test in a school setting – present during which a

potential confounder is also likely to be present. Although a pre-treatment variable of this nature is not a universal cure for all potential confounders, it can be easier to implement as opposed to thinking of and measuring potential confounders. To raise concerns of a particular confounder regarding the treatment, would require evidence that such a confounder was not present and was not captured in the pre-treatment measurement. When gauging how ensemble estimation works in the presence of a strong pre-treatment variable that may not be randomly assigned, it is sufficient to simulate a variable that is strongly correlated with the treatment variable of interest, as well as the outcome variable. A weak pre-treatment variable that is randomly assigned could be simulated using a variable that is weakly correlated with the treatment variable and the outcome variable.

5.3 Instrumental Variables

Another method that aims to address potential confounding effects and measurement error, popularized mainly in Econometrics, is called Instrumental Variables. This method is often used when controlled experiments are not feasible, such as observational studies (Angrist & Imbens, 1995). In an observational study, an individual may be more likely to receive treatment than another individual, in turn affecting the resulting outcome. In other words, random assignment does not necessarily hold. In the derivation of Ordinary Least Squares, the first order condition requires the independent variable and the error term to be uncorrelated. If this condition does not hold, Ordinary Least Squares will not provide the causal impact of the independent variable, but only the parameter that makes the resulting errors seem uncorrelated with the independent variable. This situation results in biased and inconsistent estimates using Ordinary Least Squares. (Bullock et. Al, 2010). If the correlation between the independent variable and error term is not zero, the independent variable is known as endogenous. The first

order condition that requires this correlation to zero is often known as an exogeneity condition, where the independent variable that satisfies the condition is known an exogenous. The instrumental variable method handles endogeneity by utilizing a variable that is correlated with the endogenous variable, only affects the outcome indirectly through the endogenous variable, but is not correlated with the error term. A variable that is not correlated with the error term does not suffer the problem of breaking the first order condition, but also captures the desired effect if it is correlated with the endogenous variable. This variable is called an instrumental variable and the application requires multiple steps known as stages. A common technique using instrumental variables requires two modeling steps, thus is known as two stage least squares. When gauging how ensemble estimation works in the presence of a strong instrumental variable, it is sufficient to simulate a variable that is strongly correlated with the treatment variable of interest but weakly correlated with the outcome, that is to not affect the outcome directly. In order to consider a weak instrument, it is sufficient to simulate a variable that is weakly correlated to both the treatment variable and the outcome variable.

5.4 Randomized Control Trials

A list of commonly encountered designs would not be complete without accounting for randomized control trials. In an attempt to reliably estimate unbiased treatment effects, randomized control trials utilize random assignment between treatment and control groups. When performed with fidelity, this framework allows researchers to attribute any observed difference between the treatment group and control group to the treatment effect by minimizing any other possible contamination of the treatment effect.

Randomized control trials have a long history in medical research where biased estimates may have long term consequences of high magnitude. More recently, randomized control trials

have spread to other disciplines, such as economics and social sciences. This strict structure of randomized control trials assists researchers in making a case for causality by being able to rule out any confounding effects via the control group. Imbens (2010) summarized common conceptions surrounding randomized control trials by saying, "Randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top." With that said, randomized control trials are not without drawbacks and faults. Other than the difficulty that comes with properly carrying out a strict framework and carefully constructed design, they often incur high monetary and time expenses. Randomized control trials can also suffer from the lack of generalizability. A particular trial may only consider a sample from a specific high-risk group to maximize the probability of detecting an effect, which may not be applicable to a low-risk group or the population as a whole. Randomized control trials may not be practical for urgent health issues where decisions must be made faster than a well-performed trial can permit. Although it is not uncommon for trials to last many years, that still may not be enough to assess long-term treatment effects. As randomized control trials are often viewed as more credible and rigorous than other methods, other designs often attempt to mimic randomized control trials in order to gain their benefits (Angrist & Pischke, 2010). In this spirit, Extended Ensemble Estimation can be used to help gauge how well a randomized control trial was constructed and conducted by comparing estimated effects of the treatment to estimated effects using various specifications across various model specifications. Low dispersion of estimated effects across model specifications and alignment would suggest a more sound randomized control trial implementation while a high dispersion of estimated effects across model specifications or misalignment would suggest a less sound randomized control trial implementation. That is, if a randomized control trial is well constructed and implemented, the estimated treatment effect

should align with the estimated treatment effects across the model specifications, while the estimate effects across model specifications should not vary. In order to gauge how ensemble estimation works within a randomized framework, it is sufficient to simulate a variable that is weakly correlated with the treatment variable, ideally not correlated with the treatment variable at all in the case of perfect randomization, but strongly correlated with the outcome variable.

5.5 Accounting for Selection Bias

The simulations that follow include an outcome variable (Y), treatment variable (X_1), as well as two more variables (X_2 and X_3) that will be used to account for the various scenarios described above. Since each model must include the outcome and treatment variable, there are four possible models resulting from covariate selection in each simulation (shown below).

$$Y = \beta_0 + X_1\beta_1$$

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2$$

$$Y = \beta_0 + X_1\beta_1 + X_3\beta_3$$

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3$$

Based on the relationship established by Heckman (1979) between bias due to nonrandom assignment to treatment conditions and bias due to sample selection, bias due to omitted variable can be thought of as bias due to sample selection. Furthermore, variability due to sample selection could be thought of as variability due to model specification through included or omitted covariates. In this sense, using such a limited number of controls after the treatment variable can be used to address many common concerns, including but not limited to omitted variable bias, sample selection bias and variability due to sample selection. In this sense, X_2 and it's correlations with X_1 and Y are used to help specify the various scenarios, while X_3

and it's correlations with X_1, X_2 and Y are used to stand for other potential covariates that may have been missed or left out of analysis. Model 1 is the base model, using only the treatment variable X_1 while ignoring all other controls. Model 2 includes X_2 in order to control for potential instruments, pre-treatments or random assignments. Model 3 includes X_3, standing for potentially left out control variables. Model 4 includes all potential variables. For each scenario below, a correlation matrix using standardized variables is defined. For each scenario below, a correlation matrix using standardized variables is defined. Each simulated scenario utilized the ordinary least squares estimation technique, where the mean, median and extended ensemble estimate are reported. Specifically, the Cholesky decomposition can be used to generate data under particular specifications and then a correlation matrix is calculated from which OLS estimates are obtained (Becker, 1992; Becker, 1995; Becker & Aloe, 2019; Sumiati et al., 2020)

5.5.1 Strong Pre-Treatment

Simulating a variable that is strongly correlated with the treatment and outcome, $r_{X_1,X_2} = r_{X_2,Y} = 0.8$, representing a strong pre-test, plays a large role in the resulting estimates. When the strong pre-test is omitted, the base effect of the treatment variable, $r_{X_1,Y} = 0.7$, is estimated as $\hat{\beta}_1 = 0.7$. When the strong pre-test included, the estimated effect of the treatment variable is $\hat{\beta}_1 = 0.17$. The mean and median estimated effects of the treatment are 0.28 and 0.43, respectively. The ensemble estimate of the treatment effect is 0.28, with a standard error of 0.2814. In the case of a strong pre-test, the change in estimates is representative of the effectiveness of the included pre-test. This change is also being displayed by the mean, median and ensemble estimates. Below are the tables including the correlations, model specifications, estimated effects of the treatment, estimated standard error of the treatment, mean estimated effect, median estimated effect, and extended ensemble estimate.

Table 5.1.1 Correlation structure for strong pre-treatment

Corr(. , .)	y	x1	x2	x3
y	1	0.7	0.8	0.2
x1	0.7	1	0.8	0.1
x2	0.8	0.8	1	-0.3
x3	0.2	0.1	-0.3	1

Table 5.1.2 Model specifications

Formula	X1 Estimate	X1 Standard Error	X1 Est/SE
y~ 1+x1+x2+x3	-0.44811	0.065671	-6.82359
y ~ 1+x1+x3	0.686869	0.070563	9.734172
y ~ 1+x1+x2	0.166667	0.098601	1.690309
y ~ 1+x1	0.7	0.071414	9.801961

Table 5.1.3 Extended Ensemble Estimation results

	X1 Estimate
Mean	0.276356
Median	0.426768
EEE	0.275402 (0.2814)

5.5.2 Weak Pre-Treatment

Simulating a variable that is weakly correlated with the treatment and outcome, $r_{X_1,X_2} = r_{X_2,Y} = 0.1$, representing a weak pre-test, plays a small role in the resulting estimates. When the weak pre-test is omitted, the base effect of the treatment variable, $r_{X_1,Y} = 0.7$, is estimated as $\hat{\beta}_1 = 0.7$. When the weak pre-test included, the estimated effect of the treatment variable is $\hat{\beta}_1 = 0.696$. The mean and median estimated effects of the treatment are both 0.69. The ensemble estimate of the treatment effect is 0.69, with a standard error of 0.1585774. In the case of a weak pre-test, the lack of change in estimates is due to the lack of effectiveness of the included pre-test. That is, the mean, median and ensemble estimates are robust in the presence of a weak pre-test. Below are the tables including the correlations, model specifications, estimated effects of the treatment, estimated standard error of the treatment, mean estimated effect, median estimated effect, and extended ensemble estimate.

Table 5.2.1 Correlation structure for strong weak-treatment

Corr(. , .)	y	x1	x2	x3
y	1	0.7	0.1	0.2
x1	0.7	1	0.1	0.1
x2	0.1	0.1	1	-0.3
x3	0.2	0.1	-0.3	1

Table 5.2.2 Model specifications

Model	X1 Estimate	X1 Standard Error	X1 Est/SE
y~ 1+x1+x2+x3	0.676471	0.070828	9.550863

Table 5.2.2 (cont'd)

y ~ 1+x1+x3	0.686869	0.070563	9.734172
y ~ 1+x1+x2	0.69697	0.07171	9.719274
y ~ 1+x1	0.7	0.071414	9.801961

Table 5.2.3 Extended Ensemble Estimation results

	X1 Estimate
Mean	0.690077
Median	0.691919
EEE	0.690033 (0.1585774)

5.5.3 Strong Instrumental Variable

Simulating a variable that is strongly correlated with the treatment but weakly correlated with the outcome, $r_{X_1,X_2} = 0.8$ and $r_{X_2,Y} = 0.2$, representing a strong instrument, plays a large role in the resulting estimates. When the strong instrument is omitted, the base effect of the treatment variable, $r_{X_1,Y} = 0.7$, is estimated as $\hat{\beta}_1 = 0.7$. When the strong instrument is included, the estimated effect of the treatment variable is $\hat{\beta}_1 = 1.5$. The mean and median estimated effects of the treatment are 1.2 and 1.1, respectively. The ensemble estimate of the treatment effect is 1.2 with a standard error of 0.3072394. In the case of a strong instrument, the change in estimates is representative of strong instrument when using OLS. That is, the mean, median and ensemble estimates are robust in the presence of a strong instrument. Below are the tables including the correlations, model specifications, estimated effects of the treatment, estimated

standard error of the treatment, mean estimated effect, median estimated effect and extended ensemble estimate.

Table 5.3.1 Correlation structure for strong instrument

Corr(. , .)	y	x1	x2	x3
y	1	0.7	0.2	0.2
x1	0.7	1	0.8	0.1
x2	0.2	0.8	1	-0.3
x3	0.2	0.1	-0.3	1

Table 5.3.2 Model specifications

Model	X1 Estimate	X1 Standard Error	X1 Est/SE
y~ 1+x1+x2+x3	1.900943	0.043394	43.80694
y ~ 1+x1+x3	0.686869	0.070563	9.734172
y ~ 1+x1+x2	1.5	0.06455	23.2379
y ~ 1+x1	0.7	0.071414	9.801961

Table 5.3.3 Extended Ensemble Estimation results

	X1 Estimate
Mean	1.196953
Median	1.1
EEE	1.211574 (0.3072394)

5.5.4 Weak Instrumental Variable

Simulating a variable that is weakly correlated with the treatment and outcome, $r_{X_1,X_2} = r_{X_2,Y} = 0.1$, representing a weak instrument, plays a small role in the resulting estimates. When the weak instrument is omitted, the base effect of the treatment variable, $r_{X_1,Y} = 0.7$, is estimated as $\hat{\beta}_1 = 0.7$. When the weak instrument is included, the estimated effect of the treatment variable is $\hat{\beta}_1 = 0.69$. The mean and median estimated effects of the treatment are both 0.69. The ensemble estimate of the treatment effect is 0.69 with a standard error of 0.1585774. The lack of change in estimates is due to the lack of effectiveness of the included weak instrument. The lack of change in the mean, median and ensemble estimates is evidence of robust estimation in the presence of a weak instrument. Below are the tables including the correlations, model specifications, estimated effects of the treatment, estimated standard error of the treatment, mean estimated effect, median estimated effect and extended ensemble estimate.

Table 5.4.1 Correlation structure for weak instrument

Corr(. , .)	y	x1	x2	x3
y	1	0.7	0.1	0.2
x1	0.7	1	0.1	0.1
x2	0.1	0.1	1	-0.3
x3	0.2	0.1	-0.3	1

Table 5.4.2 Model specifications

Model	X1 Estimate	X1 Standard Error	X1 Est/SE
y~ 1+x1+x2+x3	0.676471	0.070828	9.550863

Table 5.4.2 (cont'd)

y ~ 1+x1+x3	0.686869	0.070563	9.734172
y ~ 1+x1+x2	0.69697	0.07171	9.719274
y ~ 1+x1	0.7	0.071414	9.801961

Table 5.4.3 Extended Ensemble Estimation results

	X1 Estimate
Mean	0.690077
Median	0.691919
EEE	0.690033 (0.1585774)

5.5.5 Randomized Control Trial

Simulating a variable X_2 that is weakly correlated with the treatment and strongly correlated with the outcome, $r_{X_1,X_2} = 0.2$ and $r_{X_2,Y} = 0.8$, representing a randomized control trial with an ancova design, plays a moderate role in the resulting estimates. When the grouping variable is omitted, the base effect of the treatment variable, $r_{X_1,Y} = 0.7$, is estimated as $\hat{\beta}_1 = 0.7$. When X_2 is included, the estimated effect of the treatment variable is $\hat{\beta}_1 = 0.56$. The mean and median estimated effects of the treatment are 0.62 and 0.63, respectively. The ensemble estimate of the treatment effect is 0.58 with a standard error of 0.102087. The change in estimates is due to the importance of randomization. The similar results in the mean, median and ensemble estimates coincide and confirm with the estimate when randomization is present. Note that the lower estimated treatment effect by the extended ensemble estimate is due to the

weighting of the standard errors. Estimated treatment effects 0.53 and 0.56 whose models include X_2 (accounting for the RCT design) are more precise in terms of standard errors (0.0151 and 0.0242, respectively) than the estimated treatment effects of 0.69 and 0.7 whose models exclude X_2 (0.0717 and 0.0714, respectively). The Extended Ensemble Estimate, 0.57, is a result of favoring the more precise estimates more since the Extended Ensemble Estimate is weighted for precision. If interpreted in terms of information, the Extended Ensemble Estimate will favor estimates that provide more precise information.

Table 5.5.1 Correlation structure for Randomized Control Trial

Corr(. , .)	y	x1	x2	x3
y	1	0.7	0.8	0.1
x1	0.7	1	0.2	0.1
x2	0.8	0.2	1	-0.2
x3	0.1	0.1	-0.2	1

Table 5.5.2 Model specifications

Model	X1 Estimate	X1 Standard Error	X1 Est/SE
y~ 1+x1+x2+x3	0.534368	0.015052	35.50265
y ~ 1+x1+x3	0.69697	0.07171	9.719274
y ~ 1+x1+x2	0.5625	0.024206	23.2379
y ~ 1+x1	0.7	0.071414	9.801961

Table 5.5.3 Extended Ensemble Estimation results

	X1 Estimate
Mean	0.623459
Median	0.629735
EEE	0.576734 (0.102087)

5.5.6 Discussion of Simulation Results

One feature of extended ensemble estimation is the ability to utilize the precision of each estimate through standard error. When a strong pre-test is present, the estimated treatment effect decreases from 0.7 to 0.17 with consistent precision. If the pre-test is strong, researchers may want to carefully consider which effect to base conclusions on. The ensemble estimate gives researchers a framework to weigh the two estimates; is the strong pre-test evidence of a lack of actual treatment effect, or is there a treatment effect worth reporting? The extended ensemble estimate of 0.27 represents an edge towards a lack of treatment effect in the presence of a strong pre-test. In the case of a weak instrument, researchers must be careful to avoid biased and inconsistent estimates. In this scenario, the extended ensemble estimate of 0.69 is evidence favoring the base treatment effect through the weighing of precision. A treatment effect estimated by ordinary least squares in the presence of a weak instrument may suffer from a lack of precision, thus would receive less weight in the ensemble estimate. In the case of randomization, the change of estimated effect from 0.7 to 0.56 is evidence of the importance of random assignment. The extended ensemble estimate of 0.58 is representative of favoring the more precise estimate of the treatment effect with random assignment present.

CHAPTER 6

CASE STUDY – EFFECTS OF KINDERGARTEN ON CHILDRENS COGNITIVE GROWTH IN READING AND MATHEMATICS

6.1 Introduction

Retention policy in school continues to be controversial and unresolved for nearly a century (Dong 2010, Goos, Pipa & Peixoto 2021, Holmes 1989, Jackson 1975, Jimerson 2001, Park, Steiner Robertson 2021, Shepard 1989). Empirical evidence spans from supporting negative effects of retention on academic achievement and both person and social development, to no statistically detectable effect, as well as a smaller portion of studies showing supporting evidence for retention. More recently, emphasis has been put on educational standards and accountability within schools, assisting in the increased popularity of grade retention (Hauser et al. 2004, Jimerson & Kaufman 2003, McCoy and Reynolds 1999). Many states ended social promotion, in which all students are promoted to maintain homogeneity of age within classrooms, by the year 2000 with many schools having adopted grade retention at most grade levels, including kindergarten (Ellwein & Glass 1989, Hauser 1998, Roderick et. al 1999). With empirical evidence varied and suggestions unsettled, North Carolina's retention rate doubled from 1992 to 2002 (Early et al. 2003). While the evidence and opinions of researchers remains split, the rise of research on kindergarten retention in the last 20 years suggests that researchers seem to agree on the importance of kindergarten and getting retention policies right in terms of the best possible outcomes for students.

The differences across findings extends to methodologies and how effects of retention should be estimated. These discrepancies often stem from trying to account for the lack of ability and practicality to use large scale experimental designs, such as randomized control trials.

Existing studies often rely on nonexperimental data, such as observational or cross-sectional data, which raise natural issues when attempting to determine causal effects. In such a situation where the researcher cannot know how a promoted student would have performed had they been retained, they are forced to estimate a non-observable scenario, referred to as the counterfactual. Specifically, in the context of grade retention when a student has been retained, the counterfactual represents the scenario where a student had been promoted instead. Likewise, the counterfactual pertaining to a promoted student would be the case where they had been retained. Propensity score matching and propensity score stratification are often used to help estimate causal effects associated to counterfactuals by adjusting for potential selection bias (Rosenbaum & Rubin 1983). Widely cited work by Hong & Raudenbush (2005) utilizes propensity stratification by using 207 pre-treatment covariates that were found to be associated with kindergarten retention in order to estimate a child's likelihood of being retained. Work by Dong in 2010 used a control function in tandem with an instrumental variables approach to compare estimates with Nearest Neighbor Matching, proposed by Abadie & Imbens (2001). These methods rely on many assumptions and model specifications, such as the assumption of unconfoundedness. That is, that there are no unobserved covariates that play a role in selection. Although a rich and extensive set of covariates can be useful when contemplating the assumption of unconfoundedness, it does not guarantee all necessary covariates have been collected. Together, the resulting estimates produced by these various methodologies may still hinge on choices such as model specification, estimation method, which covariates to use as instruments, which covariates to include when estimating propensity scores, how many strata to use when grouping propensity scores or whether to utilize case weights. Although metanalysis have been conducted to compare results across studies on kindergarten retention, work in this area, like the

work done by Hong and Raudenbush (2005), stand as examples that may benefit from a within-study sensitivity analysis, namely Extended Ensemble Estimation, to help quantify robustness of the estimated effects. Namely, how sensitive are the findings in kindergarten retention, such as Hong and Raudenbush (2005), to specifications made by the researcher.

6.2 Description of Case Study

Hong and Raudenbush (2005) considered, among other questions, what the effect of kindergarten retention is on those students whoa retained. That is, how much more or less would kindergarten retainees have learned had they been promoted as opposed to being retained. Utilizing data from the Early Childhood Longitudinal Study Kindergarten cohort (ELCS-K) from the US National Center for Education Statistics (NCES), Hong and Raudenbush (2005) considered a nationally representative sample of 20,000 kindergarten students that included a rich set of covariates regarding the students, their families, teachers and schools across kindergarten and first-grade years from Fall 1998 to Spring 2000. This rich set of covariates served multiple purposes. The authors had a deep set of covariates to use as controls in order to account for the nested structure of students within schools, as well as to more effectively employ propensity score stratification to account for the estimating the counterfactual. Since a student is either retained or promoted, propensity score stratification uses a fixed set of covariates to compute propensity scores for each student. In this case, this refers to a student's conditional probability to be retained based on pretreatment personal and classroom characteristics, school characteristics, as well as the residual random effect of the pretreatment school of that student. Formally put, the estimation of an individual-level propensity score, q, for a retention school student i from pretreatment school j is

$$q\hat{}\_ij = P(Z\_i = 1 \dashv \mid D\_j = 1, X\_ij, W\_j, u\_j\hat{}\ *)$$

Where Z_i is an indicator of retention for student i, $D_j$is an indicator of a retention policy for school j, X_{ij} are the pretreatment personal and classroom characteristics, W_j are school level characteristics, and u_j^* is the residual random effect of the pretreatment school j. It's useful to note that in order for a student to be retained, $Z_i = 1$, it is necessary for that student's school to have a retention policy, $D_j = 1$. That also implies that no child was retained under a non-retention policy school. This propensity score stood as a gauge the risk of a student repeating kindergarten. If a student's propensity score was too low to match to a student who had been retained, the aforementioned student was considered to be at no risk of retention, while student who had such matches were considered to be at-risk of retention. To account for the varying degrees of risk of retention, the logit of the estimate propensity scores were split into 15 strata, which were balanced using 207 pretreatment covariates. It's useful to note that eight retainees in the last defined strata did not match with any students in the promoted group, thus their analysis utilized the other 14 strata.

In order to estimate the effect of kindergarten retention, Hong and Raudenbush utilized a two-level hierarchical linear model, specified below.

$$Y_{ij} = \gamma_0 + \delta_z Z_{ij} + \gamma_1 Logit(\hat{q})_{ij} + \sum_{s=2}^{15} \alpha_s L_{sij} + \gamma_2 Dur_{ij} + u_{0j} + u_{1j} Z_{ij} + e_{ij}$$

Their model included both fixed and random effects to model the outcome variable Y_{ij}, the assessment for student i in school j. They considered both math and reading assessments for outcomes, resulting in estimated effects for both math and reading. These outcome variables, found within the ECLS-K data, are scale scores calibrated using item response theory (Hambleton, Swaminathan, & Rogers, 1991). Each subject had up to four repeated assessments over the two study years, then were equated on the same scale. This standardization allows

researchers to asses math and reading growth of students over time and to compare achievement of students from different grade levels. The coefficients to be estimated for student i in school j included the binary indicator of whether or not was a kindergarten retainee, Z_{ij}, the logit of the estimated propensity score, $Logit(\hat{q})_{ij}$, the propensity strata, L_{sij}, and the duration of time between the beginning of the school year and when a student took the assessment, Dur_{ij}. The random effects included an intercept to capture the setting specific increment to a child's learning outcome, and binary indicator of whether or not was a kindergarten retainee, Z_{ij}, to capture the setting specific increment to a child's retention effect. The coefficient of the retention effect, δ_z, was the main interest in determining the answer to the authors research question; what was the effect of kindergarten retention on those students who were retained? In the extended ensemble estimation framework, the retention effect is the treatment variable of interest while the effect to be estimated, δ_z, is the estimate to examine for sensitivity and robustness across various possible alternative specifications. Namely, $\hat{\delta}_z$ is the estimated difference in a child's assessment due to being retained.

6.2.1 Findings from Hong and Raudenbush

Hong and Raudenbush estimated the fixed effect of retention to be -9.01 with a standard error of 0.68 regarding reading achievement. Specifically, if a promoted child had been retained instead, they estimated that the expected reading achievement score would be 9.01 points lower at the end of the treatment year. Hong and Raudenbush estimated the fixed effect of retention to be -5.89 with a standard error of 0.50 regarding math achievement. Specifically, if a promoted child had been retained instead, they estimated that the expected math achievement score would be 5.89 points lower at the end of the treatment year. Extended ensemble estimation could be used at this stage to estimate these effects under various specifications, other than the

specification made by the authors, to gauge the sensitivity and robustness of these estimates. In other words, how much do the negative effects of retention estimated by the authors depend on their particular specifications? Accounting for the precision of the estimated effects of retention under various specifications, what is the weighted effect of kindergarten retention on math and reading achievement scores?

6.3 Extended Ensemble Estimation

In order to carry out the extended ensemble estimation process, one may start by controlling for different covariates within the modeling stage that may explain observed differences in retainees and promoted students regarding their assessments. As noted before, ensemble techniques generally perform better when the pool of models is rich but not oversaturated by poor models, thus selecting covariates to control for outside of the authors specifications requires care as a poor pool of models may render the extended ensemble estimation process less informative. Using the rich set of covariates within the ECLS-K data, 9 covariates that include students prior math scores, reading scores, general knowledge scores, and school ID can be used to create 512 alternate specifications.

6.3.1 Results of Extended Ensemble Estimation

The distribution of estimated retention effects for the 512 alternative specifications are shown below. When excluding all controls, the estimated retention effect on reading scores is -21. The mean and median retention effect are both -10. When weighting for precision, the ensemble retention effect is -10.0433 with a standard error of 0.0842. Although the mean and median estimated effects of -10, as compared to -21 when excluding all controls, stand as evidence that there are factors that should be accounted for in terms of retention and achievement, the estimated effect of retention on reading achievement by the authors remained to

be the most conservative estimated effect of retention, even after weighting with the ensemble estimate. It is also worth noting that the difference in estimated retention effect was small when using the wealth of covariates for propensity score stratification as opposed to using pre-test controls, as the mean and median estimated effects were both -10 and the authors estimate effect was -9.

Using the Bayesian framework in Chapter 3 inspired by Frank and Min (2007), an updated retention effect on reading scores while utilizing their respective standard errors could be calculated by

$$\beta_{updated} = \left[\pi\beta_{HongRaudenbush} + (1-\pi)\beta_{EEE}\right]$$

where π could represent the efficiency of the Extended Ensemble Estimate, specifically,

$$\pi = \frac{(SE_{EEE})^2}{(SE_{EEE})^2 + \left(SE_{HongRaudenbush}\right)^2} = \frac{(0.0842)^2}{(0.0842)^2 + (0.68)^2} = 0.0151$$

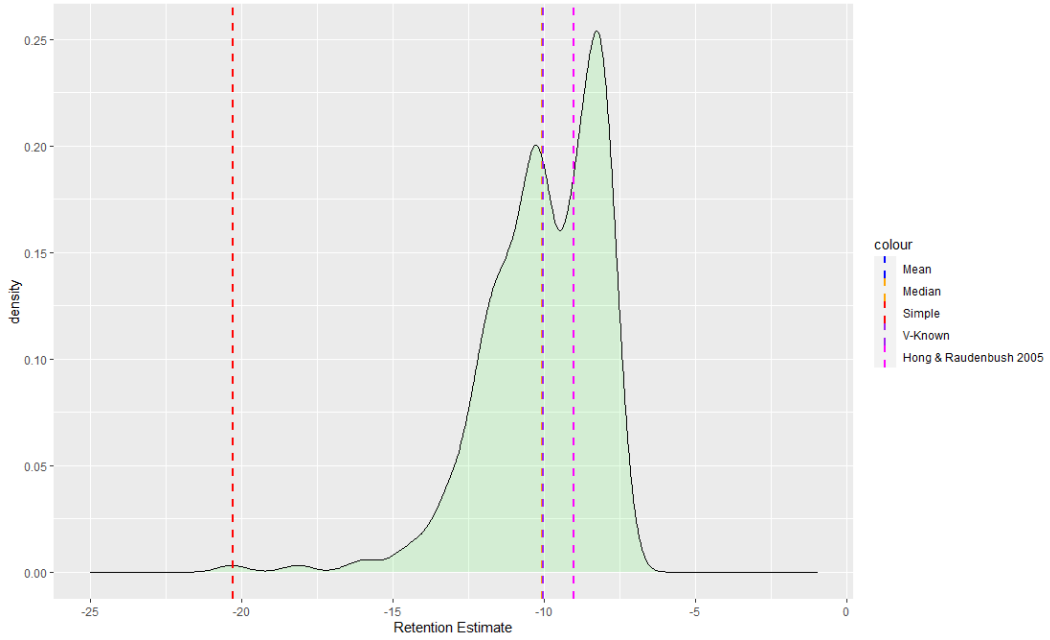
Thus, the updated estimated retention effect on reading scores is

$$\beta_{updated} = \left[\pi\beta_{HongRaudenbush} + (1-\pi)\beta_{EEE}\right]$$

$$= [(0.0151)(-9.01) + (1 - 0.0151)(-10.0433)]$$

$$= -10.0277$$

Figure 6.3.1 Distribution of estimated treatment effects



6.4 Discussion

Since students are nested within schools, students in one school may share a number of attributes that are important to account for regarding retention that may not be shared by students at other schools. From a statistical standpoint, this violation of independence may reduce the effective sample size. That is, 500 student observations that are not independent may result in a much smaller effective sample size, depending on how correlated they are in terms of measured attributes. Since standard errors are inversely proportional to sample size, ignoring dependence can result in underestimating variance or overestimating the accuracy of the effect in question (Raudenbush & Bryk, 2002; Gelman and Hill, 2007).

If retention is not independent of school level characteristics, then controlling for schools with random effects may not be sufficient to eliminate bias. Generally, clustering needs to be independent of treatment when accounting for clustering as a random effect. Randomizing treatment to students instead of schools would be one solution to this problem. When the ideal

experimental design is not possible, as is the case with kindergarten retention when randomization is not possible, Theobald and Freeman (2014) detail how controlling for student nonequivalence is crucial. This is demonstrated by Hong and Raudenbush's use of the rich set of student-level covariates.

In the case where such a specification is inappropriate or specifications leave out important covariates, such estimates should be the minority of many estimates from a pool of plausible specifications. Correlations across estimated treatment effects may indicate misspecification, specifically in the case of missing important covariates. For example, a large portion of specifications missing an important covariate could all result in under or overestimated treatment effects. Thus, examining the pool of plausible specifications for such correlated estimates to help assess the pool of plausible alternative specifications should be a priority of the user. As long as the pool of plausible alternative specifications is rich, the influence of mis-specifications are limited.

Simulations could be used to test the sensitivity of missing covariates that impact treatment effects. A step-by-step process to achieve this would start by simulating sample data from a pre-specified model with a known treatment effect. Once the sample data is simulated, one would create specifications to estimate the known treatment effect, proceeding with the extended ensemble estimation process to calculate the distribution of estimated treatment effects and the extended ensemble estimate. In order to test the sensitivity of misspecification, one would create a pool of specifications that, for example, are missing a key covariate. Then, correlations among models that omit the same covariate can be observed over many specifications. Performing extended ensemble estimation on this pool of poor specifications

would result in an extended ensemble estimate that could be compared to the known estimated treatment effect.

The findings from extended ensemble estimation regarding kindergarten retention suggest that the estimated retention effect on reading achievement reported by the authors contains a moderate level of robustness, while also erroring on the side of conservative, relative to the effects estimated using the extended ensemble estimation approach. They also suggest that since the authors findings were among the most conservative, there is evidence of a measurable effect of kindergarten retention on reading achievement that is significant.

BIBLIOGRAPHY

Angrist, J., & Imbens, G. (1995). Identification and estimation of local average treatment effects.

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, *24*(2), 3-30.

Barreto, H., & Howland, F. (2006). *Introductory econometrics: using Monte Carlo simulation with Microsoft excel*. Cambridge University Press.

Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics*, *17*(4), 341-362.

Becker, B. J. (1995). Corrections to "Using results from replicated studies to estimate linear models". *Journal of Educational and Behavioral Statistics*, *20*(1), 100-102.

Becker, B. J., & Aloe, A. M. (2019). Model-based meta-analysis and related approaches. *The handbook of research synthesis and meta-analysis*, 339-363.

Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. Journal of statistical planning and inference, 25(3), 303-328.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 317-335.

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random□effects regression model for meta□analysis. *Statistics in medicine*, *14*(4), 395-411.

Breiman, L. (1996). *Bias, variance, and arcing classifiers*. Tech. Rep. 460, Statistics Department, University of California, Berkeley, CA, USA.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1-32.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism?(don't expect an easy answer). *Journal of personality and social psychology*, *98*(4), 550.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*(2), 83-87.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*(4), 685-709.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict management and peace science*, *22*(4), 341-352.

Cohen, P., West, S. G., & Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences*. Psychology press.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational evaluation and policy analysis*, *24*(3), 175-199.

De Vaus, D. (2001). Research design in social research. *Research design in social research*, 1-296.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, *210*, 2-21.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177-188.

Dong, Y. (2010). Kept back to get ahead? Kindergarten retention and academic performance. *European Economic Review*, *54*(2), 219-236.

Early, D., Bushnell, M., Clifford, R., Konanc, E., Maxwell, K., Palsha, S., & Roberts, L. (2003). North Carolina early grade retention in the age of accountability. *Chapel Hill: The University of North Carolina, FPG Child Development Institute*, *4*.

Ellwein, M. C., & Glass, G. V. (1989). Ending social promotion in Waterford: Appearances and reality. *Flunking grades: Research and policies on retention*, 151-173.

Fisher, R. A. (1924). 035: The Distribution of the Partial Correlation Coefficient.

Frank, K. (2000). Impact of a Confounding Variable on the Inference of a Regression Coefficient. *Sociological Methods and Research*, *29*(2), 147-94.

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437-460.

Frank, K., & Min, K. S. (2007). 10. Indices of Robustness for Sample Representation. *Sociological Methodology*, *37*(1), 349-392.

Gadbury, G. L., & Allison, D. B. (2012). Inappropriate fiddling with statistical analyses to obtain a desirable p-value: tests to detect its presence in published literature.

Gelman, A., & Hill, J. (2006*). Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American scientist*, *102*(6), 460.

Goos, M., Pipa, J., & Peixoto, F. (2021). Effectiveness of grade retention: A systematic review and meta-analysis. *Educational Research Review*, *34*, 100401.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hauser, R. M. (2000). Should We End Social Promotion? Truth and Consequences.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153-161.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological methods*, *3*(4), 486.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181-188.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational evaluation and policy analysis*, *27*(3), 205-224.

Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: an application of propensity score method to multivariate, multilevel data. *Developmental Psychology*, *44*(2), 407.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic literature*, *48*(2), 399-423.

Hong, G., & Yu, B. (2007). Early-grade retention and children's reading and math learning in elementary years. *Educational evaluation and policy analysis*, *29*(4), 239-261.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School psychology review*, *30*(3), 420-437.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Publications.

Klass, G. (2008). Just plain data analysis: Common statistical fallacies in analyses of social indicator data. *Statlit. org*, *6*.

Laan, A., Madirolas, G., & De Polavieja, G. G. (2017). Rescuing collective wisdom when the average group opinion is wrong. *Frontiers in Robotics and AI*, *4*, 56.

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International journal of technology assessment in health care*, *6*(1), 5-30.

Lee, P. M. (1989). *Bayesian statistics* (pp. 54-5). London:: Oxford University Press.

LeSage, J. P., & Parent, O. (2007). Bayesian model averaging for spatial econometric models. *Geographical Analysis*, *39*(3), 241-267.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists* (Vol. 1). New York: Springer.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*(428), 1535-1546.

Madigan, D., York, J., & Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, 215-232.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750-773.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American statistical Association*, *78*(381), 47-55.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nelson, C., & Startz, R. (1988). Some further results on the exact small sample properties of the instrumental variable estimator.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, *37*(2), 187-204.

Park, S., Steiner, P. M., & Kaplan, D. (2018). Identification and sensitivity analysis for average causal mediation effects with time-varying treatments and mediators: Investigating the underlying mechanisms of kindergarten retention policy. *psychometrika*, *83*(2), 298-320.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. *The handbook of research synthesis and meta-analysis*, *2*, 295-316.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Robbins, H. E. (1992). An empirical Bayes approach to statistics. In *Breakthroughs in statistics* (pp. 388-394). Springer, New York, NY.

Robertson, R. M. (2021). To Retain or Not Retain: A Review of Literature Related to Kindergarten Retention. *Online Submission*

Roderick, M., Bryk, A. S., Jacob, B. A., Easton, J. Q., & Allensworth, E. (1999). Ending Social Promotion: Results from the First Two Years. Charting Reform in Chicago Series 1.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Jimerson, S. R., & Renshaw, T. L. (2012). Retention and social promotion. *Principal Leadership*, *13*(1), 12-16.

Saez-Rodriguez, J., Costello, J. C., Friend, S. H., Kellen, M. R., Mangravite, L., Meyer, P., ... & Stolovitzky, G. (2016). Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nature Reviews Genetics*, *17*(8), 470-486.

Shepard, L. A., & Smith, M. L. (1987). Effects of kindergarten retention at the end of first grade. *Psychology in the Schools*, *24*(4), 346-357.

Sidik, K., & Jonkman, J. N. (2005a). A note on variance estimation in random effects meta-regression. *Journal of biopharmaceutical statistics*, *15*(5), 823-838.

Sidik, K., & Jonkman, J. N. (2005b). Simple heterogeneity variance estimation for meta□ analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(2), 367-384.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2016). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, *20*(4), 518-529.

Su, H., & Chen, H. (2015). Experiments on parallel training of deep neural network using model averaging. *arXiv preprint arXiv:1507.01239*.

Sumiati, I., Handoyo, F., & Purwani, S. (2020). Multiple linear regression using Cholesky decomposition. *World Scientific News*, *140*, 12-25.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261-293.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of mathematical psychology*, *44*(1), 92-107.

Wooldridge, J. M. (2009). Omitted variable bias: the simple case. *Introductory Econometrics: A Modern Approach. Mason, OH: Cengage Learning*, 89-93.

Young, C. (2018). Model uncertainty and the crisis in science. *Socius*, *4*, 2378023117737206