THE BAYESIAN PARADIGM OF ROBUSTNESS INDICES OF CAUSAL INFERENCES

By

Tenglong Li

A DISSERTATION

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

Measurement and Quantitative Methods—Doctor of Philosophy

2018

ABSTRACT

THE BAYESIAN PARADIGM OF ROBUSTNESS INDICES OF CAUSAL INFERENCES

By

Tenglong Li

The validity of a causal inference hinges on a research design with both strong internal validity and strong external validity (Shadish et al. 2002). Unfortunately, such research is rare so that causality is typically inferred through a small-scale randomized experiment or a large-scale observational study (Schneider et al. 2007). In light of this gap, the robustness indices of causal inferences have been proposed by Frank et al. (2013) to measure the robustness of causal inference by quantifying the proportion of the observed sample that needs to be replaced with unfavorable unobserved cases.

Drawing on the Bayesian discussion in Frank & Min (2007), this dissertation purposes developing the Bayesian framework of the robustness indices of causal inferences for causal research with either limited internal validity or limited external validity. This dissertation has two chapters: The first chapter lays the foundation of the Bayesian paradigm of robustness indices by formally defining prior as distribution built on an unobserved sample. For a particular family of prior and likelihood distributions, the posterior can be interpreted as distribution built on an ideal sample. The Bayesian paradigm of robustness indices of causal inferences focuses on the relationship between the posterior probability of invalidating an inference and the unobserved sample statistics and the central task is to locate the threshold of an unobserved sample statistics with regard to a given value of the posterior probability of invalidating an inference. Considering the first chapter targets the simple group-mean-difference estimator only, the second chapter extends the Bayesian paradigm of robustness indices to regression models. This dissertation

promotes the scientific discourse of causality and critical thinking by linking the probability of invalidating an inference to detailed thought experiments characterized by the thresholds of sufficient statistics pertaining to an unobserved sample.

Copyright by
TENGLONG LI
2018

ACKNOWLEDGEMENTS

Throughout my doctoral study, I have been so blessed to received tremendous help and support from so many people. First and foremost, I am extremely grateful to my advisor Prof. Kenneth Frank who has instilled a scientific attitude as well as rigorous thought in me and imparted me immense knowledge on causal inferences and robustness indices. He has demonstrated to me what is a good researcher not only through advising but also through his own practice. Particularly, this dissertation cannot exist, let alone evolve into its current shape, without his guidance and support. I also own a debt of gratitude to Prof. Maier who has offered me so many insights and advices on Bayesian statistics and to Prof. Konstantopoulos who has given me quite a few invaluable suggestions on causal inference and teaching. I am indebted to Prof. Imberman who introduces me to the economic perspective of causal inference and relevant literature, which really broadens my horizon.

This is the best opportunity to express my sincere gratitude to my beloved family. I cannot concentrate on my work and make any progress without the support and love of my wife Senmu Zheng. Finally, my parents deserve special thanks for bringing me to this world, encouraging me to pursue my passion, supporting me wholeheartedly and much more.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

Chapter 1: The Bayesian paradigm of robustness indices of causal inferences 1
- 1-Introduction 1
  - 1.1-The robustness indices of causal inferences 1
  - 1.2-The conceptualization of unobserved sample 4
  - 1.3-Previous work on the Bayesian framework of the robustness indices 13
  - 1.4-Purposes of this study 15
- 2-The unifying framework of the robustness indices of causal inferences 16
  - 2.1-The frequentist recipe 18
  - 2.2-The Bayesian recipe 20
- 3-The Bayesian models of robustness indices for internal validity 27
- 4-The Bayesian models of robustness indices for external validity 32
- 5-Statistical threshold and Bayesian models for replacing observed cases 39
  - 5.1-Appropriate statistical threshold $\delta^{\#}$ for Bayesian models of robustness indices 40
  - 5.2-The Bayesian models of robustness indices for replacing observed cases 43
- 6-Demonstrative examples 46
  - 6.1-The Bayesian robustness indices of the effect of OCR on reading achievement 46
  - 6.2-The Bayesian robustness indices of the effect of kindergarten retention on reading achievement 59
- 7-Discussion and conclusion 68
  - 7.1-Features of the Bayesian paradigm of robustness indices 68
  - 7.2-Comparisons with other similar approaches 70
    - 7.2.1-The robustness indices in Frank et al. (2013) 70
    - 7.2.2-The robustness indices in Frank & Min (2007) 71
    - 7.2.3-The bounds on treatment effect in Manski (1990) and Lee (2009) 71
  - 7.3-Limitations 72
  - 7.4-Conclusion 73

Chapter 2: The Bayesian paradigm of robustness indices of causal inferences for regression models 74
- 1-Introduction 74
  - 1.1-Regression-based causal inference 74
  - 1.2-The philosophy of robustness indices 76
  - 1.3-Research objectives 84
- 2-The unifying framework of robustness indices for regression-based causal inference 86
  - 2.1-Setting and notation 86
  - 2.2-The frequentist recipe 87
  - 2.3-The Bayesian recipe 91
- 3-Bayesian models of robustness indices for raw data 94

3.1-Data and the sample statistics 94
3.2-The posterior distribution of γ_w for raw data 98
3.3-Probit models for the probability of invalidating an inference 101
3.4-External validity and internal validity 103
4-Bayesian models of robustness indices for centered and standardized data 107
4.1-For centered data 107
4.2-For standardized data 109
5-Appropriate statistical threshold and Bayesian models for replacing observed cases 111
5.1-Appropriate statistical threshold 111
5.2-The Bayesian models of robustness indices for replacing observed cases 113
6-Illustrative examples 115
6.1-The Bayesian robustness indices of the effect of Open Court Reading on reading achievement 115
6.2-The Bayesian robustness indices of the effect of kindergarten retention on reading achievement 126
7-Discussion 134
7.1-A summary of the Bayesian paradigm of robustness indices for regression-based causal inference 134
7.2-Comparisons with other similar approaches 136
7.2.1-The impact thresholds in Frank (2000) 136
7.2.2-The robustness indices in Frank & Min (2007) 138
7.2.3-The robustness indices in Frank et al. (2013) 138
7.3-Limitations 139
7.4-Conclusion 140

APPENDICES 141
Appendix A: Proofs of Theorem 1 and Theorem 2 142
Appendix B: The Algebraic Equivalence Between Theorem 1 and Common Expressions of Regression Coefficients 149

REFERENCES 153

LIST OF TABLES

Table 1.1: Thresholds of α when π_R is fixed as 0.46 55

Table 1.2: Thresholds of π_R when α is fixed as 1 56

Table 1.3: Thresholds of α when π is fixed as 0.0617 65

Table 2.1: Thresholds of n^{un} assuming $\hat{\sigma}_{WY}^{un} = 0$ 122

Table 2.2: Thresholds of r_{WY}^{un} 131

LIST OF FIGURES

Figure 1.1: The structure of ideal population in Hong & Raudenbush (2005) 11

Figure 1.2: The structure of ideal population in Borman et al. (2008) 12

Figure 1.3: The structure of ideal population in both scenarios 13

Figure 1.4: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Borman et al. (2008) (π_R is fixed as 0.46) 57

Figure 1.5: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Borman et al. (2008) (α is fixed as 1) 58

Figure 1.6: The relationship between α and the probability of invalidating the inference of Hong & Raudenbush (2005) 66

Figure 1.7: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Hong & Raudenbush (2005) 67

Figure 2.1: The structure of ideal population in Hong & Raudenbush (2005) 82

Figure 2.2: The structure of ideal population in Borman et al. (2008) 83

Figure 2.3: The structure of ideal population in both scenarios 84

Figure 2.4: The relationship between n^{un} and the probability of invalidating the inference of Borman et al. (2008) 123

Figure 2.5: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Borman et al. (2008) 124

Figure 2.6: The relationship between r_{WY}^{un} and the probability of invalidating the inference of Hong & Raudenbush (2005) 132

Figure 2.7: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Hong & Raudenbush (2005) 133

Chapter 1: The Bayesian paradigm of robustness indices of causal inferences

1-Introduction

1.1-The robustness indices of causal inferences

The issues of reproducibility and generalizability have plagued the scientific community. For example, Open Science Collaboration (2015) has reported that a substantial proportion of the selected psychological studies failed to be replicated by other parties. To promote the replicability and possibly generalizability of published research, various scholars and organizations have called on enforcing higher standards and rigorous checks of the research designs and statistical analytical procedures.

Particularly, when it comes to research which attempts to support causal inferences, the concerns about reproducibility and generalizability become even stronger since one has to wrestle with both the internal validity and external validity of his design. Due to the nature of causal inference, researchers can never rule out all possible threats to both internal validity and external validity. Therefore, oftentimes they are uncertain about the degrees to which they can justify or reject their conclusions. In light of such a headache, the analyses of robustness or sensitivity of causal inference have been proposed by different scholars. The robustness indices suggested by Frank et al. (2013) is of particular interest as it naturally arises from the context of the empirical research.

The idea of the robustness indices is straightforward in Frank et al. (2013). There are three key quantities in this framework of the robustness indices, namely the estimated effect $\hat{\delta}$, the threshold $\delta^{\#}$ and the population effect δ. The estimated effect is the effect researchers estimate based on their obtained samples and research designs. The threshold is a fixed value predetermined by the researchers so that they can compare their estimated effect with the thresholds they set. For an instance, in order to claim that attending Catholic high schools can enhance the academic achievement of students, one has to get an estimated effect of the attendance of Catholic schools on students' test scores and prove it is larger than the threshold he set up in his research. Usually the aforementioned threshold is chosen to be the same as the threshold determining the statistical significance for specific research hypothesis and collected sample. The population effect will remain unknown as always in empirical research.

According to Frank et al. (2013), the inference will be invalid if the following condition is satisfied:

$$\hat{\delta} > \delta^{\#} > \delta \tag{1.1}$$

Or equivalently, if β is used to denote the bias:

$$\beta = \hat{\delta} - \delta > \hat{\delta} - \delta^{\#} \tag{1.2}$$

It is necessary to point out that the above formulae will only apply to the situations of inferring positive effects. The counterparts of formulae (1.1) and (1.2) for inferring the negative effects are easy to be derived as follows, from the same reasoning:

$$\hat{\delta} < \delta^{\#} < \delta \tag{1.3}$$

$$\beta = \hat{\delta} - \delta < \hat{\delta} - \delta^{\#} \tag{1.4}$$

The rest arguments of Frank et al. (2013) directly follow from the preceding rules, and by partitioning the sample into the parts with and without bias, the robustness indices could be expressed as the proportion of the sample to be replaced by the new data for which the treatment effect is zero. Such proportion is interpreted as the replacement that is necessary to invalidate the inference.

In an empirical context, the estimated effect $\hat{\delta}$ is fixed and the true causal effect δ is a parameter. The threshold $\delta^{\#}$ could be a subjective choice based on policy implication or an objective choice based on level of significance, which means $\delta^{\#}$ is not necessarily fixed. Given the natures of $\hat{\delta}$, δ and $\delta^{\#}$, it's possible to simplify the decision rules in (1.2) and (1.4) further as follows:

$$\begin{aligned} &\delta > \delta^{\#} \text{ for inferring a positive effect} \\ &\delta < \delta^{\#} \text{ for inferring a negative effect} \end{aligned} \tag{1.5}$$

Frank et al. (2013) offered two examples, namely Hong & Raudenbush (2005) and Borman et al. (2008), to illustrate the procedure of quantifying necessary bias to invalidate an inference using the decision rules above.

Hong & Raudenbush (2005) is a research whose goal was to evaluate the effect of kindergarten retention on academic achievement. In this example, it was impossible to randomly assign the sampled students to the conditions of being retained in kindergarten and being promoted to the first grade. According to Rubin Causal Model (RCM), every sampled student should have two potential outcomes, namely, one outcome under the condition of being retained and one outcome under the condition of being promoted. Draw on RCM, the only sample that will lead to true causal inference, is supposed to be composed of reading scores of all sampled students assuming they were all retained in kindergarten and reading scores of all sampled students assuming they were all promoted to the first grade. Such sample is very ideal since no students could be retained and promoted simultaneously. In this case, the bias is induced by gap between the ideal sample which consists of potential reading score of every sample student under both retention and promotion and the observed sample which only has reading score of every sampled student under either retention or promotion (but not both).

Borman et al. (2008) studied the effect of Open Court Reading (OCR) curriculum on students' reading achievement by randomly drawing schools which showed strong interest in this program to the publisher of OCR and volunteered in their study. Particularly, Frank et al. (2013) pointed out that it would be questionable to generalize the inference made based on the observed sample to the population of schools that didn't volunteer in this program in the first place, since it's possible that the volunteered schools might benefit more from OCR because they had better plans and more experience comparing to the population of schools which were less attracted to this curriculum and didn't volunteer in the OCR program in the first place. Consequently, the observe sample might not be well-represented of the entire population of schools, which includes both volunteered schools and non-volunteered schools. In this case bias is induced by the gap between a random sample of the entire population of schools and the observed sample which can only represents volunteered schools.

Each of both examples epitomizes a distinguished scenario where a causal inference is prone to bias and invalidation. Specifically, Hong & Raudenbush (2005) typifies a scenario where external validity is strong because the observed sample is representative of the target population but internal validity is weak due to a lack of randomization. This scenario is referred to as "the first scenario" throughout this paper. On the other hand, Borman et al. (2008) exemplifies another scenario where internal validity is sound because of randomization but external validity is worrisome as the observed sample can only represents a part of the target population. This scenario is referred to as "the second scenario" henceforth.

1.2-The conceptualization of unobserved sample

The gist of the framework of robustness indices of causal inferences put forth by Frank & Min (2007) and Frank et al. (2013) is that bias β is induced by the gap between the observed sample and the sample one is supposed to obtain for his inference and conclusion. In this study, I intend to address and fill this gap through the conceptualization of unobserved sample. Relying on Rubin Causal Model and especially its potential outcome framework, I have the following definitions:

Definition 1.1: **A real or non-counterfactual outcome** refers to an outcome which is observable, i.e., an outcome of a controlled subject under the condition of control or an outcome of a treated subject under the condition of treatment.

A real outcome in Hong & Raudenbush (2005) could either be a reading score of a retained child John under the condition of he was retained in kindergarten or a reading score of a promoted child Mary under the condition of she was promoted to first grade.

Definition 1.2: **A counterfactual outcome** of a subject refers to an imaginary outcome that would be observed under a condition which is different from what this subject actually received. In Hong & Raudenbush (2005), the counterfactual outcome of John who was retained in the kindergarten would be his reading score had he been promoted to first grade. Likewise, the counterfactual outcome of Mary who was promoted to first grade would be her reading score had she been retained in kindergarten.

Next, I define potential outcome for the first scenario based on definition 1.1 and 1.2:

Definition 1.3.1: **A potential outcome of a subject in the first scenario** refers to either his/her real outcome or his/her counterfactual outcome.

In Hong & Raudenbush (2005), every student had two potential outcomes. For example, John (who was actually retained) had two potential outcomes, which were his reading score (real outcome) under the condition of being retained in kindergarten and his reading score (counterfactual outcome) under the condition of being promoted to first grade. Similarly, Mary (who was actually promoted) had two potential outcomes which were her reading score (real outcome) under the condition of being promoted to first grade and her reading score (counterfactual outcome) under the condition of being retained in kindergarten.

Before I proceed to define a potential outcome for the second scenario, it’s vital to appreciate the difference between the first scenario and the second scenario: In the first scenario, the lack of randomization means that real outcomes and counterfactual outcomes are fundamentally different and therefore should not be treated as equals. In the second scenario, counterfactual outcomes can be considered to be equivalent to real outcomes in the long run due to randomization. This suggests the discussion and definition of potential outcomes in the second scenario can be confined to real outcomes only. Hence, I have the following definition of potential outcomes for the second scenario:

Definition 1.3.2: **A potential outcome in the second scenario** refers to a real outcome which could be potentially drawn from the target population.

Given the target population of Borman et al. (2008) consists of both volunteered and non-volunteered schools, a potential outcome in Borman et al. (2008) could be either the mean reading score of a classroom which belonged to a volunteered school in their study or the mean reading score of a classroom that could be potentially drawn from non-volunteered schools. It’s remarkable that definition 1.3.2 implies a random assignment of classrooms to the groups of Open Court Reading and control in either volunteered schools or non-volunteered schools.

Both the first scenario and the second scenario share the same definition of ideal population, which is provided next:

Definition 1.4: **An ideal population** refers to the collection of all possible potential outcomes of the target population.

The ideal population of Hong & Raudenbush (2005) is the collection of reading scores of all U.S. kindergarten children under both conditions of retention and promotion. Likewise, the ideal population of Borman et al. (2008) is the collection of mean reading scores of all U.S. classrooms. I remark here that the ideal population of Hong & Raudenbush (2005) contains counterfactual outcomes while the ideal population of Borman et al. (2008) comprises real outcomes only.

To fathom the bias invalidating causal inference in both scenarios and its creation, I further decompose an ideal population into two parts, namely the observed part and the unobserved part and distinguish them with the following two definitions:

Definition 1.5.1: **The unobserved part of an ideal population in the first scenario** refers to the collection of all counterfactual outcomes of the target population. Naturally, **the observed part of an ideal population in the first scenario** refers to the collection of all real outcomes of the target population.

Definition 1.5.2: **The unobserved or non-representable part of an ideal population in the second scenario** refers to the collection of all potential outcomes of the part of the target population that cannot be represented by the observed sample. Conversely, **the observed or representable part of an ideal population in the second scenario** refers to the collection of all potential outcomes of the part of the target population that was deemed to be logically represented by the observed sample.

Again, I use Hong & Raudenbush (2005) and Borman et al. (2008) to concretize the above two definitions. The unobserved part of the ideal population of Hong & Raudenbush (2005) would be the collection of counterfactual reading scores of all U.S. kindergarten students, i.e., the reading scores of all U.S. kindergarten students under retention when they were all promoted to first grade and the reading scores of all U.S. kindergarten students under promotion if they were all retained in kindergarten. The observed part of the ideal population of Hong & Raudenbush (2005) would be the collection of real reading scores of all U.S. kindergarten students, namely the reading scores of all U.S. kindergarten students under promotion when they were all promoted to first grade and the reading scores of all U.S. kindergarten students under retention if they were all retained in kindergarten. Furthermore, the unobserved (non-representable) part of the ideal population of Borman et al. (2008) would be the collection of the mean reading scores (real outcome) of all classrooms in the non-volunteered schools. The observed (representable) part of the ideal population of Borman et al. (2008) would be the collection of the mean reading scores (real outcome) of all classrooms in the volunteered schools.

Equipped with all aforementioned definitions, it's ready to conceptualize an unobserved sample as a random sample from the unobserved part of ideal population and formalize it with the following definitions:

Definition 1.6.1: **An unobserved sample in the first scenario** refers to the collection of counterfactual outcomes of all sampled subjects. **An unobserved treated sample in the first scenario** refers to the collection of counterfactual outcomes of sampled subjects who actually received control, that is, the collection of outcomes of treated subjects had they participated in the control group instead. **An unobserved control sample in the first scenario** refers to the collection of counterfactual outcomes of sampled subjects who actually received treatment, i.e., the collection of outcomes of control subjects had they switched to the treatment group.

Definition 1.6.2: **An unobserved sample in the second scenario** refers to an imaginary random sample which is drawn from the non-representable part of an ideal population and consists of real outcomes. I assume a subsequent randomization is carried out on this unobserved sample, and resultantly the proportion of treated subjects in this unobserved sample is the same as the proportion of the treated subjects in the observed sample. **An unobserved treated sample in the second scenario** refers to the collection of real outcomes of subjects who were assigned to the treatment group in this imaginary random sample. **An unobserved control sample in the second scenario** refers to the collection of real outcomes of subjects who were assigned to the control group in this imaginary random sample.

Definition 1.7: **An ideal sample** refers to the combination of the observed sample and an unobserved sample. **An ideal treated sample** refers to the combination of the observed treated sample and an unobserved treated sample. **An ideal control sample** refers to the combination of the observed control sample and an unobserved control sample.

According to definition 1.6.1, an unobserved sample of Hong & Raudenbush (2005) is the collection of counterfactual reading scores of sampled students in their study. Specifically, this unobserved sample can be decomposed into an unobserved control sample which is the collection of reading scores of retained students had they all been promoted to first grade and an unobserved treated sample which is the collection of reading scores of promoted students had they all been retained in kindergarten. According to definition 1.6.2, an unobserved sample of Borman et al. (2008) is an imaginary sample of classrooms which were randomly drawn from non-volunteered schools and subsequently randomly assigned to the Open Court Reading (OCR) group or the control group. This unobserved sample comprises an unobserved treated sample which is the collection of mean reading scores of the sampled classrooms in the OCR group and an unobserved control sample which is the collection of mean reading scores of the sampled classrooms in the control group.

Figure 1.1 details the structure of ideal population in Hong & Raudenbush (2005). Notationally speaking, I use Y to denote the outcome. The subscript of Y has two parts separated by a comma: The first part is used to denote which group this outcome belongs to and the second part is used to denote which subject this outcome pertains to. The superscript of Y signals which kind of sample this outcome belongs to. For example, the reading score of John (or any other student who was retained in kindergarten) is symbolized by $Y_{r,i}^{ob}$ as John was observed as the ith retained student. The conceptualization of an unobserved sample (represented by the arrows with a label '1') requires to project his reading score had he been promoted to first grade, which is denoted by $Y_{p,i}^{un}$. In this case, $Y_{r,i}^{ob}$ becomes an element of the observed treated sample and $Y_{p,i}^{un}$ is a member of an unobserved control sample. The reading score of Mary (or any other student who was promoted to first grade) is symbolized by $Y_{p,j}^{ob}$ and the conceptualization of an unobserved sample demands a projection of his reading score had she been retained in kindergarten, which is symbolized by $Y_{r,j}^{un}$. Consequently, $Y_{p,j}^{ob}$ is one element of the observed control sample and $Y_{r,j}^{un}$ is one element of an unobserved treated sample.

Figure 1.1: The structure of ideal population in Hong & Raudenbush (2005)

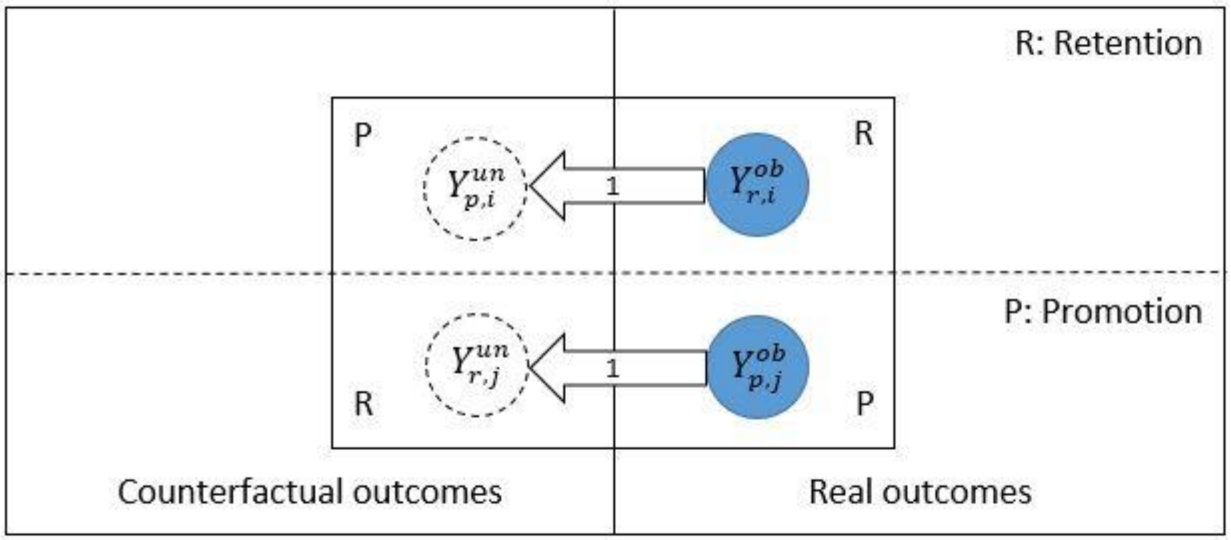


Figure 1.2 elaborates on the structure of ideal population in Borman et al. (2008). $Y^{ob}_{o,i}$ represents the mean reading score of an Open Court Reading classroom sampled from volunteered schools and it could be any single element of the observed treated sample. $Y^{ob}_{c,j}$ denotes the mean reading score of a control classroom sampled from volunteered schools and it could be any single element of the observed control sample. To generalize the conclusion of Borman et al. (2008) convincingly to non-volunteered schools, one needs the conceptualization of an unobserved sample (represented by the arrows with a label ‘2’) which is defined as an imaginary random sample of classrooms from non-volunteered schools. After an imaginary random assignment of classrooms in this unobserved sample to Open Court Reading or control, the mean reading score of an Open Court Reading classroom in this unobserved sample is $Y^{un}_{o,k}$ which could be any single element of an unobserved treated sample. The mean reading score of a control classroom in this unobserved sample is $Y^{un}_{c,l}$ which could be any single element of an unobserved control sample. As specified in definition 1.6.2, the proportion of OCR classrooms in this unobserved sample should be equal to the proportion of OCR classrooms in the observed sample.

Figure 1.2: The structure of ideal population in Borman et al. (2008)

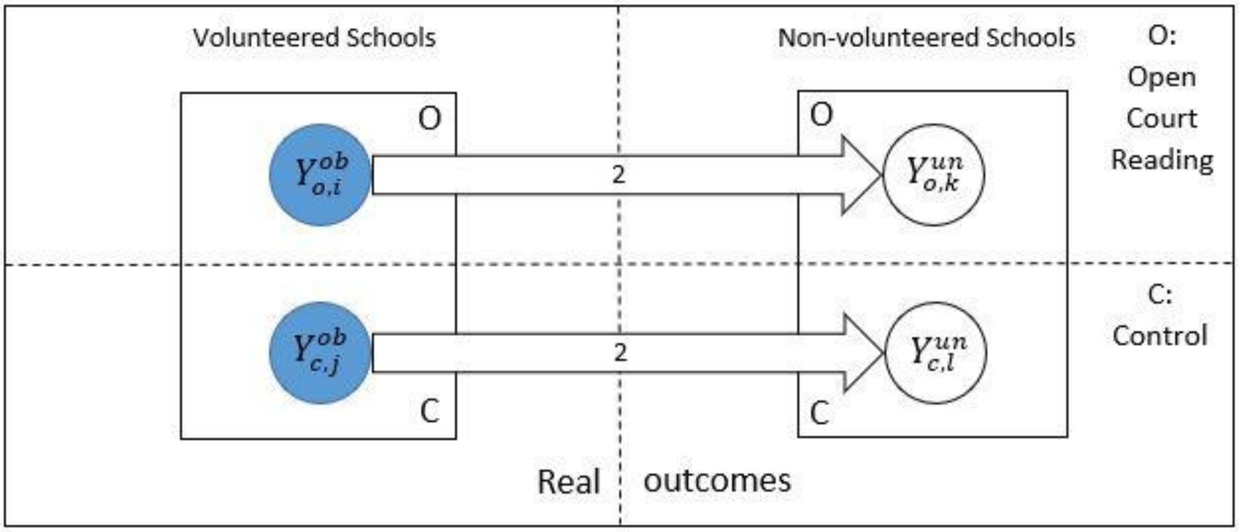
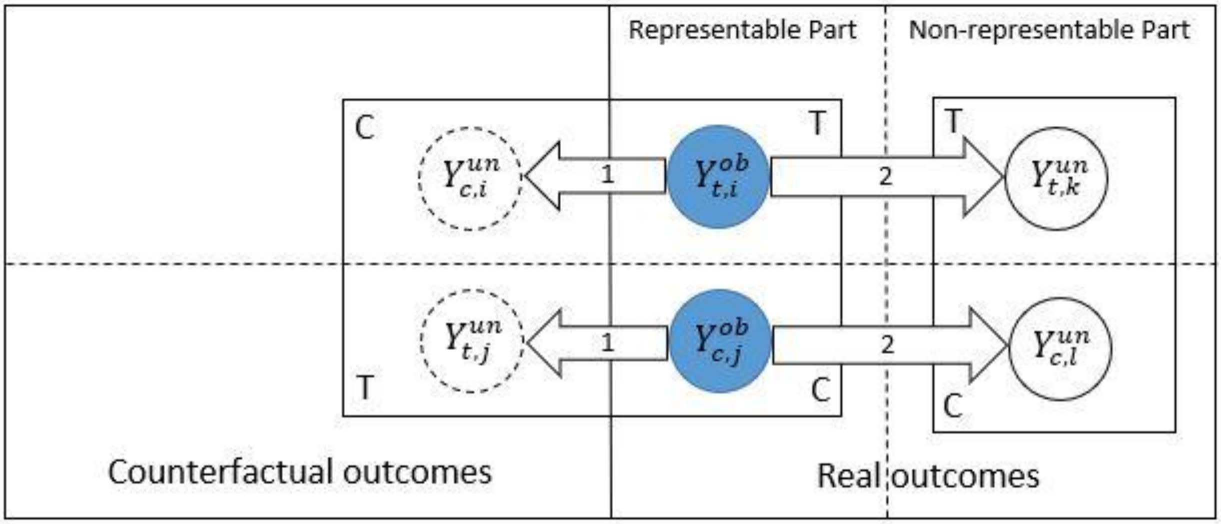


Figure 1.3 synthesizes above two figures and portrays the structure of ideal population in both scenarios. $Y^{ob}_{t,i}$ signifies the outcome belongs to the observed sample and subject i which is a member of the treatment group. In other words, $Y^{ob}_{t,i}$ could be associated with any member in the observed treated sample. In the first scenario, the conceptualization of an unobserved sample is tantamount to the projection of a counterfactual outcome (dashed circle) for each real outcome (blue-shaded circle) in the observed sample. For example, for treated outcome of subject i in the observed sample, it's necessary to project this subject i's counterfactual outcome had he participated in control group (i.e., $Y^{un}_{c,i}$, which is an element of an unobserved control sample). In the second scenario, the conceptualization of an unobserved sample is a process of projecting a random sample in the non-representable part of ideal population and conceptually forming treatment and control group within this random sample by random treatment assignment. In this case, the outcomes in the observed sample (blue-shaded circles) and the outcomes in an unobserved sample (solid unshaded circles) are both real outcomes as they pertain to different subjects (as manifested by their different subscripts i, j, k, l).

Figure 1.3: The structure of ideal population in both scenarios



To summarize, it is worthy to point out that every study is associated with an ideal sample when its robustness of causal inference is of main concern. An ideal sample is thought to be comprised of the observed sample and an unobserved sample. The observed sample represents the observed part of ideal population while an unobserved sample is thought of as a random sample from the unobserved part of ideal population which cannot provide any real observed data even though it is essential for causal inference. Throughout this paper, the observed sample is considered as fixed while an unobserved sample must be varying instead of fixed. Holding the observed sample fixed, the estimate based on the observed sample will be "contaminated" when I expand the observed samples with an unobserved sample. As a result, it is this unobserved sample that alter the sample statistics (see Frank & Min, 2007) and induce bias which renders internal validity or external validity vulnerable. Therefore, the conceptualization and modeling of an unobserved sample is indispensable in quantifying the robustness of causal inference.

1.3-Previous work on the Bayesian framework of the robustness indices

A Bayesian framework of the robustness indices has been offered by Frank & Min (2007) in a slightly different setting than the robustness indices I have discussed so far. However, their argument about the formation of this Bayesian framework is quite illuminating. Specifically, they defined the sampling distribution of the correlation computed based on an unobserved sample as the prior and modeled the sampling distribution of the correlation computed based on the observed sample as the likelihood. Therefore, the prior and likelihood can be combined to generate the posterior distribution in an ordinary Bayesian fashion. Most importantly, the generated posterior distribution could be interpreted as the sampling distribution of the correlation for an ideal sample, which is consisted of both observed sample and unobserved sample. Such interpretation is consistent with the fact that the posterior distribution is just the compromise between the prior and the likelihood. The Bayesian framework propounded by Frank & Min (2007) lays the foundation of construction and interpretation of the Bayesian paradigm of the robustness indices in this study.

Fundamentally, the Bayesian framework of Frank & Min (2007) resides in the Bayesian causal inference world pioneered by Rubin (1978), which proposed to impute missing counterfactual outcomes based on their predictive posterior distribution(s) conditional on the assignment mechanism, real outcomes and covariate values. This procedure is implemented by sampling counterfactual outcomes from their predictive posterior distribution(s) and re-estimate average treatment effect as the mean of individual differences between real outcomes and imputed counterfactual outcomes. It’s noteworthy that, the Bayesian framework of Frank & Min (2007), just like other literature inheriting Rubin (1978)’s Bayesian perspective of addressing causal problems (Imbens & Rubin, 1997, 2015; Rubin & Zell, 2010; Zajonc, 2012; Espinosa et al., 2016), considers counterfactual outcomes as missing data and imputes a sample of them through an underlying Bayesian model.

1.4-Purposes of this study

The robustness indices are quite user-friendly and suitable for most empirical research since they inform the researchers about how robust their causal inference could be to the potential design and sampling bias, by making them think about the sample with no effect at all and the proportion of such sample is needed if it is used to replace the original sample to invalidate the inference. Nevertheless, it would be even more straightforward if one can manage to answer the question "How likely is my inference invalid" instead of the question "What is the proportion of my sample to be replaced to necessarily alter my conclusion", since the previous question allow us to directly quantify the robustness of the causal inference as the probability of nullifying the inference. In fact, research on the probability of replicability/reproducibility of a specific study has been advanced and advocated in different fields, and the probability of invalidating an inference proposed in this paper is essentially a form of probability of replicability which has been becoming an advisable choice of statistic in scholarly publishing and reporting (See Greenwald et al., 1996; Thompson, 1996; Sohn, 1998; Killeen, 2005; Psychological Science editorial board, 2005; Miller, 2009; Iverson et al., 2010).

To express the robustness indices probabilistically, I draw on the Bayesian framework provided by Frank & Min (2007) and extend it further to the robustness indices defined in Frank et al. (2013). It's important to note here that Frank & Min only focused on the Bayesian models of the robustness indices for biased sampling and the correlation coefficient as the measurement of effect size. To make this Bayesian framework more comprehensive and applicable to the problems discussed in Frank et al. (2013), I first propose a unifying Bayesian framework of the robustness indices, which is logically identical to the framework put forth by Frank & Min (2007). I will show this unifying Bayesian framework will lead to the posterior distribution of the bias, which allows one to calculate the probability that the bias exceeds its corresponding threshold for a certain study by utilizing the rules of overturning the inference as defined in (1.2) and (1.4). Given the motivations and mechanisms of the bias concerning the internal validity and external validity are different, separate Bayesian models for those two distinct kinds of bias are subsequently developed from the unifying Bayesian framework of the robustness indices.

In the following section, I will first present the unifying framework of the robustness indices of causal inferences. This unifying framework contains two recipes of preparing robustness indices, namely a frequentist recipe and a Bayesian recipe. In the third section, I define the Bayesian models of robustness indices (the Bayesian recipe) specifically in terms of a research which has limited internal validity. Such Bayesian models typically will be applied to an observational/quasi-experimental study. In the fourth section, I particularly define the Bayesian models of robustness indices with regard to a research which has limited external validity. This set of Bayesian models have appropriate applications in randomized experiments. The fifth section discusses the appropriate statistical threshold $\delta^{\#}$ for the Bayesian models of robustness indices as well as replacing observed cases with unobserved ones as an alternative sampling scenario. In the section of demonstrative examples, the robustness of the inferences made by Borman et al. (2008) as well as Hong & Raudenbush (2005), which has been evaluated by Frank et al. (2013), is reassessed within the corresponding Bayesian frameworks for external validity and internal validity provided in this paper. I conclude this study with a summary of the findings and point out the limitations and possibly their implications for the future research.

2-The unifying framework of the robustness indices of causal inferences

My discussion throughout this paper on the robustness indices of causal inferences is limited to the following setting: I assume there are only two groups in comparison, i.e., a treatment group whose participants received a treatment of main interest (like OCR or kindergarten retention) and a control group of subjects who didn't receive such treatment. I further assume that, contingent on the two-group design, the difference between the mean of all observed treated outcomes and the mean of all observed control outcomes is the estimate of average treatment effect $\hat{\delta}$. Throughout the text I adopt the following notations: $\mathbf{Y_t^{un}}$ is an unobserved treated sample and $\mathbf{Y_c^{un}}$ is an unobserved control sample. Moreover, the observed treated sample and the observed control sample are denoted by $\mathbf{Y_t^{ob}}$ and $\mathbf{Y_c^{ob}}$respectively. Likewise, the ideal treated sample and the ideal control sample are denoted by $\mathbf{Y_t^{ideal}}$ and $\mathbf{Y_c^{ideal}}$ respectively. The sample means of $\mathbf{Y_t^{un}}, \mathbf{Y_c^{un}}, \mathbf{Y_t^{ob}}, \mathbf{Y_c^{ob}}$ are correspondingly represented by $\bar{Y}_t^{un}, \bar{Y}_c^{un}, \bar{Y}_t^{ob}, \bar{Y}_c^{ob}$. Probabilistically, the value of a single outcome under the condition of treatment (we can call it a treated outcome) can be treated as a random variable Y_t and the value of a single outcome under the condition of control (we can call it a control outcome) can be treated as a random variable Y_c. Furthermore, the value of a treated outcome which might appear in the observed sample is a random variable $Y_t^{ob}$and the value of a control outcome which might appear in the observed sample is also a random variable Y_c^{ob}. The expectations of the distributions of Y_t and Y_c are symbolized by μ_t and μ_c respectively. Similarly, μ_t^{ob} and μ_c^{ob} stand for the expectations of the distributions of Y_t^{ob} and Y_c^{ob}. Finally, μ_t^{ideal} and μ_c^{ideal} are two random variables whose distribution are respectively conditional on an ideal treated sample and an ideal control sample. (so they are not expectations).

2.1-The frequentist recipe

It is the definition of the bias that motivates construction of the Bayesian models of the robustness indices. (Frank et al. 2013 Appendix). The bias, according to Frank et al (2013), is uniformly defined as follows:

$$\begin{aligned}\beta &= \mathrm{E}[\hat{\delta}] - \mathrm{E}[\bar{\delta}] = \{\mathrm{E}[Y_t^{ob}] - \mathrm{E}[Y_c^{ob}]\} - \{\mathrm{E}[Y_t] - \mathrm{E}[Y_c]\} \\ &= (\mu_t^{ob} - \mu_c^{ob}) - (\mu_t - \mu_c)\end{aligned} \tag{2.1}$$

To elaborate on the definition above, (2.1) is partitioned into two series of differences. The first series of differences imply that the estimate of the treatment effect based on an observed sample, which is denoted as $\hat{\delta}$ in (2.1), has an expectation equal to $\mu_t^{ob} - \mu_c^{ob}$. The second series of differences suggest that the estimate of the treatment effect based on an ideal sample, which is represented by $\bar{\delta}$ in (2.1), should have an expectation equal to $\mu_t - \mu_c$ which is the true treatment effect.

The operationalization of the above definition of bias relies on the strategy of molding β as a random variable. First, given $\hat{\delta}$ is supposed to be fixed when considering the bias associated with an estimate of causal effect, I simply use $\bar{Y}_t^{ob} - \bar{Y}_c^{ob}$ to substitute $\mu_t^{ob} - \mu_c^{ob}$ in (2.1) at the cost of ignoring the random sampling error associated with the observed sample. Second, an unobserved sample needs to be taken into account as it is the source of the bias β as we discussed earlier. Third, the estimate of causal effect based on the observed sample, i.e., $\bar{Y}_t^{ob} - \bar{Y}_c^{ob}$, should be compared with the true causal effect $\mu_t - \mu_c$. To achieve this purpose, I model the distributions of μ_t^{ideal} and μ_c^{ideal} conditional on an imaginary ideal sample, in order to account for the uncertainty brought by an unobserved sample. Consequently, the bias β can be recast as a random variable conditional on $\mathbf{Y}_{\mathbf{t}}^{\mathbf{ideal}}$ and $\mathbf{Y}_{\mathbf{c}}^{\mathbf{ideal}}$:

$$\beta \mid \mathbf{Y}_{\mathbf{t}}^{\mathbf{ideal}}, \mathbf{Y}_{\mathbf{c}}^{\mathbf{ideal}} = \overline{Y}_t^{ob} - \overline{Y}_c^{ob} - (\mu_t^{ideal} - \mu_c^{ideal}) \tag{2.2}$$

It's noteworthy that $\overline{Y}_t^{ob} - \overline{Y}_c^{ob}$ is exactly the estimate of the treatment effect based on the observed sample, i.e., $\hat{\delta}$, and it is unbiased for $\mu_t^{ob} - \mu_c^{ob}$ .Here I treat $\overline{Y}_t^{ob} - \overline{Y}_c^{ob}$ as a fixed constant because the observed sample is fixed. The randomness of μ_t^{ideal} and μ_c^{ideal} is due to random sampling error of an ideal sample because of its imaginary nature.

Comparing to the original definition of bias in (2.1), the new definition of bias (2.2) has two meaningful distinctions: First, the definition (2.2) is a frequentist version of bias which is built on finite samples rather than the whole populations. This permits us to ignore random sampling error of the observed samples and thereby focus on their nonrandom sampling error in the discussion of robustness indices henceforth. Additionally, a distribution of bias conditional on ideal samples is accessible through the definition (2.2) whereby quantifying the robustness indices as probabilities of invalidating an inference is feasible based on it.

The decision rules in (1.5) could be restated conditional on imaginary ideal samples as rules of invalidating an inference as follows:

$$\begin{array}{l} \mu_t^{ideal} - \mu_c^{ideal} < \delta^{\#} \text{ for inferring positive effects} \\ \mu_t^{ideal} - \mu_c^{ideal} > \delta^{\#} \text{ for inferring negative effects} \end{array} \tag{2.3}$$

given $\hat{\delta} = \overline{Y}_t^{ob} - \overline{Y}_c^{ob}$ is fixed and statistically significant. Finally, I propose the robustness indices of causal inferences as probabilities of invalidating an inference as below, according to (2.3):

$$
\begin{aligned}
&P(\mu_t^{ideal} - \mu_c^{ideal} < \delta^{\#}) \text{ for inferring positive effects}\\
&P(\mu_t^{ideal} - \mu_c^{ideal} > \delta^{\#}) \text{ for inferring negative effects}
\end{aligned}
\tag{2.4}
$$

2.2-The Bayesian recipe

The frequentist approach is attractive only if unobserved samples becomes observable. However, this will never happen, which renders the frequentist recipe implausible. An alternative approach is conceptualizing and modeling unobserved samples in the prior distributions by Bayesian reasoning as introduced by Frank & Min (2007). For this purpose, the definition of bias in (2.2) is modified so that it adapts to Bayesian world:

$$\beta \mid \mathbf{Y_t^{ob}}, \mathbf{Y_c^{ob}} = \overline{Y}_t^{ob} - \overline{Y}_c^{ob} - (\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}}) \tag{2.5}$$

The main difference between the Bayesian definition (2.5) and the frequentist definition (2.2) is that the former can and only can depend on the observed sample. It would be illegitimate to think a parameter is conditional on something unobservable like an ideal sample in Bayesian inference. Generically, the Bayesian models which are interpretatively equivalent to the Bayesian framework of Frank & Min (2007) can be formulated as follows:

$$
\begin{aligned}
&\textit{Prior: } \theta \sim F_\theta(\boldsymbol{\eta_0})\\
&\textit{Likelihood: } Y \mid \theta \sim G_Y(\theta)\\
&\textit{Posterior: } \theta \mid Y \sim F_\theta(\boldsymbol{\eta})
\end{aligned}
\tag{2.6}
$$

where $F_\theta(\boldsymbol{\eta_0})$ is the prior distribution of the parameter θ with prior parameters $\boldsymbol{\eta_0}$ and $G_Y(\theta)$ is the likelihood function of the outcome Y with the parameter θ. Hoff (2009) (also see Diaconis & Ylvisaker, 1979, 1985) has shown that when $G_Y(\theta)$ belongs to exponential family and $F_\theta(\boldsymbol{\eta_0})$ is conjugate to $G_Y(\theta)$ (i.e., the posterior distribution $F_\theta(\boldsymbol{\eta})$ and the prior distribution $F_\theta(\boldsymbol{\eta_0})$ are the same distribution with different parametric values) the prior distribution can be interpreted as a distribution built on an unobserved sample whose sample size and sufficient statistics are considered as prior parameters. By construction, any member of exponential family (for example: normal, Poisson, exponential, binomial and multinomial) that has a conjugate prior is appropriate for the Bayesian paradigm of robustness indices of causal inferences and hereafter I only consider the case where likelihood function as well as prior distribution are normal. (Some common distributions that do not belong to exponential family include: T distribution, F distribution, Cauchy, Logistic, mixture models and compounded distributions like beta-binomial and Dirichlet-multinomial distribution).

The construction of Bayesian paradigm of robustness indices begins with the formulation of likelihood functions, which are generally described as the distributions of the treated outcome and the control outcome of ideal population:

$$\begin{aligned} Y_t &\sim N(\mu_t, \sigma_t^2) \\ Y_c &\sim N(\mu_c, \sigma_c^2) \end{aligned} \tag{2.7}$$

The parameters of interest in the likelihood functions (2.7) are μ_t and μ_c, which are defined as the expected value of the treated outcomes of ideal population and the expected value of the control outcomes of ideal population. The variances of both distributions, denoted by σ_t^2 and σ_c^2 , are assumed known. The likelihood functions can be thought of as distributions founded on the observed samples as argued by Frank & Min (2007) even though they are defined for the ideal populations, since practically they are what the real observed data is fitted to.

The Bayesian theory stipulates that the parameters of interest, in this case μ_t and μ_c, should follow some prior distributions and by the logic of Frank & Min (2007) these prior distributions could be conceived as representations of prior knowledge one would learn through unobserved sample. Such prior knowledge is vital and indispensable in modeling the robustness indices of causal inferences as bias is engendered by an unobserved sample. To elaborate on this, it’s imperial to conceptualize an unobserved treated sample whose sample size is n_t and an unobserved control sample whose sample size is n_c. Central limit theorem then suggests the following distributions for μ_t and μ_c conditional on such unobserved treated sample and such unobserved control sample:

$$\begin{aligned} \mu_t &\sim N(\bar{Y}_t^{un}, \frac{\sigma_t^2}{n_t}) \\ \mu_c &\sim N(\bar{Y}_c^{un}, \frac{\sigma_c^2}{n_c}) \end{aligned} \qquad (2.8)$$

The distributions in (2.8) is what I am seeking for prior distributions, that is, prior knowledge which is founded on an unobserved sample. Consolidating the prior distributions (2.8) and the likelihood functions (2.7) gives the complete Bayesian models of robustness indices of causal inferences when the observed sample has N_t subjects in the treatment group and N_c subjects in the control group:

$$\begin{aligned}
&\mu_t \sim N(\bar{Y}_t^{un}, \frac{\sigma_t^2}{n_t}) \\
&Y_t \sim N(\mu_t, \sigma_t^2) \\
&\mu_t \mid \mathbf{Y_t^{ob}} \sim N(\theta_t, \phi_t) \\
&\mu_c \sim N(\bar{Y}_c^{un}, \frac{\sigma_c^2}{n_c}) \\
&Y_c \sim N(\mu_c, \sigma_c^2) \\
&\mu_c \mid \mathbf{Y_c^{ob}} \sim N(\theta_c, \phi_c)
\end{aligned} \tag{2.9}$$

Where:

$$\begin{aligned}
&\theta_t = \frac{n_t}{N_t + n_t}\bar{Y}_t^{un} + \frac{N_t}{N_t + n_t}\bar{Y}_t^{ob} \\
&\phi_t = \frac{\sigma_t^2}{N_t + n_t} \\
&\theta_c = \frac{n_c}{N_c + n_c}\bar{Y}_c^{un} + \frac{N_c}{N_c + n_c}\bar{Y}_c^{ob} \\
&\phi_c = \frac{\sigma_c^2}{N_c + n_c}
\end{aligned} \tag{2.10}$$

To demonstrate the posterior distribution is identical to the distribution upon which the frequentist inference relies, it’s necessary to present the distribution of μ_t^{ideal} and μ_c^{ideal} when an ideal sample is available:

$$\mu_t^{ideal} \sim N\left(\frac{n_t}{N_t+n_t}\bar{Y}_t^{un} + \frac{N_t}{N_t+n_t}\bar{Y}_t^{ob}, \frac{\sigma_t^2}{N_t+n_t}\right)$$
$$\mu_c^{ideal} \sim N\left(\frac{n_c}{N_c+n_c}\bar{Y}_c^{un} + \frac{N_c}{N_c+n_c}\bar{Y}_c^{ob}, \frac{\sigma_c^2}{N_c+n_c}\right) \qquad (2.11)$$

The derivation of distributions in (2.11) is straightforward by central limit theorem given the ideal treated and control sample means are:

$$\bar{Y}_t^{ideal} = \frac{n_t}{N_t+n_t}\bar{Y}_t^{un} + \frac{N_t}{N_t+n_t}\bar{Y}_t^{ob}$$
$$\bar{Y}_c^{ideal} = \frac{n_c}{N_c+n_c}\bar{Y}_c^{un} + \frac{N_c}{N_c+n_c}\bar{Y}_c^{ob} \qquad (2.12)$$

and variances associated with those means are:

$$Var(\bar{Y}_t^{ideal}) = \frac{\sigma_t^2}{N_t+n_t}$$
$$Var(\bar{Y}_c^{ideal}) = \frac{\sigma_c^2}{N_c+n_c} \qquad (2.13)$$

What (2.11) uncovers is that the posterior distribution in the Bayesian recipe (i.e., the distribution of $\mu_t \mid \mathbf{Y_t^{ob}}$ or $\mu_c \mid \mathbf{Y_c^{ob}}$) and the distribution of parameter built on an ideal sample (i.e., the distribution of μ_t^{ideal} or μ_c^{ideal}) in the frequentist recipe are identical when a normal likelihood function with the mean as the only parameter and normal prior are considered in the Bayesian paradigm. However, I caution readers that this result will remain valid only for a certain type of likelihood and prior (exponential family with conjugate prior) and in this case the Bayesian recipe and the frequentist recipe are still distinct in many aspects.

If the independence between the treated outcome Y_t and the control outcome Y_c is posited, the distributions of $\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}}$ and β have explicit forms:

$$\begin{aligned} &\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}} \sim N(\theta_t - \theta_c, \phi_t + \phi_c) \\ &\beta \mid \mathbf{Y_t^{ob}}, \mathbf{Y_c^{ob}} \sim N(\bar{Y}_t^{ob} - \bar{Y}_c^{ob} - (\theta_t - \theta_c), \phi_t + \phi_c) \end{aligned} \tag{2.14}$$

with $\theta_t, \theta_c, \phi_t, \phi_c$ quantified as in (2.10). An inference is invalidated if one of the following conditions are true:

$$\begin{aligned} &\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}} < \delta^{\#} \text{ for inferring positive effects} \\ &\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}} > \delta^{\#} \text{ for inferring negative effects} \end{aligned} \tag{2.15}$$

Capitalize on the distribution of $\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}}$, the probability of invalidating an inference is defined as follows:

$$\begin{aligned} &P(\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}} < \delta^{\#}) \text{ for inferring positive effects} \\ &P(\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}} > \delta^{\#}) \text{ for inferring negative effects} \end{aligned} \tag{2.16}$$

Given the threshold of making an inference and the values of the parameters θ_t , θ_c , ϕ_t and ϕ_c, this probability could be directly calculated as the function of those parameters and employed as the measurement of the robustness for any single study. Furthermore, one can calculate the probabilities of invalidating the inference for different but parallel studies and compare their robustness in terms of those probabilities.

I caution readers here that the probability of invalidating an inference should not be confused with the p-value in hypothesis testing. Unfortunately, the overwhelming misinterpretations of p-value often make researchers treat those two distinct indices as parallel ones even though they are in fact telling completely different stories. A particularly relevant misinterpretation in the context is perceiving p-value as one indicator of the robustness or replicability of an inference, and this misinterpretation is scientifically detrimental and blurs the boundary between p-value and true robustness indices.

It’s worthy to emphasize that p-value can never become an index of the robustness of any inferences for mainly two reasons. First, p-value only deals with random sampling error and it evaluates the degree to which a similar finding will occur in another equivalent random sample drawn by repeated random sampling. It largely quantifies the significance of a result when random sampling error is the only concern. Nonetheless, random sampling error has never been the focus of the analysis of robustness since it virtually exists in every study and every inference. Quite the opposite, robustness indices usually highlight the errors due to sources other than random sampling such as nonrandom sampling, nonrandom assignment and omission of important confounding variables. The probability of invalidating an inference is an index of robustness because it takes either nonrandom sampling error or nonrandom assignment error into account by considering prior distributions as ones built on unobserved samples.

Equally importantly, p-value is unqualified for a measurement of robustness because it is only valid when the null hypothesis is true. In contrast, the robustness indices invented by Frank et al. (2013) and the probability of invalidating an inference are useful regardless of the condition specified in null hypothesis. For example, Frank et al. (2013) mentioned that one can change the null hypothesis and compute the corresponding robustness indices by modifying the threshold value accordingly. The same thing can be done in computing the posterior probability of invalidating an inference as it depends on not only the posterior distribution of average treatment effect but also the threshold. Testing null hypotheses of nonzero values can always be achieved by adjusting the threshold $\delta^{\#}$.

The construction of the probability of invalidating an inference is built on three assumptions. First, the random sampling error associated with the observed sample is ignored in the Bayesian models so that researchers should be aware that this probability can only indicate how likely an inference will be invalidated due to bias induced by either nonrandom sampling or nonrandom assignment. Second, the distributions of the treated outcome and the control outcome are assumed to be normal. Third, the treated and the control outcome are assumed to be independent. In summary, the Bayesian recipe exemplifies the Bayesian framework raised in Frank & Min (2007). A prior distribution whose definition is the distribution carries the information of one's belief about the parameters prior to observing the data, could be conceptualized as the distribution of a focal parameter conditional on an unobserved sample since it exactly reflects the belief about the inferred parameter and is solely motivated and shaped by such belief. Neither a distribution based on an unobserved sample nor a typical prior distribution in a Bayesian context contains any information about the observed sample. The likelihood function in the Bayesian models, which serves as a generic characterization of the ideal population, is in fact completely driven by the observed sample. Furthermore, the problem of checking the robustness of the inference by varying the mean and sample size of an unobserved sample is transformed into a problem of checking the influence of a prior on its corresponding posterior distribution while holding the observed sample and the likelihood function fixed.

3-The Bayesian models of robustness indices for internal validity

The unifying Bayesian framework of robustness indices of causal inferences can be recast as the Bayesian models of robustness indices particularly for internal validity, by deliberately define the observed (treated/control) sample and an unobserved (treated/control) sample in the following way:

$$\begin{aligned}
\mathbf{Y_t^{ob}} &= \left\{ Y_{t,i}^{ob} : i \in T \right\} \\
\mathbf{Y_t^{un}} &= \left\{ Y_{t,j}^{un} : j \in C \right\} \\
\mathbf{Y_c^{ob}} &= \left\{ Y_{c,j}^{ob} : j \in C \right\} \\
\mathbf{Y_c^{un}} &= \left\{ Y_{c,i}^{un} : i \in T \right\}
\end{aligned} \tag{3.1}$$

Definition (3.1) shows that the observed treated sample for the studies with questionable internal validity will be real outcomes of the subjects who indeed received the treatment. An unobserved treated sample in this case will be counterfactual outcomes of the subjects in the control group had they been assigned to the treatment group instead. Likewise, the observed control sample for the studies with questionable internal validity will be real outcomes of the subjects who actually received the control and an unobserved control sample for the same studies will be counterfactual outcomes of the subjects in the treatment group had they switched to the control group. Obviously, definition (3.1) just mathematically restates the definition 1.6.1 which formalizes the concepts of unobserved and observed sample in the first scenario where internal validity is limited. For example, to conceptualize an unobserved treated sample in the study of Hong & Raudenbush (2005), we need to ask a question like “what if a promoted child did not get promoted in the first place” and how it can affect his test score. Similarly, an unobserved control sample would answer a question like “what would the academic achievement of a retained student be if he had been promoted”.

Aside from this definition, everything else of the Bayesian models of robustness indices for internal validity will remain the same as the unifying Bayesian framework. Draw on (3.1), this model has the identical definition of bias as in (2.5), identical Bayesian formulations as in (2.9) and (2.10), together with the identical distributions of $\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}}$ and β as in (2.14).

Moreover, as discussed earlier, the sample sizes of unobserved treat and control sample are fixed for the case of internal validity. The sample sizes of an unobserved treated sample and the observed control sample should be equal, and the sample sizes of an unobserved control sample and the observed treated sample should be equal as well. To impose the aforementioned restrictions on the models (2.9) and (2.10), a new set of models are proposed next with one additional parameter π defined as the proportion of subjects who get the treatment in the whole sample:

$$\begin{aligned}
&\mu_t \sim N(\bar{Y}_t^{un}, \frac{\sigma_t^2}{n_t}) \\
&Y_t \sim N(\mu_t, \sigma_t^2) \\
&\mu_t \mid \mathbf{Y_t^{ob}} \sim N(\theta_t, \phi_t) \\
&\mu_c \sim N(\bar{Y}_c^{un}, \frac{\sigma_c^2}{n_c}) \\
&Y_c \sim N(\mu_c, \sigma_c^2) \\
&\mu_c \mid \mathbf{Y_c^{ob}} \sim N(\theta_c, \phi_c)
\end{aligned} \tag{3.2}$$

Where:

$$\begin{aligned}
n_t &= \left(\frac{1-\pi}{\pi}\right) N_t = N_c \\
n_c &= \left(\frac{\pi}{1-\pi}\right) N_c = N_t
\end{aligned} \tag{3.3}$$

And:

$$\begin{aligned}
\theta_t &= (1-\pi)\bar{Y}_t^{un} + \pi\bar{Y}_t^{ob} \\
\phi_t &= \pi\frac{\sigma_t^2}{N_t} = \frac{\sigma_t^2}{N} \\
\theta_c &= \pi\bar{Y}_c^{un} + (1-\pi)\bar{Y}_c^{ob} \\
\phi_c &= (1-\pi)\frac{\sigma_c^2}{N_c} = \frac{\sigma_c^2}{N}
\end{aligned} \tag{3.4}$$

In the formula above, N is the total observed sample size, i.e., $N = N_t + N_c$ .The denominators in the second and fourth equations in (3.4) become N simply because π by definition is the ratio between N_t and N . Given a designated threshold $\delta^{\#}$ and a chosen decision rule in (2.15), the posterior distribution in (2.14) will naturally generate the probability of invalidating an inference due to limited internal validity, as a function of the parameters in (3.4). It's imperative to keep in mind that this probability is built on three assumptions, namely the assumption of no random sampling error for the observed sample, the normality assumption for the distributions of treated and control outcome and the assumption of independence between treated outcome and control outcome.

By introducing a new parameter α as the ratio between $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$, the relationship between the probability of invalidating an inference due to inadequate internal validity and the parameters mentioned in (3.2) through (3.4) is proven to be a probit function in the following form, for a targeted negative effect:

$$probit(p) = \frac{\sqrt{N}}{\sqrt{\sigma_t^2 + \sigma_c^2}}\left[\bar{Y}_c^{un}(1-\pi)\cdot\alpha + (\bar{Y}_t^{ob} + \bar{Y}_c^{ob} - \bar{Y}_c^{un})\cdot\pi - \bar{Y}_c^{ob} - \delta^{\#}\right] \tag{3.5}$$

For a targeted positive effect, we just need to reverse the signs of the coefficient of α and constant presented in (3.5), which leads to the equation below:

$$probit(p)=\frac{\sqrt{N}}{\sqrt{\sigma_t^2+\sigma_c^2}}\left[\bar{Y}_c^{un}(\pi-1)\cdot\alpha-(\bar{Y}_t^{ob}+\bar{Y}_c^{ob}-\bar{Y}_c^{un})\cdot\pi+\bar{Y}_c^{ob}+\delta^{\#}\right] \tag{3.6}$$

p in (3.5) (or (3.6)) symbolizes the probability of invalidating an inference, which is computed as the probability that $\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}}$ is larger (or smaller, depends on the sign of inferred effect) than a threshold $\delta^{\#}$. This probability should be straightforward as we have learned that the distribution of $\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}}$ is normal with mean $\theta_t - \theta_c$ and variance $\phi_t + \phi_c$ described in (3.4).

What (3.5) and (3.6) demonstrate is that the probit link function of the probability of invalidating an inference due to limited internal validity is a linear function of α. Therefore, given the values of $N, \sigma_t^2, \sigma_c^2, \bar{Y}_c^{un}, \bar{Y}_t^{ob}, \bar{Y}_c^{ob}, \pi$ and the threshold $\delta^{\#}$, the probit link function of the probability of invalidating an inference due to limited internal validity can be explicitly expressed as a linear function of α. I will draw on this feature to elicit answers of some very meaningful questions, such as finding out the how large/small α could be conditional on a set of value of parameters $N, \sigma_t^2, \sigma_c^2, \bar{Y}_c^{un}, \bar{Y}_t^{ob}, \bar{Y}_c^{ob}, \pi, \delta^{\#}$ that makes the probability of invalidating an inference smaller than a certain value (for example, 0.3). Normally, one would extract $\bar{Y}_t^{ob}, \bar{Y}_c^{ob}, N$ from the observed sample and select some fixed constants for σ_t^2, σ_c^2. $\bar{Y}_c^{un}$, the mean of an unobserved control sample, is conceptualized as a number which are not necessarily fixed in this approach. Together with the variable α, $\bar{Y}_c^{un}$ characterizes unobserved treated and control sample which of paramount concern in my Bayesian models.

The Bayesian models (3.2)-(3.4) can be recast as Rubin Causal Model (RCM). Suppose in one observational study there are N subjects in total. Moreover, there are N_t participants in the treatment group and N_c participants in the control group. In other words, we have N_t observed treated outcomes and N_c observed control outcomes. According to RCM, every participant in the treatment group would have had a counterfactual outcome if he had been assigned to the control group. Likewise, every participant in the control group would have had a counterfactual outcome if he had been assigned to the treatment group. This means there should be N_t unobserved control outcomes and N_c unobserved treated outcomes in total. In the Bayesian models of robustness indices for internal validity, the ideal treated sample could be thought to be consisted of the N_t observed treated outcomes as the observed treated sample and N_c unobserved treated outcomes as an unobserved treated sample. Similarly, the ideal control sample could be perceived as a composition of the N_c observed control outcomes as the observed control sample and N_t unobserved control outcomes as an unobserved control sample. To summarize the Bayesian models of robustness indices for internal validity that are presented in this section, it's necessary now to review the perspective of Rubin Causal Model (RCM) associated with them. The Rubin Causal Model conceptualizes the observational studies as a missing data problem and the assignment mechanism as the mechanism of how the missing data is generated. Specifically, follow the logic of the Rubin Causal Model (RCM), every individual has one observed outcome and one missing outcome.

4-The Bayesian models of robustness indices for external validity

The Bayesian models of robustness indices for external validity, just like the Bayesian models of robustness indices for internal validity, is a descendant of the unifying Bayesian framework. There are two key differences between the models for external validity and the models for internal validity. The first key difference is that the models for external validity and internal validity have distinct definitions regarding an unobserved (treated/control) sample and the observed (treated/control) sample. For research whose major concern is external validity, the observed (treated/control) sample and an unobserved (treated/control) sample are defined as follows:

$$\begin{aligned}\mathbf{Y_t^{ob}} &= \left\{Y_{t,i}^{ob} : i \in R\right\} \\ \mathbf{Y_t^{un}} &= \left\{Y_{t,k}^{un} : k \in R^{'}\right\} \\ \mathbf{Y_c^{ob}} &= \left\{Y_{c,j}^{ob} : j \in R\right\} \\ \mathbf{Y_c^{un}} &= \left\{Y_{c,l}^{un} : l \in R^{'}\right\}\end{aligned} \tag{4.1}$$

Definition (4.1) is just the mathematical equivalent of the definitions 1.6.2 that formalizes the concepts of unobserved sample and observed sample in the second scenario where research has limited external validity. Here I use R to denote the representable part of ideal population and $R^{'}$ to denote the non-representable part of ideal population. A pivotal difference between the models for external validity and internal validity is that one need a new parameter π_R to operationalize the definition in (4.1). π_R represents the proportion of the representable part of ideal population R in the whole ideal population to which an inference is intended to generalize. To quantify the parameter π_R one need judicious conceptualizations of the size of R relative to its corresponding ideal population. For example, π_R would be the proportion of volunteered schools (and arguably schools which are similar to volunteered schools) in Borman et al. (2008) in the population of all U.S. schools.

By this logic, the expectations of ideal treated/control population (i.e., $\mathrm{E}[Y_t]$ and $E[Y_c]$) can be rewritten as functions of π_R as below:

$$\begin{aligned}\mathrm{E}[Y_t] &= \mathrm{E}[\bar{Y}_t^{ob}]\pi_R + (1-\pi_R)\,\mathrm{E}[\bar{Y}_t^{un}] \\ \mathrm{E}[Y_c] &= \mathrm{E}[\bar{Y}_c^{ob}]\pi_R + (1-\pi_R)\mathrm{E}[\bar{Y}_c^{un}]\end{aligned} \tag{4.2}$$

Recall that the ideal treated and control sample means have been presented in (2.11) and their expected values need to match the expectations listed in (4.2), which leads to the following equations:

$$\begin{aligned}\bar{Y}_t^{ideal} &= \frac{n_t}{N_t+n_t}\bar{Y}_t^{un} + \frac{N_t}{N_t+n_t}\bar{Y}_t^{ob} = \bar{Y}_t^{ob}\pi_R + (1-\pi_R)\bar{Y}_t^{un} \\ \bar{Y}_c^{ideal} &= \frac{n_c}{N_c+n_c}\bar{Y}_c^{un} + \frac{N_c}{N_c+n_c}\bar{Y}_c^{ob} = \bar{Y}_c^{ob}\pi_R + (1-\pi_R)\bar{Y}_c^{un}\end{aligned} \tag{4.3}$$

Equation (4.3) reveals the following constraints for the unobserved sample sizes:

$$\begin{aligned}n_t &= \left(\frac{1-\pi_R}{\pi_R}\right)N_t \\ n_c &= \left(\frac{1-\pi_R}{\pi_R}\right)N_c\end{aligned} \tag{4.4}$$

An appropriate conceptualization of (4.1) through (4.4) would be envisaging that an unobserved sample of subjects is randomly drawn from the non-representable part of ideal population and subsequently a random assignment which results in the same proportion of treated subjects as in the observed sample is carried out for this unobserved sample. The treated outcomes of those treated subjects in this unobserved sample are therefore grouped as an unobserved treated sample, and the control outcomes of the remaining subjects in this unobserved sample (that is, people who receive control) will form an unobserved control sample. I warn readers about the difference in the formation of unobserved treated/control sample between the scenarios of internal validity and external validity.

Now I construct the Bayesian models of robustness indices for external validity, by utilizing the likelihood function listed in (2.7) and the prior distribution advanced in (2.8). This too will yield the same form as the Bayesian formulation suggested in (2.9) and (2.10), as we have already seen in the previous section. The Bayesian models below again rely on the assumptions of no random sampling error for the observed samples, normality, and independence between treated and control outcomes:

$$\begin{aligned}
&\mu_t \sim N(\bar{Y}_t^{un}, \frac{\sigma_t^2}{n_t}) \\
&Y_t \sim N(\mu_t, \sigma_t^2) \\
&\mu_t \mid \mathbf{Y_t^{ob}} \sim N(\theta_t, \phi_t) \\
&\mu_c \sim N(\bar{Y}_c^{un}, \frac{\sigma_c^2}{n_c}) \\
&Y_c \sim N(\mu_c, \sigma_c^2) \\
&\mu_c \mid \mathbf{Y_c^{ob}} \sim N(\theta_c, \phi_c)
\end{aligned} \tag{4.5}$$

Where:

$$\begin{aligned}
&\theta_t = (1-\pi_R)\bar{Y}_t^{un} + \pi_R \bar{Y}_t^{ob} \\
&\phi_t = \pi_R \frac{\sigma_t^2}{N_t} \\
&\theta_c = (1-\pi_R)\bar{Y}_c^{un} + \pi_R \bar{Y}_c^{ob} \\
&\phi_c = \pi_R \frac{\sigma_c^2}{N_c}
\end{aligned} \tag{4.6}$$

As always, the Bayesian models in (4.5) and (4.6) will effectuate the posterior distributions displayed in (2.14) and the probability of invalidating an inference due to limited external validity as a function of the parameters presented in (4.6), once a threshold and a preselected decision rule are set up.

Again by defining α as the ratio between $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$, the probit link function of the probability of invalidating an inference due to limited external validity is shown to be a nonlinear function of α and π_R, depending on the signs of the focal treatment effect. When inferring a negative effect, the probit model is:

$$probit(p)=\frac{1}{\sqrt{\frac{\sigma_t^2}{N_t}+\frac{\sigma_c^2}{N_c}}}\left[\bar{Y}_c^{un}\cdot\alpha\pi_R^{-\frac{1}{2}}+(\bar{Y}_t^{ob}-\bar{Y}_c^{ob}+\bar{Y}_c^{un})\cdot\pi_R^{\frac{1}{2}}-\bar{Y}_c^{un}\cdot\alpha\pi_R^{\frac{1}{2}}-(\bar{Y}_c^{un}+\delta^{\#})\cdot\pi_R^{-\frac{1}{2}}\right]$$

(4.7)

And when inferring a positive effect, the probit model becomes:

$$probit(p)=\frac{1}{\sqrt{\frac{\sigma_t^2}{N_t}+\frac{\sigma_c^2}{N_c}}}\left[\bar{Y}_c^{un}\cdot\alpha\pi_R^{\frac{1}{2}}-\bar{Y}_c^{un}\cdot\alpha\pi_R^{-\frac{1}{2}}-(\bar{Y}_t^{ob}-\bar{Y}_c^{ob}+\bar{Y}_c^{un})\cdot\pi_R^{\frac{1}{2}}+(\bar{Y}_c^{un}+\delta^{\#})\cdot\pi_R^{-\frac{1}{2}}\right]$$

(4.8)

The probit models in (4.7) and (4.8) share the same feature and notations as their counterparts in the case of internal validity. For example, p in (4.7) and (4.8) denotes the probability of invalidating an inference, which is simply calculated based on the distribution of $\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}}$ with mean $\theta_t - \theta_c$ and variance $\phi_t + \phi_c$ listed in (4.6) as the probability of $\mu_t \mid Y_t^{ob} - \mu_c \mid Y_c^{ob}$is larger than a threshold $\delta^{\#}$. Typically, σ_t^2, σ_c^2 are predetermined and

$N_t, N_c, \bar{Y}_t^{ob}, \bar{Y}_c^{ob}$ are information contained in the observed sample. Most importantly, I distinguish and summarize unobserved treated and control sample with $\bar{Y}_c^{ob}$ and α.

The probit models (4.7) and (4.8) can be employed to answer questions like "how large/small α has to be conditional on a value of π_R such that the probability of invalidating an inference is smaller than 0.3?" or "how large/small π_R has to be conditional on a value of α such that the probability of invalidating an inference is smaller than 0.2?", as soon as the values of parameters $N_t, N_c, \sigma_t^2, \sigma_c^2, \bar{Y}_c^{un}, \bar{Y}_t^{ob}, \bar{Y}_c^{ob}$ and the threshold $\delta^{\#}$ are chosen by the researcher.

From a sampling perspective, the Bayesian models of robustness indices for external validity is tantamount to the following sampling process: The observed sample is first drawn from the representable part of ideal population and fixed henceforth. Then an unobserved sample is thought to be drawn from the non-representable part of ideal population and it is not necessarily to be fixed. The unobserved sample size, i.e., $n = n_t + n_c$ is determined by the observed sample size $N = N_t + N_c$ and π_R, i.e., $n = \frac{1-\pi_R}{\pi_R} N$. All subjects in this unobserved sample will be then randomly assigned to a treatment group or control group, and I do maintain that the proportion of treated subjects in this unobserved sample will be equal to the proportion of treated subjects in the observed sample, that is, $\frac{n_t}{n} = \frac{N_t}{N}$.

I again emphasize the difference between the Bayesian models concerning the internal validity and external validity is that they address different central questions. For the internal validity, the unobserved part of ideal population is the collection of counterfactuals brought by the assignment mechanism. In this case, one does not seek to generalize his inference to other populations of subjects that are not accessible. For example, the data from Early Childhood

Longitudinal Study Kindergarten (ECLS-K) used by Hong & Raudenbush (2005) is nationally representative and the paramount concern of this study is the lack of random assignment of kindergarten students to the conditions of being retained or promoted, in this case the Bayesian models of the robustness indices for the internal validity will be quite appropriate to employ. Nonetheless, for the external validity, the unobserved (non-representable) part of ideal population is not the counterfactuals, which though exist but does not affect the inference. Rather, it is occasioned by the overgeneralization, that is, the researchers attempt to generalize their conclusions beyond the populations they have sampled from. For an instance, Borman et al. (2008) conducted a cluster randomized trial to examine the efficacy of OCR curriculum in the six schools they randomly sampled from the schools that volunteered in this curriculum and the results pertaining to this experiment is intended for students across the whole nation. The unobserved (non-representable) part of ideal population for Borman et al. (2008) would be the schools in the U.S. which did not volunteer in this research. The Bayesian models of the robustness indices for the external validity takes this unobserved (non-representable) part of ideal population into consideration and modify the inference accordingly.

In spite of the important distinctions between those two classes of Bayesian models I just discussed, it is pivotal to appreciate the commonness shared by the both sets of Bayesian models when learning and utilizing them. First, the definition of the bias and the rules of judging the inference invalid are the same for both sets of models, and they are the starting points of the construction of the both kinds of Bayesian models as they form the base of calculating the probability of invalidating an inference. Second, both sets of models share the same model structure and the same group of parameters. Specifically, the distribution of μ_t based on an unobserved treated sample and the distribution of μ_c based on an unobserved control sample are the priors and the generic descriptions of treated outcome Y_t and control outcome Y_c are the likelihood functions in both kinds of models. The parameters in the Bayesian models include unobserved treated and control sample means, observed treated and control sample means, and the variances of Y_t and Y_c . In addition, one parameter symbolizing the relative size of an unobserved sample in an ideal sample will be needed and its definition does depend on the context of whether internal validity or external validity is the focus. Third, the interpretations of those two classes of Bayesian models are in nearly the same fashion. That is, we conceptualize one unobserved samples is randomly drawn from the unobserved part of ideal population, and this unobserved sample is then integrated with the observed sample to form an ideal sample. The ideal treated (control) sample mean are just the weighted average of unobserved treated (control) sample mean and the observed treated (control) sample mean, where the weights are just the proportions of these samples in an ideal sample. Therefore, the Bayesian models of robustness indices assume one can augment the observed sample with an unobserved sample and update the inference over this augmented sample.

5-Statistical threshold and Bayesian models for replacing observed cases

An empirical researcher who tries to decide whether an inference is invalidated based on his observed sample and chosen statistical threshold $\delta^{\#}$ could be easily entrapped in an inferential pitfall about Bayesian models of robustness indices. This occurs when one compute the statistical threshold with the variance of average treatment effect estimate based on the observed sample instead of with the variance of average treatment effect estimate based on an ideal sample, and is unaware of the key difference between those two types of variances. In fact, the Bayesian models of robustness indices of causal inferences I have discussed so far assume one obtain an unobserved sample and incorporate this unobserved sample into his observed sample to form an ideal sample. Given the standard deviation of average treatment effect estimate computed from an ideal sample has taken the sample sizes and observations of both unobserved and observed sample into consideration, it becomes more appropriate than its counterpart extracted from the observed sample in quantifying a statistical threshold.

There are two ways of addressing this issue: The first way is to calculate the statistical threshold $\delta^{\#}$ as a product of the chosen critical value of standard normal distribution (by convention it is 1.96) and the standard error of this average treatment effect estimate based on an ideal sample. The second approach redefines an ideal sample as a sample furnished by replacing a proportion of observed cases with an unobserved sample so as to keep the ideal sample size identical to the observed sample size. Consequently, this approach requires an utter shift in sampling perspective from adding unobserved cases (to the observed sample) to replacing a part of the observed sample with unobserved cases. Such shift in sampling perspective further necessitates some modifications of the Bayesian framework.

5.1-Appropriate statistical threshold $\delta^{\#}$ for Bayesian models of robustness indices

Identifying the ideal sample variance of the average treatment effect estimate is a prerequisite for the calculation of appropriate statistical threshold for Bayesian models of robustness indices. Recall that in (2.14) I have presented the distribution of $\mu_t \mid \mathbf{Y_t^{ob}} - \mu_c \mid \mathbf{Y_c^{ob}}$ which is equivalent to the distribution of average treatment effect estimate based on an ideal sample, and therefore the ideal sample variance of the average treatment effect estimate is informed by this distribution as $\phi_{\mathrm{t}} + \phi_{\mathrm{c}}$.

For the Bayesian models of robustness indices for internal validity, $\phi_{\mathrm{t}} + \phi_{\mathrm{c}}$ equals:

$$\phi_t + \phi_c = \frac{\sigma_t^2 + \sigma_c^2}{N} \tag{5.1}$$

For the Bayesian models of robustness indices for external validity, $\phi_\mathrm{t}+\phi_c$ becomes:

$$\phi_t+\phi_c=\pi_R\left(\frac{\sigma_t^2}{N_t}+\frac{\sigma_c^2}{N_c}\right) \tag{5.2}$$

Draw on (5.1), the appropriate statistical threshold $\delta^{\#}$ for the Bayesian models of robustness indices for internal validity should be computed as follows:

$$\begin{aligned}\delta^{\#}&=1.96*\sqrt{\frac{\sigma_t^2+\sigma_c^2}{N}}\text{ for inferring a positive effect}\\ \delta^{\#}&=-1.96*\sqrt{\frac{\sigma_t^2+\sigma_c^2}{N}}\text{ for inferring a negative effect}\end{aligned} \tag{5.3}$$

Likewise, the following statistical threshold $\delta^{\#}$ is recommended for the Bayesian models of robustness indices for external validity:

$$\begin{aligned}\delta^{\#}&=1.96*\sqrt{\pi_R\left(\frac{\sigma_t^2}{N_t}+\frac{\sigma_c^2}{N_c}\right)}\text{ for inferring a positive effect}\\ \delta^{\#}&=-1.96*\sqrt{\pi_R\left(\frac{\sigma_t^2}{N_t}+\frac{\sigma_c^2}{N_c}\right)}\text{ for inferring a negative effect}\end{aligned} \tag{5.4}$$

Based on (5.3), the probit models for the probability of invalidating an inference due to limited internal validity are rewritten as follows:

For inferring a negative effect:

$$probit(p)=\frac{\sqrt{N}}{\sqrt{\sigma_t^2+\sigma_c^2}}\left[\bar{Y}_c^{un}(1-\pi)\cdot\alpha+(\bar{Y}_t^{ob}+\bar{Y}_c^{ob}-\bar{Y}_c^{un})\cdot\pi-\bar{Y}_c^{ob}\right]+1.96 \tag{5.5}$$

For inferring a positive effect:

$$probit(p)=\frac{\sqrt{N}}{\sqrt{\sigma_t^2+\sigma_c^2}}\left[\bar{Y}_c^{un}(\pi-1)\cdot\alpha-(\bar{Y}_t^{ob}+\bar{Y}_c^{ob}-\bar{Y}_c^{un})\cdot\pi+\bar{Y}_c^{ob}\right]+1.96 \quad (5.6)$$

Similarly, plugging the appropriate statistical thresholds in (5.4) will update the probit models for the probability of invalidating an inference due to limited external validity as below:

For inferring a negative effect:

$$probit(p)=\frac{1}{\sqrt{\frac{\sigma_t^2}{N_t}+\frac{\sigma_c^2}{N_c}}}\left[\bar{Y}_c^{un}\cdot\alpha\pi_R^{-\frac{1}{2}}+(\bar{Y}_t^{ob}-\bar{Y}_c^{ob}+\bar{Y}_c^{un})\cdot\pi_R^{\frac{1}{2}}-\bar{Y}_c^{un}\cdot\alpha\pi_R^{\frac{1}{2}}-\bar{Y}_c^{un}\pi_R^{-\frac{1}{2}}\right]+1.96$$

(5.7)

For inferring a positive effect:

$$probit(p)=\frac{1}{\sqrt{\frac{\sigma_t^2}{N_t}+\frac{\sigma_c^2}{N_c}}}\left[\bar{Y}_c^{un}\cdot\alpha\pi_R^{\frac{1}{2}}-\bar{Y}_c^{un}\cdot\alpha\pi_R^{-\frac{1}{2}}-(\bar{Y}_t^{ob}-\bar{Y}_c^{ob}+\bar{Y}_c^{un})\cdot\pi_R^{\frac{1}{2}}+\bar{Y}_c^{un}\cdot\pi_R^{-\frac{1}{2}}\right]+1.96$$

(5.8)

I do recognize that the threshold $\delta^{\#}$ could be a non-statistical one rather than a statistical one, as typically empirical researchers would set the threshold through a multifaceted and pragmatic decision-making process. The threshold $\delta^{\#}$ tends to be non-statistical when, for example, a benchmark in effect size is available through literature review or research synthesis. Therefore, the formulae of (5.1) through (5.4) can only serve as the guidelines of determining the threshold $\delta^{\#}$ based solely on statistical significance. A more general guidance has been offered by Frank et al. (2013) to shed a light on choosing a threshold based on the transaction costs of proposed actions.

5.2-The Bayesian models of robustness indices for replacing observed cases

Up to now we have delved into the Bayesian models of robustness indices where an unobserved sample is modeled by a prior distribution and correspondingly an ideal sample is formed and represented by the posterior distribution. Such Bayesian models are tantamount to a sampling procedure where one first obtains an observed sample and then adds an unobserved sample to this observed sample to construct an ideal sample. However, I point out that adding unobserved cases is not the only way of generating an ideal sample considering an ideal sample can also be shaped by replacing a proportion of the observed sample with an unobserved sample, as proposed by Frank & Min (2007).

To articulate Bayesian models concerning replacing a part of the observed sample with an unobserved sample, I introduce following notations for the sampling scheme of replacing observed cases:

For an individual who joined the treatment group, I_i^t is an indicator of whether he is retained in an ideal sample (and thus he is not replaced with an unobserved case). Therefore, when $I_i^t = 1$ this individual i (say his name is Tom) is not replaced with an unobserved case and when $I_i^t = 0$ Tom belongs to the part of the observed sample which is to be replaced with an unobserved sample. Likewise, I_j^c is a binary indicator of whether an individual j (say her name is Ashley) who participates in the control group (symbolized by the superscript ‘c’) is remained in an ideal sample (so she is not replaced with an unobserved case either).

Next, I define s_t as an ideal treated sample and s_c as an ideal control sample. Operationally, s_t can be represented by a collection of $I_i^t,\ i = 1, 2, \ldots, N_t$ and s_c can be represented by a collection of $I_j^c,\ j=1,2,\ldots,N_c$, as the collections of I_i^t and I_j^c would inform us which observations are kept in an ideal sample and which ones are to be replaced with an unobserved sample.

Finally, I define π_r as the proportion of cases to be replaced with an unobserved sample in the observed sample and thus unobserved sample size becomes the product between observed sample size and π_r:

$$\begin{aligned} n_t &= \pi_r N_t \\ n_c &= \pi_r N_c \end{aligned} \tag{5.9}$$

Upon the sampling outlook and definitions, the Bayesian models of robustness indices of causal inferences for replacing observed cases is formalized here:

$$\begin{aligned} &\mu_t \sim N(\bar{Y}_t^{un}, \frac{\sigma_t^2}{n_t}) \\ &Y_{t\cdot i} \mid I_i^t = 1 \sim N(\mu_t, \sigma_t^2) \\ &\mu_t \mid \mathbf{Y_t^{ob}}, s_t \sim N(\theta_t^r, \phi_t^r) \\ &\mu_c \sim N(\bar{Y}_c^{un}, \frac{\sigma_c^2}{n_c}) \\ &Y_{c\cdot j} \mid I_j^c = 1 \sim N(\mu_c, \sigma_c^2) \\ &\mu_c \mid \mathbf{Y_c^{ob}}, s_c \sim N(\theta_c^r, \phi_c^r) \end{aligned} \tag{5.10}$$

where:

$$\begin{aligned} s_t &= [I_1^t, I_2^t, \ldots, I_{N_t}^t] \\ s_c &= [I_1^c, I_2^c, \ldots, I_{N_c}^c] \end{aligned} \tag{5.11}$$

and:

$$
\begin{aligned}
\theta_t^r &= \pi_r \bar{Y}_t^{un} + \frac{\sum_{i=1}^{N_t} I_i^t Y_{t\cdot i}}{N_t} \\
\phi_t^r &= \frac{\sigma_t^2}{N_t} \\
\theta_c^r &= \pi_r \bar{Y}_c^{un} + \frac{\sum_{j=1}^{N_c} I_j^c Y_{c\cdot j}}{N_c} \\
\phi_c^r &= \frac{\sigma_c^2}{N_c}
\end{aligned}
\tag{5.12}
$$

Precaution is needed when one chooses the above Bayesian models since the posterior distribution is conditional not only on the observed sample but also on which observations are kept and which ones are exchanged with unobserved cases in an ideal sample. To derive a posterior distribution which depends solely on the observed sample, the expectations of the posterior distributions $\mu_t \mid \mathbf{Y_t^{ob}}, s_t$ and $\mu_c \mid \mathbf{Y_c^{ob}}, s_c$ can be computed over the distributions of s_t and s_c respectively, which results in the following posterior distributions:

$$
\begin{aligned}
\mu_t \mid \mathbf{Y_t^{ob}} &= E_{s_t}[\mu_t \mid \mathbf{Y_t^{ob}}, s_t] \sim N(\theta_t^r, \phi_t^r) \\
\mu_c \mid \mathbf{Y}_c^{\mathbf{ob}} &= E_{s_c}[\mu_c \mid \mathbf{Y}_c^{\mathbf{ob}}, s_c] \sim N(\theta_c^r, \phi_c^r)
\end{aligned}
\tag{5.13}
$$

where the posterior means and variances are now become:

$$
\begin{aligned}
\theta_t^r &= \pi_r \bar{Y}_t^{un} + (1-\pi_r)\bar{Y}_t^{ob} \\
\phi_t^r &= \frac{\sigma_t^2}{N_t} \\
\theta_c^r &= \pi_r \bar{Y}_c^{un} + (1-\pi_r)\bar{Y}_c^{ob} \\
\phi_c^r &= \frac{\sigma_c^2}{N_c}
\end{aligned}
\tag{5.14}
$$

Guided by the posterior distributions in (5.13) and (5.14), the probability of invalidating an inference is readily accessible through (2.16).

6-Demonstrative examples

6.1-The Bayesian robustness indices of the effect of OCR on reading achievement

According to Borman et al. (2008), the Open Court Reading (OCR) curriculum "has been widely used since 1960s and offers a phonics-based K-6 curriculum that is grounded in the research-based practices cited in the National Reading Panel report (National Reading Panel, 2000)." Therefore, Borman et al. (2008) argued that the OCR program had a potential to enhance instructional quality and thus reading achievement as it was rooted in research-based practices that had been advanced by federal educational programs like Reading First and No Child Left Behind. To arrive at a reliable inference for the effect of OCR program on the reading achievement of elementary school students, Borman et al. (2008) designed a multisite, cluster-randomized controlled trial, considering "OCR has never been evaluated rigorously through a randomized trial".

Borman et al. (2008) randomly drew 6 schools from the schools had contacted and shown their interest to SRA/McGraw Hill, the publisher of OCR curriculum. Those 6 schools came from six different states (Florida, Georgia, Idaho, Indiana, North Carolina and Texas) and they were considered to be geographically, ethnically and socioeconomically representative of the schools in US. Within each school and each grade level, classrooms were randomly assigned to the group that was treated with OCR program or the control group. With strong confidence in the internal validity of the design, Borman et al. (2008) estimated the effect of OCR curriculum on student reading composite scores as 7.95, which is statistically significant with an effect size equal to 0.16. Based on the result and design, Borman et al. (2008) went on and concluded that "the outcomes from these analyses provided not only evidence of the promising 1-year effects of OCR on students' reading outcomes but also suggest that these effects may be replicated across varying contexts with rather consistent and positive results".

Nevertheless, the strong internal validity endowed by randomization cannot preempt the debate about external validity, especially when a conclusion is hinged on strong external validity as the one made by Borman et al. (2008). As pointed out by Frank et al. (2013), the study population of Borman et al. (2008) are essentially schools which were volunteered in their research on the effect of OCR program since Borman et al. (2008) only sampled schools from the list of schools that had reached out to the publisher of OCR curriculum. However, it would be suspicious to think the effect of OCR program is the same as the one reported by Borman et al. (2008) when their study is conducted in non-volunteered schools, possibly because volunteered schools were more experienced and capable to carry out programs like OCR and therefore might think OCR program was advantageous for them in particular. In this case, the effect of OCR curriculum was apparently overestimated and the conclusion drawn by Borman et al. (2008) may not be warranted for non-volunteered schools. I will apply the Bayesian model of robustness indices for external validity to Borman et al. (2008) next to quantify the robustness of its inference as well as identify the situations where this inference becomes intolerably fragile.

The analysis of the inferential robustness of Borman et al. (2008) starts with the following sampling process. First, Borman et al. (2008) had 27 classrooms randomly sampled and assigned to the OCR group and 22 classrooms randomly sampled and assigned to the control group, as I described earlier. In addition, the Bayesian models of robustness indices require one to conceptualize the proportion of the relative size of the population of volunteered schools in the population of all US schools, which is denoted as π_R. Suppose π_R is thought to be 0.5, that is, roughly half of the US schools were fundamentally different from the volunteered schools, which is the observed part of ideal population in the study of Borman et al. (2008). Furthermore, an imaginary sampling process took place in the non-representable part of ideal population for Borman et al. (2008), i.e., the half of US schools that were considerably distinct from the volunteered schools. This imaginary sampling process should be mostly identical to the observed sampling process in Borman et al. (2008), namely drawing 5 or 6 schools (or equivalently 49 classrooms) from its non-representable part of ideal population and then randomly assigning 27 classrooms to the OCR group and 22 classrooms to the control group in those unobserved sampled schools. In general, for a given π_R I conceptualize that $49*\frac{1-\pi_R}{\pi_R}$ classrooms were drawn from the non-representable part of ideal population of Borman et al. (2008) and subsequently roughly $27*\frac{1-\pi_R}{\pi_R}$ classrooms were randomly assigned to the OCR group. (so there should be $22*\frac{1-\pi_R}{\pi_R}$ classrooms in the control group).

Draw on this imaginary sampling procedure, I conceptualize the mean reading composite scores for the classrooms randomly assigned to the control group and randomly sampled from the US schools that were fundamentally different from the volunteered ones in Borman et al. (2008) as

611.5. I further conceptualize the mean reading composite scores for the classrooms randomly assigned to the OCR group and randomly drawn from the US schools that were fundamentally different from the volunteered ones in Borman et al. (2008) as 611.5*α. The value of 611.5 is chosen as it is the overall of mean of the whole sample in Borman et al. (2008) and the case of $\alpha = 1$ typifies the null hypothesis which states the average treatment effect is 0.

The prior and the likelihood functions for Borman et al. (2008) are built based on the information of unobserved treated and control sample and Frank et al. (2013) (see pg. 444):

$$\begin{aligned} &\mu_t \sim N(611.5 * \alpha, \frac{45}{n_t}) \\ &Y_t \sim N(\mu_t, 45) \\ &\mu_c \sim N(611.5, \frac{45}{n_c}) \\ &Y_c \sim N(\mu_c, 45) \end{aligned} \tag{6.1}$$

Where the unobserved treated sample size n_t and the unobserved control sample size n_c are:

$$\begin{aligned} n_t &= 27 * \frac{1 - \pi_R}{\pi_R} \\ n_c &= 22 * \frac{1 - \pi_R}{\pi_R} \end{aligned} \tag{6.2}$$

Next, I capitalize on the probit function established in (5.8) to inform the thresholds of π_R or α for the probability of invalidating the inference made by Borman et al. (2008) to be smaller than a desired level. The following list of parameters are contained in the Bayesian model (6.1) and (6.2) and to be plugged into the probit model (4.8) (Also see Frank et al. (2013)):

$$
\begin{aligned}
\bar{Y}_c^{un} &= 611.5 \\
\bar{Y}_t^{ob} &= 615 \\
\bar{Y}_c^{ob} &= 607 \\
\sigma_t^2 &= 45 \\
\sigma_c^2 &= 45 \\
N_t &= 27 \\
N_c &= 22
\end{aligned}
\tag{6.3}
$$

The final step is to quantify an appropriate statistical threshold to account for the added unobserved samples. By plugging the parametric values in (6.3) into the generic expression in (5.4), this threshold is obtained as $1.96 * \sqrt{\pi_R\left(\frac{45}{27}+\frac{45}{22}\right)}$.

The probit model corresponds to the parametric values assumed in (6.3) is:

$$probit(p) = 317.38\alpha\pi_R^{\frac{1}{2}} - 317.38\alpha\pi_R^{-\frac{1}{2}} - 321.54\pi_R^{\frac{1}{2}} + 317.38 \cdot \pi_R^{-\frac{1}{2}} + 1.96 \tag{6.4}$$

The above probit function is utilized in the following fashion: first, one needs to set up a desired level of probability of invalidating the inference made by Borman et al. (2008), for example, as 0.5. This means he would like to find out the threshold for α or π_R such that the probability of invalidating the inference of Borman et al. (2008) is smaller than 0.5. Moreover, the threshold for α is conditional on the value of π_R and vice versa. Specifically, the threshold for α is first calculated as a function of π_R based on the desired level of probability of invalidating the inference and subsequently instantiated with some selected values of π_R so that it could be quantified as numbers instead of as a function. The threshold for π_R is approached with the same procedure except that it is contingent on the value of α.

From (6.4), the boundary line that separates the area within which probability of invalidating the inference is larger than 0.5 and the area within which probability of invalidating the inference is smaller than 0.5 is:

$$317.38\alpha\pi_R^{\frac{1}{2}} - 317.38\alpha\pi_R^{-\frac{1}{2}} - 321.54\pi_R^{\frac{1}{2}} + 317.38\cdot\pi_R^{-\frac{1}{2}} < -1.96 \qquad (6.5)$$

More importantly, the inequality (6.5) leads to the following quadratic inequality for $\pi_R^{0.5}$ when α is a given fixed number:

$$(317.38\alpha - 321.54)\pi_R + 1.96\pi_R^{0.5} + (317.38 - 317.38\alpha) < 0 \qquad (6.6)$$

Assuming $\alpha = 1$, the quadratic inequality (6.6) will generate the following lower bound for π_R in order to keep the probability of invalidating the inference of Borman et al. (2008) lower than 0.5:

$$\pi_R > 0.22 \qquad (6.7)$$

which suggests that the proportion of the observed sample in an ideal sample should be larger than 0.22 so as to keep the probability of invalidating the inference of Borman et al. (2008) smaller than 0.5.

From the boundary function (6.5), the inequality for α can be derived as follows:

$$\alpha > \frac{321.54\pi_R^{0.5} - 317.38\pi_R^{-0.5} - 1.96}{317.38(\pi_R^{0.5} - \pi_R^{-0.5})} \qquad (6.8)$$

For an instance, conditional on $\pi_R = 0.46$ the above inequality suggests α should be larger than 0.9966 in order to make the probability of invalidating the inference of Borman et al. (2008) smaller than 0.5.

The inequality (6.6) reveals that the bounds of π_R can be computed through (6.6) as long as a value of α is given, for the purpose of keeping the probability of invalidating the inference of

Borman et al. (2008) under 0.5. Conditional on $\alpha = 1$, which means the mean reading composite scores of both unobserved treated and control sample is 611.5, π_R needs to be larger than 0.22 for the probability of invalidating the inference made by Borman et al. (2008) to be smaller than 0.5. An interpretation of this lower bound 0.22 would be that one can add an unobserved sample potentially drawn from the non-volunteered schools to the observed sample but this unobserved sample can contain at most 95 OCR classrooms and 78 control classrooms assuming the effect of Open Court Reading is absolutely zero for those unobserved classrooms, i.e., the mean reading scores of those 95 unobserved OCR classrooms and of those 78 unobserved control classrooms are both 611.5.

Equally meaningful, the inequality (6.8) suggests that α must be larger than the ratio on the right-hand side, which is a function of π_R only, so as to keep the probability of invalidating the inference made by Borman et al. (2005) under 0.5. This threshold of α can be evaluated at every given number of π_R. For example, one can fix the value of π_R at 0.46 and the resultant lower bound for α is 0.9966 in order to make the probability of invalidating the inference made by Borman et al. (2008) smaller than 0.5, which requires the mean reading score of the classrooms which were randomly assigned to the Open Court Reading classrooms and randomly sampled from the non-volunteered schools to be at least 609.42. This is about two points lower than the mean reading score of the students in the classrooms which were randomly assigned to the control classrooms and randomly drawn from the non-volunteered schools.

The threshold of π_R (or α) can be repeatedly calculated for the desired probability of invalidating the inference of your choice conditional on any fixed sensible value of α (or π_R). Table 1.1 and Table 1.2 provide thresholds of α and thresholds of π_R when the desired level of probability of invalidating the inference is from 0.1 to 0.9. It further provides the threshold of average treatment effect based on an ideal sample, which is just $\theta_t - \theta_c$ in (4.6), to help researchers interpret those levels of probability of invalidating the inference as the desired levels of the estimate of average treatment effect. For an instance, α needs to be larger than 1.0017 for the probability of nullifying the inference in Borman et al. (2008) to be smaller than 0.1 holding π_R constant as 0.46. Meanwhile, this threshold of α suggests the estimate of average treatment effect of OCR in an ideal sample should be larger than 4.24, for the probability of nullifying Borman et al. (2008)'s inference to be lower than 0.1. Choosing a desired level of probability of invalidating an inference, just as choosing a threshold related to a decision about an intervention or policy or program discussed in Frank et al. (2013), should be based on the features and the specific context of a research design.

Figure 1.4 illustrates the relationship of testing null hypothesis and the posterior probability of invalidating Borman et al. (2008)'s inference when we iteratively plug in the thresholds of α tabulated in table 1 into the probit model (6.4) conditional on $\pi_R = 0.46$. It's evident that as the threshold of α decreases the posterior distribution (red curve) moves towards the distribution corresponding to null hypothesis (black curve). As a result, the posterior probability of invalidating Borman et al. (2008)'s inference is growing. Essentially, the posterior probability of invalidating the inference of Borman et al. (2008) is type II error of retesting null hypothesis: $\mu_t - \mu_c = 0$ against the alternative hypothesis: $\mu_t - \mu_c$ follows the posterior distribution in (2.13), when an unobserved sample randomly drawn from the non-volunteered schools is available and added to their observed sample. Figure 1.5 unfolds the same relationship between testing null hypothesis and posterior probability of invalidating the inference of Borman et al. (2008), except that it is built on the thresholds tabulated in table 1.2 and conditional on $\alpha = 1$. In figure 2.5, the statistical threshold as well as the posterior variance decreases when π_R decreases, which indicates the unobserved sample size as well as the ideal sample size is enlarging. We also observe the posterior distribution is shifting towards the distribution corresponding to null hypothesis when π_R is decreasing.

Table 1.1: Thresholds of α when π_R is fixed as 0.46

Level of probability	Threshold of α	Threshold of the mean of an unobserved treated sample	The estimate of average treatment effect based on an ideal sample
0.1	1.0017	612.54	4.24
0.2	0.9999	611.44	3.65
0.3	0.9987	610.71	3.25
0.4	0.9976	610.03	2.89
0.5	0.9966	609.42	2.56
0.6	0.9956	608.81	2.23
0.7	0.9945	608.14	1.86
0.8	0.9933	607.4	1.47
0.9	0.9915	606.3	0.87

Table 1.2: Thresholds of π_R when α is fixed as 1

Level of probability	Threshold for π_R	The estimate of average treatment effect based on an ideal sample
0.1	0.6095	4.88
0.2	0.4553	3.64
0.3	0.358	2.86
0.4	0.284	2.27
0.5	0.2228	1.78
0.6	0.1689	1.35
0.7	0.1195	0.96
0.8	0.0725	0.58
0.9	0.0267	0.21

Figure 1.4: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Borman et al. (2008) (π_R is fixed as 0.46)

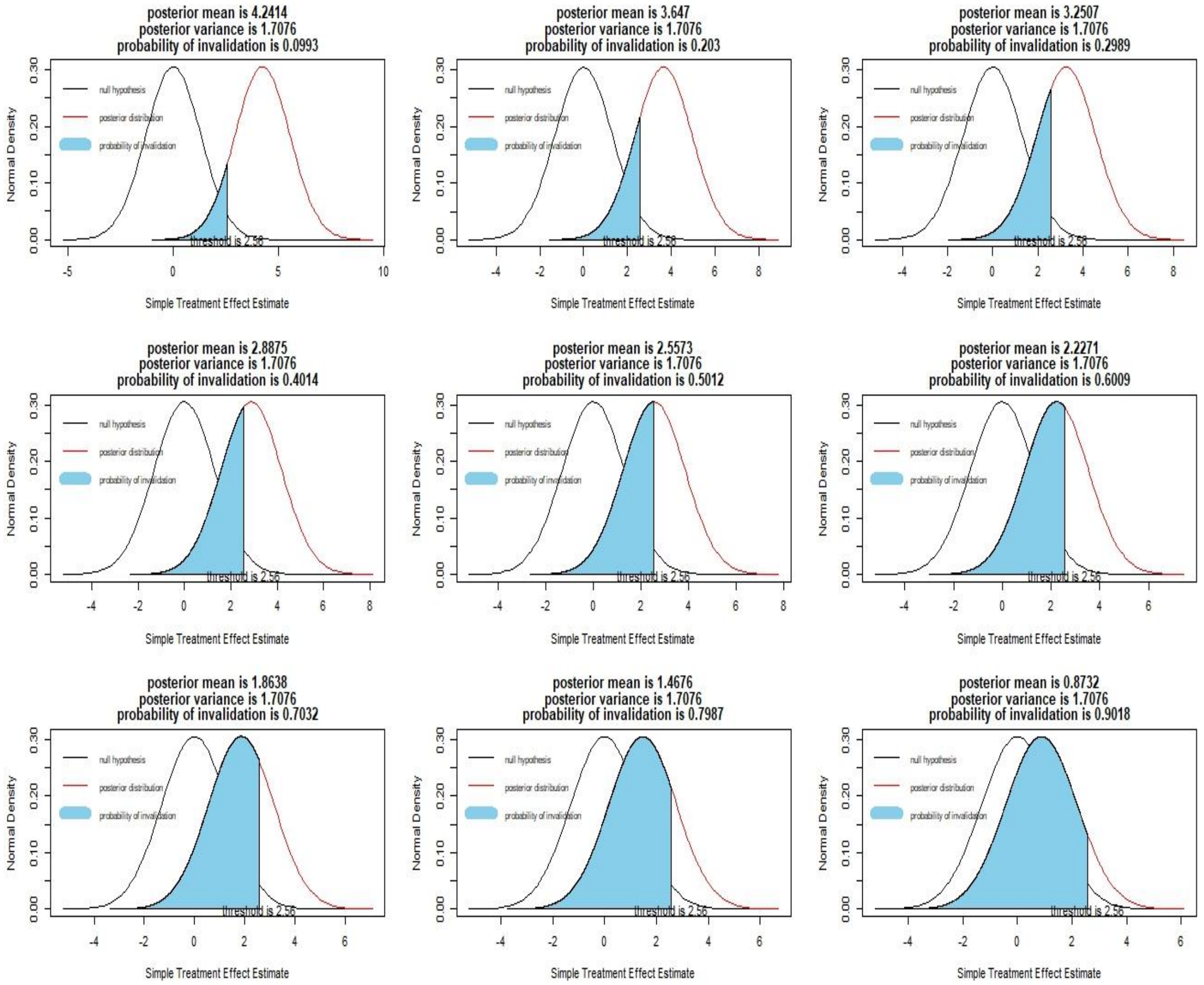
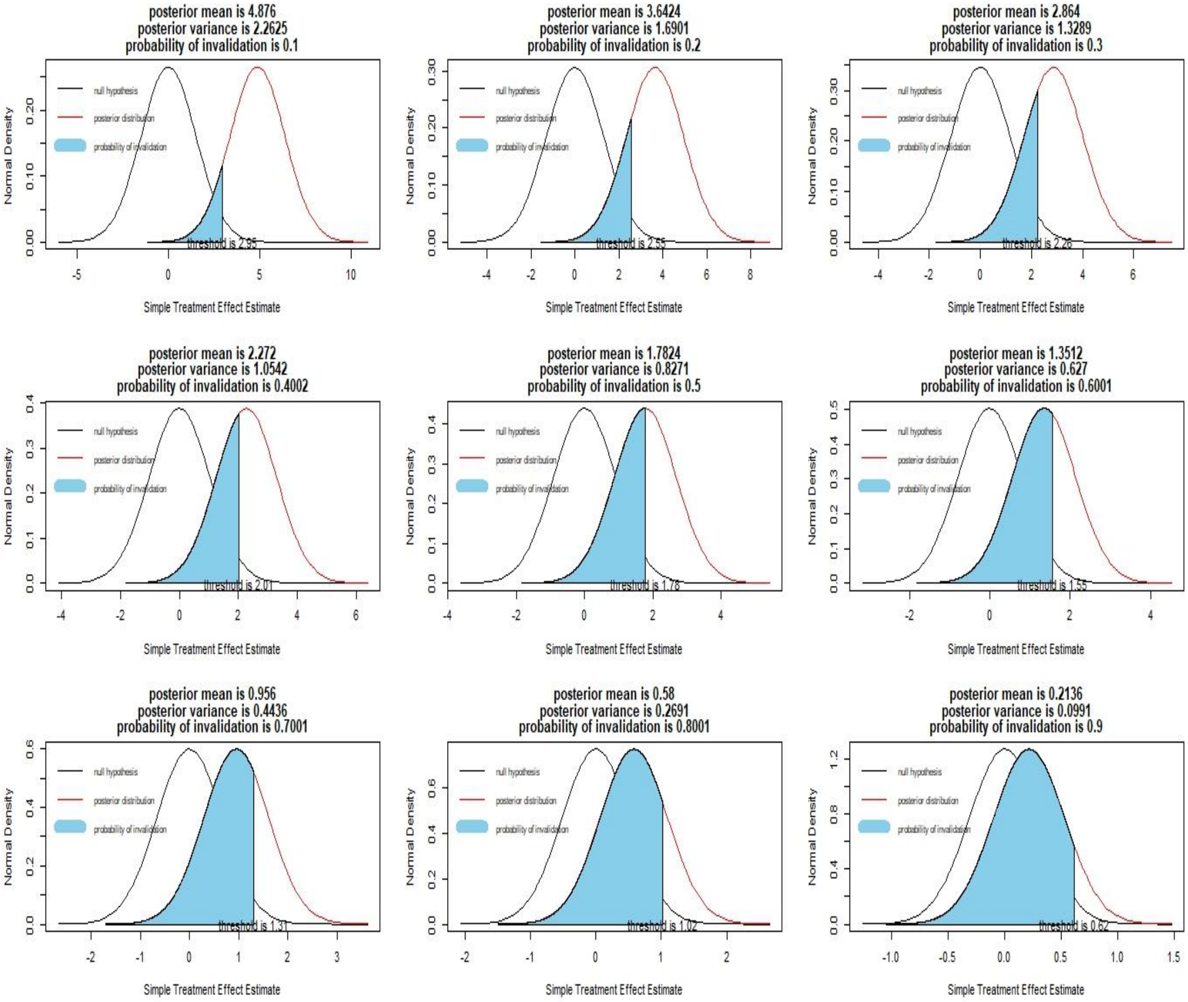


Figure 1.5: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Borman et al. (2008) (α is fixed as 1)



6.2-The Bayesian robustness indices of the effect of kindergarten retention on reading achievement

Hong & Raudenbush (2005) and Frank et al. (2013) have pointed out that kindergarten retention is a widespread phenomenon in the US and its impact could be profound for both promoted children and retained children, and therefore it has long been a controversial issue. To resolve such controversy, Hong & Raudenbush (2005) conducted the analysis which combined the multilevel model controlling for propensity scores and additional propensity score stratification to evaluate the effects of kindergarten retention policy and actual kindergarten retention on students' academic achievement. Such analysis is necessary and possibly effective for the purpose of reducing the selection bias due to the lack of randomization in this kind of studies. Draw on this method, Hong & Raudenbush (2005) estimated the effect of kindergarten retention on students' reading achievement as -9.01 and its standard error as 0.68, which is tantamount to a significant effect whose size is about 0.67. In light of this considerable effect, Hong & Raudenbush (2005) concluded that "children who were retained would have learned more had they been promoted" and therefore "kindergarten retention treatment leaves most retainees even further behind".

Nevertheless, the method proposed by Hong & Raudenbush (2005) does not prevent the selection bias from persisting for two reasons: First, propensity score analysis is built on the assumption of igorability, which basically says all confounding variables are able to be observed and controlled in the causal model. However, as argued by Frank et al. (2013), some confounding variables such as motivation of a child may not be measured and controlled in the causal model of Hong & Raudenbush (2005), and this will result in violation of the assumption of ignorability and incur the selection bias of their estimate. Second, to ensure that quasi-experimental design is a plausible approximation of randomized experiment, the estimated propensity scores need to be good balancing scores, which means most if not all controlled covariates have to be balanced conditional on the estimated propensity score. Even though Hong & Raudenbush (2005) reported 97% of the covariates had achieved balance and argued that the existence of the remaining imbalanced covariates "could largely be attributed to the Type I error related to sampling fluctuation", there is little evidence to show that such imbalance of those 3% of the covariates is due to sampling error and not consequential. Most importantly, the credibility of quasi-experimental design will be greatly undermined if the imbalanced covariates are happened to be the most influential covariates. (See the draft of Maroulis, Frank & Duong). In cases such that motivation was negatively correlated with kindergarten retention and positively correlated with reading achievement and promoted children had significantly higher pretest readings scores than retained children did in some propensity score strata, the negative effect of kindergarten retention could be mitigated or even reversed.

The aforementioned innate limitations of quasi-experimental design prompt us to capitalize on the Bayesian model of robustness indices for internal validity to express the robustness of the inference made by Hong & Raudenbush (2005) as the probability of invalidating their inference. Furthermore, for a chosen desired level of this probability (say 0.5), a threshold characterizing an unobserved sample can be computed to determine when the probability of invalidating their inference will exceed this desired level. As in the example of Borman et al. (2008), the underlying sampling process of the Bayesian model of robustness indices for the internal validity of Hong & Raudenbush (2005) is conceptualized as follows:

1-The observed treated sample is constituted of 471 retained children and the observed control sample is constituted of 7168 promoted children, according to Hong & Raudenbush (2005).

2-From Rubin Causal Model (RCM), the unobserved part of treated population is the collection of the reading scores of promoted children had they all been retained instead and the reading scores of retained children had they all been promoted instead. In the terminology of RCM, the unobserved part of ideal population contains all possible values of the counterfactuals for the promoted students and all possible values of the counterfactuals for the retained students.

3-An unobserved sample should be randomly drawn from the unobserved part of ideal population. Furthermore, it can be decomposed into an unobserved treated sample and an unobserved control sample. This unobserved treated sample is a group of reading scores of all sampled promoted children had they been retained instead, and therefore its sample size should be 7168. Likewise, this unobserved control sample is a group of reading scores of all sampled retained children had they been promoted instead, and thus its sample size should be 471.

Dependent on the above sampling procedure, I assume the mean reading test scores of an unobserved control sample and an unobserved treated sample are 45.2 and 45.2*α respectively. As mentioned earlier, the case of $\alpha = 1$ corresponds to the null hypothesis that asserts the average treatment effect of kindergarten retention is 0.

Again, the prior and likelihood functions are constructed based on those unobserved samples, Hong & Raudenbush (2005) (pg.216) and Frank et al. (2013) (pg.448):

$$\begin{aligned} &\mu_t \sim N(45.2 * \alpha, \frac{143.26}{n_t}) \\ &Y_t \sim N(\mu_t, 143.26) \\ &\mu_c \sim N(45.2, \frac{138.83}{n_c}) \\ &Y_c \sim N(\mu_c, 138.83) \end{aligned} \tag{6.9}$$

Where:

$$\begin{aligned} n_t &= 7168 \\ n_c &= 471 \end{aligned} \tag{6.10}$$

To find out the threshold of α such that it is a switching point of whether the probability of invalidating the inference of Hong & Raudenbush (2005) is smaller than a preselected desired value (say 0.5), I utilize the probit model (3.5) and extract following parametric values from Hong & Raudenbush (2005) and Frank et al. (2013):

$$\begin{aligned} \bar{Y}_c^{un} &= 45.2 \\ \bar{Y}_t^{ob} &= 36.77 \\ \bar{Y}_c^{ob} &= 45.78 \\ \sigma_t^2 &= 143.26 \\ \sigma_c^2 &= 138.83 \\ N &= 7639 \\ \pi &= 0.0617 \end{aligned} \tag{6.11}$$

Guided by (5.3) and parametric values above, the appropriate statistical threshold is determined as $-1.96\sqrt{\frac{143.26+138.83}{7639}}$ which equals -0.38.

Based on (6.11), the probit model can be explicitly written as this:

$$probit(p) = 221.49\alpha - 225.09 \tag{6.12}$$

From (6.12), the threshold of α can be located conditional on the parametric values as assumed in (6.11), once the desired level of probability is given. I note here that the threshold of α can surely repeatedly calculated contingent on various desired levels of probability of invalidating the inference of Hong & Raudenbush (2005) while holding values in (6.11) fixed.

Again, when the desired value of probability is set to be 0.5, the boundary line separating the region where the probability of invalidating the inference of Hong & Raudenbush (2005) is larger than 0.5 and the region where the probability of invalidating the inference of Hong & Raudenbush (2005) is smaller than 0.5 should be:

$$221.49\alpha - 225.09 < 0 \qquad (6.13)$$

It could be learned from (6.13) that α needs to be smaller than 1.0162 so as to make the probability of invalidating an inference smaller than 0.5, assuming π is 0.0617 and the mean reading score of the retained children had all of them been promoted instead is 45.2. Equivalently, this means $\bar{Y}_t^{un}$, i.e., the mean reading score of the promoted children had all of them been retained instead, has to be smaller than 45.93 for the probability of invalidating an inference lower than 0.5. Moreover, this threshold of μ_t^{un} can be recast as the threshold of average treatment effect based on an ideal sample, i.e., $\theta_t - \theta_c$, since it is a function of $\bar{Y}_t^{un}$ and the parametric values in (6.11). In the setting of current example, the threshold of average treatment effect based on an ideal sample is -0.38, which is exactly the appropriate statistical threshold.

The threshold of α can be obtained for any given desired level of the probability of invalidating the inference. Table 1.3 tabulates the thresholds of α, μ_t^{un} and $\theta_t - \theta_c$ when the desired level of the probability of invalidating the inference of Hong & Raudenbush (2005) is 0.1, 0.2, …, 0.9. For example, α could be at most as large as 1.0138 for the sake of keeping the probability of invalidating their inference under 0.3, which indicates that the mean reading score of promoted students had they all been retained needs to be smaller than 45.82 and the estimate of average treatment effect acquired from an ideal sample should be even more extreme than -0.48, given π as 0.0617 and the mean reading score of retained students had they all been promoted as 45.2. However, in an empirical research, decision about the desired level of the probability of nullifying its inference should be a rational choice based on its cost and policy/behavioral implications as argued by Frank et al. (2013) rather than a haphazard choice.

One may notice that the thresholds for α in table 1.3 are all very close to 1, which means for almost any level of probability of invalidating the inference of Hong & Raudenbush (2005) the means of unobserved treated and control sample should be roughly equal. This may appear to be unintuitive, however, is not surprising in the case of Hong & Raudenbush (2005) as their sample size is considerable. Ordinarily, the probability of invalidating an inference will be quite sensitive and jumps/drops sharply within a certain range of α as to a study with questionable internal validity and large sample size, as depicted in figure 1.6.

Again, one main research goal is to learn the relationship between testing null hypothesis and the posterior probability of invalidating the inference of Hong & Raudenbush (2005). For this purpose, figure 1.7 is presented. The general pattern is, when α increases the posterior distribution moves toward the distribution corresponding to the null hypothesis and therefore the posterior probability of invalidating the inference of Hong & Raudenbush (2005) becomes larger. As discussed earlier, such relationship parallels the relationship between testing null hypothesis versus alternative hypothesis and type II error. The posterior probability of invalidating the inference of Hong & Raudenbush, is tantamount to type II error of retesting null hypothesis: $\mu_t - \mu_c = 0$ when an unobserved sample (i.e., a collection of counterfactual outcomes of all their sampled students) is actualized and added to their observed sample.

Table 1.3: Thresholds of α when π is fixed as 0.0617

Level of probability	Threshold for α	Threshold for the mean of an unobserved treated sample	The estimate of average treatment effect based on an ideal sample
0.1	1.0104	45.67	-0.62
0.2	1.0124	45.76	-0.54
0.3	1.0138	45.82	-0.48
0.4	1.015	45.88	-0.43
0.5	1.0162	45.93	-0.38
0.6	1.0174	45.99	-0.33
0.7	1.0186	46.04	-0.28
0.8	1.02	46.1	-0.22
0.9	1.022	46.19	-0.13

Figure 1.6: The relationship between α and the probability of invalidating the inference of Hong & Raudenbush (2005)

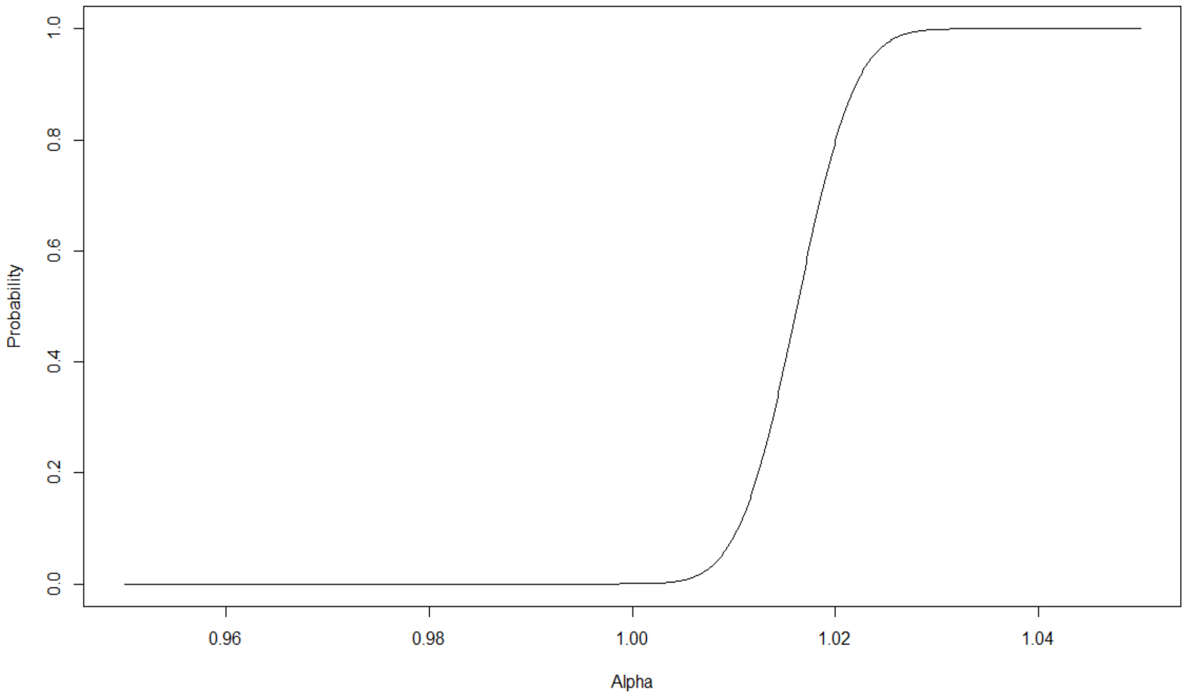
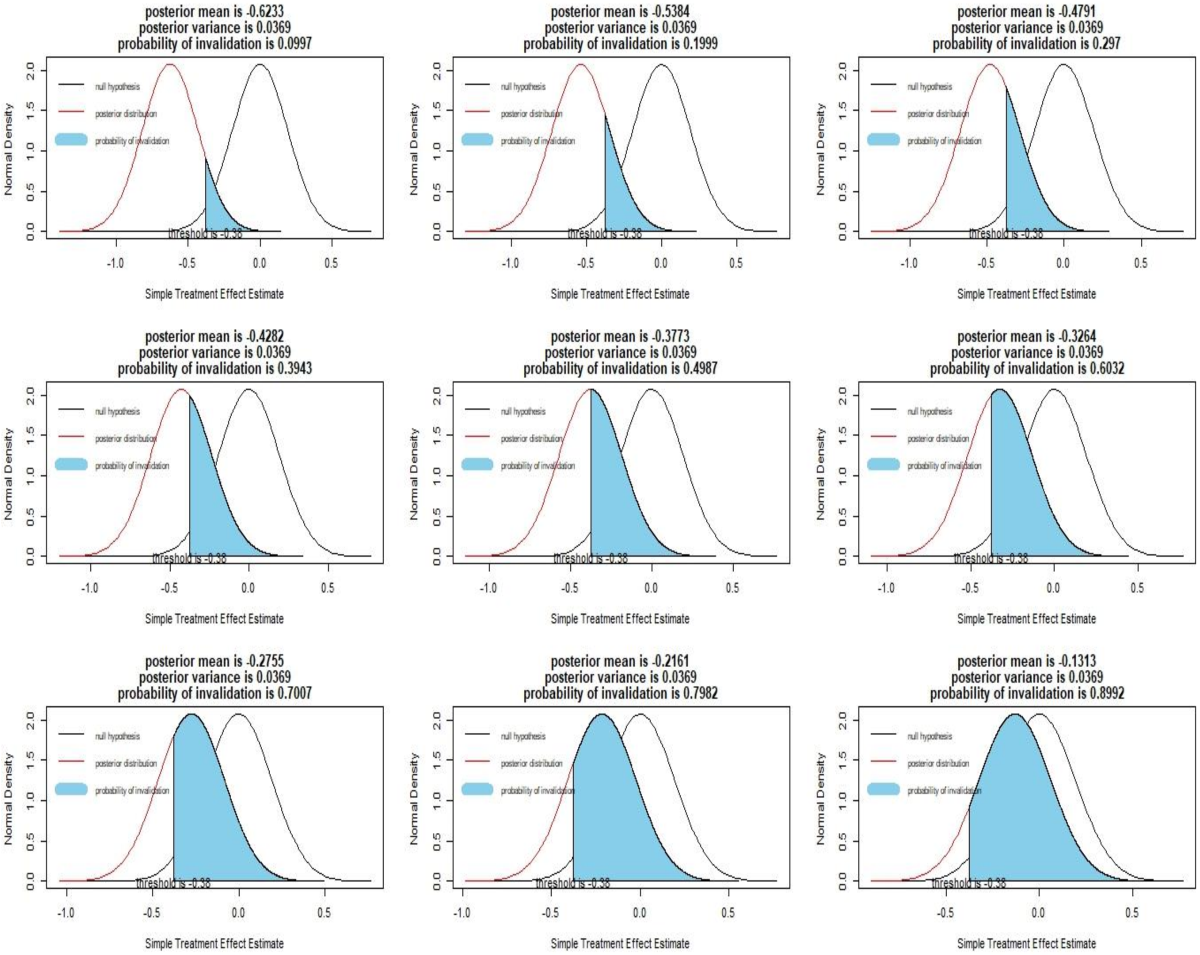


Figure 1.7: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Hong & Raudenbush (2005)


posterior mean is -0.6233
posterior variance is 0.0369
probability of invalidation is 0.0997

posterior mean is -0.5384
posterior variance is 0.0369
probability of invalidation is 0.1999

posterior mean is -0.4791
posterior variance is 0.0369
probability of invalidation is 0.297

posterior mean is -0.4282
posterior variance is 0.0369
probability of invalidation is 0.3943

posterior mean is -0.3773
posterior variance is 0.0369
probability of invalidation is 0.4987

posterior mean is -0.3264
posterior variance is 0.0369
probability of invalidation is 0.6032

posterior mean is -0.2755
posterior variance is 0.0369
probability of invalidation is 0.7007

posterior mean is -0.2161
posterior variance is 0.0369
probability of invalidation is 0.7982

posterior mean is -0.1313
posterior variance is 0.0369
probability of invalidation is 0.8992

7-Discussion and conclusion

7.1-Features of the Bayesian paradigm of robustness indices

The Bayesian paradigm proposed in this paper has some remarkable characteristics. First, it treats the problem of causal inference as a missing data issue, which exactly is the essence of causal inference according to RCM. Specifically, I define the "missing data" as an unobserved sample which could be thought as a sample randomly drawn from the unobserved part of ideal population. The definition of unobserved sample depends on which one of internal and external validity is of central concern for researchers. For example, it could be a random sample from the schools which didn't show interest in the study of OCR curriculum when the external validity of Borman et al (2008) is challenged or the counterfactuals of the test scores of kindergarten students in the study of kindergarten retention when the internal validity of Hong & Raudenbush (2005) is disputable. An ideal sample is formed by adding an unobserved sample to the observed one and this ideal sample should be able to lead to an unbiased estimate of the treatment effect. Following Frank & Min (2007), the posterior distribution can be interpreted as the distribution of the estimate based on an ideal sample, by treating the distribution of the estimate based on an unobserved sample as prior distribution and constructing likelihood function which should contain information of the observed sample. I have demonstrated that, it is posterior distribution that yields the probability of invalidating the inference and express this probability as a function of the parameters in prior and likelihood distribution.

Another notable feature of the Bayesian paradigm is that it is in fact Bayesian sensitivity analysis which manipulates the posterior distribution by inputting different informative priors while holding the observed data and likelihood function fixed. Defining distribution of the imaginary unobserved sample statistic as prior distribution is totally legitimate in the Bayesian world.

Recall that a Bayesian model typically requires the prior distribution to be based on one's belief about the parameter before he actually collects and analyzes the data. My Bayesian models demand a reasonable conjecture about the unobserved part of ideal population such that it is possible to be true if the unobserved part of ideal population can somehow be reached. This requires, in Bayesian language, prior is subjective (or objective and informative) instead of noninformative because an noninformative prior will make the probability of invalidating an inference noninformative as well. By this logic, the prior mean is usually a meaningful quantity and the prior variance is usually relatively small in practice. Essentially, the Bayesian paradigm of robustness indices is about checking the influence of the parameters of prior distribution on posterior distribution, i.e., how the changes in prior distribution affect posterior distribution and the inference built on it. This is exactly the spirit of Bayesian sensitivity analysis. I propose the probability of invalidating the inference as a new index of the robustness of causal inference in this paper since it quantifies the condition under which (prior) and the degree to which (probability) a particular inference is robust.

Furthermore, the Bayesian paradigm enhances the interpretability of the framework of robustness indices. First of all, the underlying sampling process of the Bayesian paradigm could be conceptualized as follows: conditional on the fixed observed sample, one sample can be randomly drawn from the unobserved part of ideal population and merged into the observed sample to construct an imaginary ideal sample. This procedure can be implemented many different times and thus it can generate many different ideal samples, with the observed sample being the same and fixed. Equally importantly, the expression of the mean of the ideal sample in (2.10) evidence that the ideal sample mean can be interpreted as the weighted average between the unobserved sample mean and observed sample mean, which is consistent with the arguments made by Frank & Min (2007). Specifically, those weights will solely rely on the sample sizes of unobserved and observed sample. As a result, the Bayesian paradigm itself could be interpreted as a sampling scheme where one ideal sample is comprised of heterogeneous subsamples with different sample sizes.

7.2-Comparisons with other similar approaches

7.2.1-The robustness indices in Frank et al. (2013)

By asking the question "what would it take to change your inference", Frank et al. (2013) initiated the robustness indices which were the proportion of the original data that was necessary to be replaced with the hypothetical data of zero treatment effect, for the purpose of invalidating the inference. The Bayesian paradigm has three basic distinctions from the robustness indices in Frank et al. (2013): First, the robustness indices in Frank et al. (2013) and the Bayesian paradigm are quantified in different forms. Specifically, the robustenss indices in the Bayesian paradigm are posterior probabilities of invalidating an inference rather than a proportion of data need to be replaced in Frank et al (2013). Second, the derivation of robustness indices is different in those two frameworks. My Bayesian approach adopts a distributional thinking, i.e., it demands a distribution built on an unobserved sample randomly drawn from the unobserved part of ideal population for its expected value. On the contrary, Frank et al. (2013) usually focuses on the estimate of average treatment effect in the unobserved part of ideal population and directly assumes it to be a certain value. Third, the sampling mechanisms implied by those two kinds of robustness indices are dissimilar. The Bayesian paradigm, as explained in the last section, is actually drawing and including an unobserved sample into the observed sample, instead of replacing a portion of the observed sample with a hypothetical unobserved sample, which is embeded in Frank et al. (2013). Still, it is noteworthy that, both kinds of robustness indices are cognate in that they share the same definition of bias, both consider modeling of the unobserved population to be central and quantify the robustness of an inference as a threshold when it is likely to be overturned.

7.2.2-The robustness indices in Frank & Min (2007)

There are two major differences between my Bayesian paradigm and the robustness indices in Frank & Min (2007). One is that the Bayesian paradigm of robustness indices addressess both questions about internal validity and external validity while the paper of Frank & Min is only intended for the question about external validity. Additionally, Frank & Min proposed two sampling schemes that can explain the induction of bias and derivation of robustness indices, namely neutralization by replacement and neutralization by addition. The robustness indices of Frank et al. (2013) is well situated in the former one while the Bayesian paradigm is well situated in the latter one.

7.2.3-The bounds on treatment effect in Manski (1990) and Lee (2009)

As robustness indices of causal inferences, the bounds on treatment effect proposed by Manski (1990) and Lee (2009) target the potential bias associated with the point estimate of treatment effect and highlight the identification issue of such point estimate due to confounded sample selection. However, the approach of bounding effect is different from the robustness indices of causal inferences in three main aspects: First, the robustness indices are rooted in pragmatism and decision-making while the bounds of treatment effect are intended for the estimation problem. Specifically, the purpose of judging whether an effect is significant (or equivalently whether a null hypothesis should be rejected) and whether a decision should be made thereupon is better served by the robustness indices of causal inferences. The bounds on treatment effect may better inform researchers about what data and assumptions can do and what they cannot do.

Second, the robustness indices of causal inferences rely on thought experiments, i.e., conceptualizations of unobserved samples while the bounds on treatment effect do not. Third, the robustness of causal inferences is quantified through thresholds of invalidating an inference by the robustness indices rather than bounds on treatment effect offered by Manski (1990) or Lee (2009).

7.3-Limitations

Although the Bayesian paradigm of robustness indices is an useful tool as to quantifying the robustness of a causal inference, I caution the readers about its limitations here so that one can decide how to implement it by weighing the gains and risks. First, my Bayesian paradigm has focused exclusively on the estimate of average treatment effect. With that being said, the Bayesian paradigm is not intended for the bias in the estimates of average treatment effect for the treated and average treatment effect for the control, and consequently, it should not be used to quantify the robustness of causal inferences due to such kind of bias. Generally speaking, the Bayesian paradigm will not be suitable for modeling the robustness of causal inferences occasioned by the bias of estimate of any differential causal effect, i.e., the treatment effect conditional on any covariates. For example, it would be inappropriate to employ the Bayesian paradigm of robustness indices on the inference of the average treatment effect of OCR curriculum for students with high social-econimical status or the average treatment effect of kindergarten retention for girls. Second, there are other sources of bias that can undermine causal inferences besides insufficient internal validity and external validity, such as measurement error and violation of SUTVA. I emphasize here that the Bayesian paradigm is designated only for measuring the degree to which a causal inference is affected by its debatable internal validity or external validity.

7.4-Conclusion

The Bayesian paradigm of robustness indices is an addition to the current literature of robustness indices of causal inferences, which purpose to bridge statistical inference and causal inference by guiding researchers when their inferences are too delicate to uphold as the conceptualization of unobserved sample is varying. Cohen (1990) pointed out that "A successful piece of research doesn't conclusively settle an issue, it just makes some theoretical proposition to some degree more likely. Only successful future replication in the same and different settings provides an approach to settling the issue". (pg.1311). Indeed, even a statistical inference based on a careful design like Borman et al. (2008) or Hong & Raudenbush (2005) should not be deemed as a established causal inference without further inquiry into the sources of bias. Starting at the definition of bias due to limited internal validity or external validity, the Bayesian paradigm of robustness indices is managed to ask and answer the question "What would an unobserved sample have to be for the probability of invalidating my inference is small enough (than a predetermined desired value of mine)?", or equivalently "How different can I afford for an unobserved sample to be from the observed one so that the probability of invalidating my inference is small enough?".

Essentially, the Bayesian paradigm of robustness indices is consistent with the argument of Cohen (1990) in that it quantifies the robustness through the modeling of unobserved sample and thereby simulates the replications of the same study in various contexts and the probability of an replication is successful. It is my hope that through the Bayesian paradigm of robustness indices proposed in this paper, researchers are able to cast their conclusions in terms of the degree to which their inferences will be valid under what circumstances and therefore contribute to the scientific discourse of a particular causal relationship.

Chapter 2: The Bayesian paradigm of robustness indices of causal inferences for regression models

1-Introduction

1.1-Regression-based causal inference

A causal question is nearly impossible to be convincingly resolved unless a research raising such a question is deemed to have both indisputable internal and external validity. Empirically this means that a randomized experiment with a representative sample is a prerequisite for answering any causal question. Under this ideal condition, extensive literature has justified the usage of regression-based causal inference, i.e., the approach of treating the outcome as the dependent variable and an binary treatment indicator as an independent variable in regression. According to Imbens & Rubin (2015), regression-based causal inference, when subjects are randomly assigned to the treatment and control group, can generate consistent and efficient estimate of a true average treatment effect. Some simulation studies have also indicated that regression-based causal inference is as good as any other methodologies in causal inference under certain assumptions (Morgan & Winships, 2007; Shadish et al., 2008; Steiner et al, 2010; Imbens & Rubin, 2015).

While regression-based causal inference and its offshoots have been predominant in addressing causal questions, critics of regression-based causal inference have questioned the validity of inference brought by this approach (Shadish, Cook and Campbell, 2002). Specifically, when randomization is lacked in a research design, the validity of regression-based causal inference is solely built on the assumption of strong ignorability (Rosenbaum & Rubin, 1983) that is not justifiable or testable (Morgan & Winship, 2007). In this case, it is natural to suspect the internal validity of regression-based causal inference. Moreover, the validity of regression-based causal inference can even be shattered in a randomized experiment, when a research conclusion targets a population of which the observed sample is not fully representative. The cases where regression-based causal inference is potentially invalid can be categorized into two scenarios, which I elaborate next.

The first scenario typically refers to an observational study or quasi-experiment with a representative sample of the target population. Such research shall be labeled as one with strong external validity and yet limited internal validity. The validity of regression-based causal inference hinges on the assumption of strong ignorability. That is, one need to conjecture and do his best to justify the independence between the treatment and the outcome conditional on a set of measured covariates. In addition, the probabilities of selecting/being assigned to the treatment of all subjects have to be strictly smaller than 1 and bigger than 0 conditional on the same set of measured covariates, in order to identify the average treatment effect. The potential pitfall of conducting regression-based causal inference under the strong ignorability assumption, is that one can never prove or disprove this assumption and thus can never completely legitimize using regression-based causal inference under it. Some practical issues, such as checking the overlap of distributions of propensity scores (or logits of them) and the balance of covariates conditional on propensity score, can still exist and potentially compromise the validity of regression-based causal inference even if the strong ignorability assumption is plausible (Gelman & Hill, 2007). Hong & Raudenbush (2005), which evaluated the impact of kindergarten retention on academic achievement, exemplifies this scenario as a random assignment of kindergarten children to retention and promotion groups was impossible while a nationally representative sample from ECLS-K study was available in this research.

The second scenario features any randomized experiment with a nonrandom sample drawn from its target population. A nonrandom sample, as discussed in Wooldridge (2010, 2013), can bias estimates of regression coefficients and therefore make regression-based causal inference inconsistent and biased for true average treatment effect. Gelman & Hill (2007) argued that, in this case "causal inferences are still justified but inferences no longer generalize to the entire population". They suggested that regression-based causal inference is only valid for an imaginary subpopulation and "further modeling is needed to generalize to any other population". I will illustrate this scenario by discussing Borman et al. (2008), which conducted a multisite cluster randomized trial to examine the effect of Open Court Reading (OCR) curriculum with a random sample from the schools which volunteered in this study. Apparently, Borman et al. (2008) enjoyed strong internal validity brought by randomized assignment to OCR and control groups and yet suffered from limited external validity since they attempted to generalize their conclusions to both volunteered and non-volunteered schools.

1.2-The philosophy of robustness indices

The robustness indices proposed by Frank & Min (2007) and Frank et al. (2013) are built on a philosophy that there exists, at least conceptually, an ideal population for any single study planning a causal inference. To elaborate, the following definitions for the first scenario is needed:

Definition 1.1: **A real or non-counterfactual observation** refers to an observation which is observable, i.e., an observation of a controlled subject under the condition of control or an observation of a treated subject under the condition of treatment.

A real or non-counterfactual observation in Hong & Raudenbush (2005) could be an observation of John who was retained in kindergarten or an observation of Mary who was promoted to first grade. It’s noteworthy here that those observations are real since one can only obtain John’s observation when he was retained and Mary’s observation when she was promoted.

Definition 1.2: **A counterfactual observation** of a subject refers to an imaginary observation where his outcome is counterfactual, his membership is different than what is actually observed and his covariates’ values are identical to the ones in his real observation.

In Hong & Raudenbush (2005), a counterfactual observation of John who was retained in the kindergarten would be the observation where the outcome was John’s potential reading score had he been promoted to first grade, the binary indicator of treatment status was 0 (since he is imagined as a promoted student) and the covariates were remained the same as the ones in his real observation. Likewise, the counterfactual of Mary who was promoted to first grade would be the observation where the outcome was Mary’s potential reading score had she been retained in kindergarten, the binary indicator of treatment status was 1 (since she is imagined as a retained student) and the covariates were identical to the ones in her real observation.

Definition 1.3.1: **A potential observation of a subject in the first scenario** refers to either his/her real observation or his/her counterfactual observation.

In Hong & Raudenbush (2005), every student had two potential observations. For example, John had two potential observations, namely his real observation under the condition of being retained in kindergarten and his counterfactual observation under the condition of being promoted to first grade. Similarly, Mary had two potential observations which were her real observation under the condition of being promoted to first grade and her counterfactual observation under the condition of being retained in kindergarten.

With regard to the second scenario, definitions and conceptualizations about real and counterfactual observations are unnecessary since in the long run randomization would guarantee the equivalence between real observations and counterfactual observations. Still, it’s instrumental to offer a different version of the definition of potential observations for the second scenario as follows:

Definition 1.3.2: **A potential observation in the second scenario** refers to a real observation which could be potentially drawn from the target population.

Given the target population of Borman et al. (2008) is both volunteered and non-volunteered schools, a potential observation in Borman et al. (2008) could be either an observation of a classroom (along with the observations of students sat in it) which belonged to a volunteered school in their study or an observation of a classroom (along with the observations of students sat in it) which could be potentially drawn from non-volunteered schools.

Built on previous definitions, the definition of ideal population is formalized next for both the first scenario and the second scenario:

Definition 1.4: **An ideal population** refers to the collection of all possible potential observations of the target population.

The operationalization of this definition depends on the specific context of the research and the scenarios that are discussed earlier. For example, the ideal population of Hong & Raudenbush (2005) is the collection of potential observations of all kindergarten students in the U.S. Likewise, the ideal population of Borman et al. (2008) is the collection of observations of all classrooms in the volunteered and non-volunteered schools.

To understand the bias which invalidates regression-based causal inference in those two scenarios, it’s necessary to decompose an ideal population into an unobserved part and an observed part and differentiate between them. For this purpose, I have the following two definitions:

Definition 1.5.1: **The unobserved part of an ideal population in the first scenario** refers to the collection of all counterfactual observations of the target population. Naturally, **the observed part of an ideal population in the first scenario** refers to the collection of all real observations of the target population.

Definition 1.5.2: **The unobserved or non-representable part of an ideal population in the second scenario** refers to the collection of all potential observations of the part of the target population that cannot be represented by the observed sample. Conversely, **the observed or representable part of an ideal population in the second scenario** refers to the collection of all potential observations of the part of the target population that was deemed to be logically represented by the observed sample.

According to definition 1.5.1, the unobserved part of ideal population of Hong & Raudenbush (2005) is the collection of counterfactual observations of all kindergarten students in the U.S. and the observed part of ideal population of Hong & Raudenbush (2005) is the collection of real observations of all kindergarten students in the U.S. According to definition 1.5.2, the non-representable part of the ideal population of Borman et al. (2008) would be the collection of observations of all classrooms in the non-volunteered schools and the representable part of the ideal population of Borman et al. (2008) would be the collection of observations of all classrooms in the volunteered schools. More importantly, a random sample of the unobserved (non-representable) part of ideal population is the main target of the Bayesian framework and it is defined as an unobserved sample as follows:

Definition 1.6.1: **An unobserved sample in the first scenario** refers to the collection of counterfactual observations of sampled subjects.

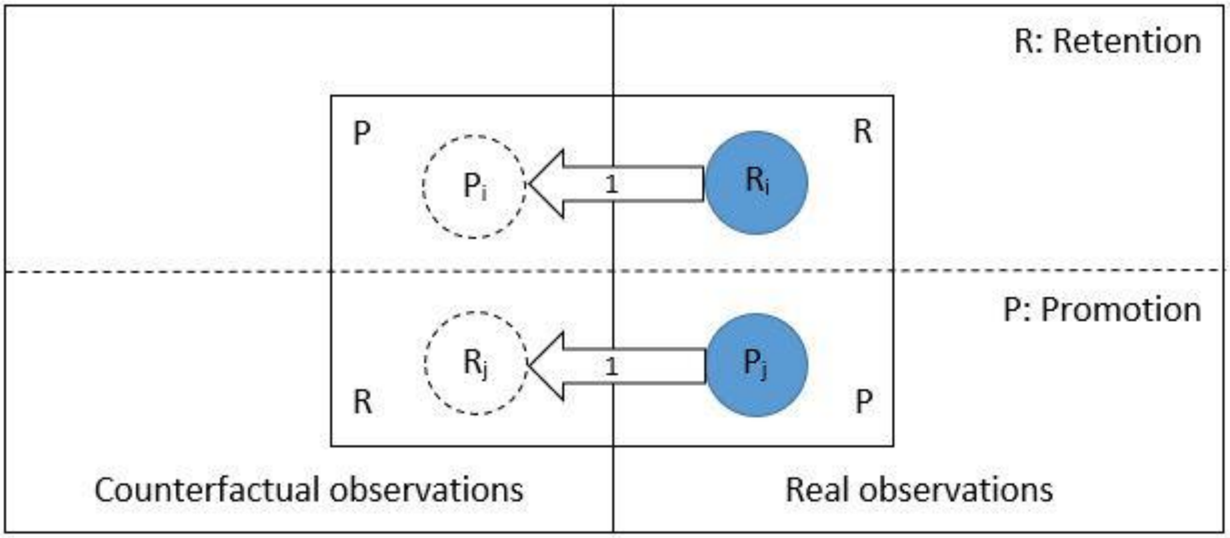
Definition 1.6.2: **An unobserved sample in the second scenario** refers to an imaginary random sample which is drawn from the non-representable part of an ideal population and consists of real observations. I assume a subsequent randomization is carried out on this unobserved sample.

Definition 1.7: **An ideal sample** refers to the combination of the observed sample and an unobserved sample.

Based on definition 1.6.1, an unobserved sample of Hong & Raudenbush (2005) is the collection of all counterfactual observations of their sampled kindergarten students. Even though Hong & Raudenbush (2005) did get a random sample from their target population, i.e., the kindergarten students in America, this unobserved sample were still missing and not ignorable since their sampled students were not randomly retained in kindergarten or promoted to first grade. Furthermore, as argued by Frank et al. (2013), although Hong & Raudenbush (2005) has measured and controlled most relevant covariates, such as kindergarten children's pretest scores, demographical features, psychological qualities, and family backgrounds, they still might leave some significant confounders unmeasured, such as their cognitive abilities and motivations. As a result, this unobserved sample that were missing in Hong & Raudenbush (2005) might not be ignorable conditional on their measured covariates, which equivalently disproves the strong ignorability assumption and poses a threat to its internal validity.

Moreover, based on definition 1.6.2, an unobserved sample of Borman et al. (2008) is an imaginary sample of classrooms which were randomly drawn from the non-volunteered schools. By assumption, classrooms in this unobserved sample had been already randomly assigned to either Open Court Reading group or the control group. I note here that the observed sample of Borman et al. (2008) can only represent the volunteered schools since it came from six schools which were randomly drawn from volunteered schools in their study. The non-volunteered schools, which is an indispensable part of their target population, would be represented by this unobserved sample rather than the observed sample of Borman et al. (2008) and exhibits the discrepancy between their target population of schools and the population of schools can be represented by their sample. Due to this missing unobserved sample, this discrepancy constitutes a nonrandom sampling from their target population and poses a threat to its external validity. Figure 2.1 shows that the observed sample in Hong & Raudenbush (2005) had two groups: retention group (students who were retained in kindergarten) and promotion group (students who were promoted to first grade). For every retained student (the blue-shaded circle R_i), there is a counterfactual observation of his in an unobserved sample (the dashed circle P_i) had he been promoted instead. Similarly, an unobserved sample also keeps the counterfactual observation of every promoted student (the blue-shaded circle P_j) had he been retained instead. An ideal sample is represented by the rectangle formed by conjoining the small rectangle with dashed circles (an unobserved sample) and the small rectangle with solid blue-shaded circles (the observed sample), and it consists of real and counterfactual observations of all sampled students in Hong & Raudenbush (2005). It's remarkable that those two small rectangles adjoin each other, which indicates the observed sample and an unobserved sample refer to the same group of subjects and share the same values of the covariates for those subjects in the first scenario.

Figure 2.1: The structure of ideal population in Hong & Raudenbush (2005)



What figure 2.2 displays is that the ideal population of Borman et al. (2008) is the collection of all real observations of classrooms that could be potentially drawn from American schools. The representable part of this ideal population is the collection of classrooms of schools which volunteered in their research, since the observed sample is the classrooms of schools which were randomly drawn from the volunteered schools. Automatically, the non-representable part of this ideal population is the collection of classrooms of schools which didn’t volunteer in this research. An unobserved sample in this case is thought of as a random sample from the non-representable part of this ideal population. Classrooms of this unobserved sample are thought to be subsequently randomly assigned to the Open Court Reading group or the control group. An ideal sample is the combination of the small rectangle with solid blue-shaded circles (the observed sample) and the small rectangle with solid unshaded circles (an unobserved sample), and it is composed of real observations of classrooms drawn from volunteered and non-volunteered schools. In figure 3 those two small rectangles do not adjoin each other, which reveals that the observed sample and an unobserved sample pertain to different groups of subjects in the second scenario.

Figure 2.2: The structure of ideal population in Borman et al. (2008)

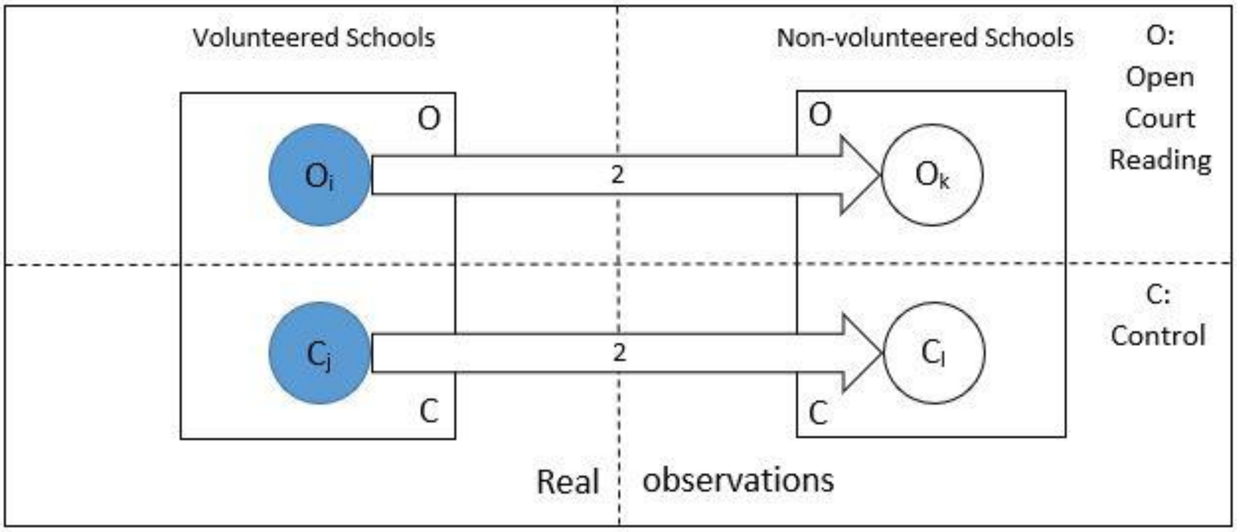
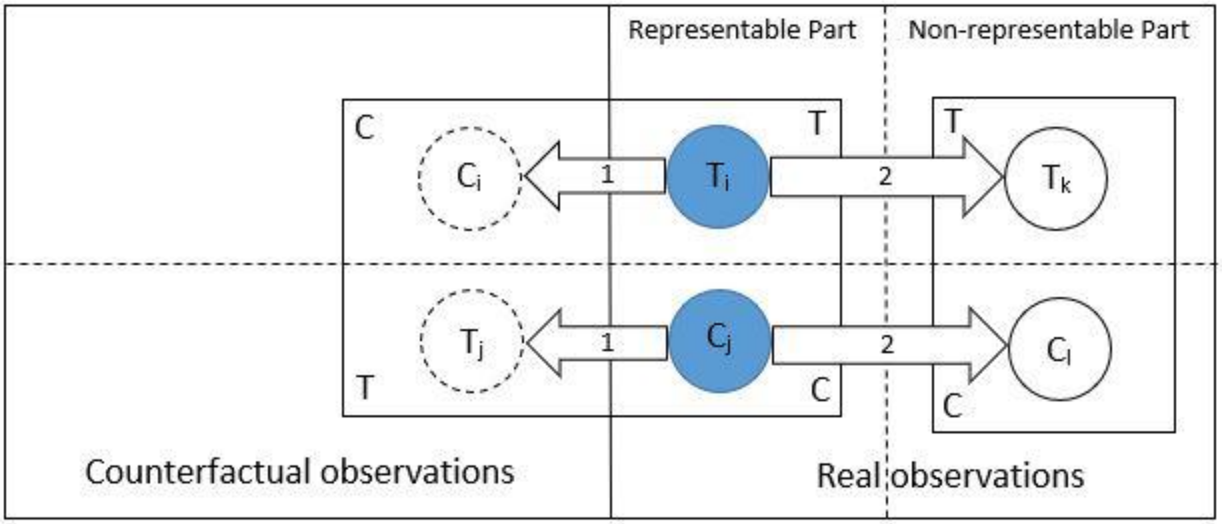


Figure 2.3 synthesizes above two figures and presents the structure of an ideal population in both scenarios. The rectangle which contains two small blue-shaded circles is the observed sample whose upper part is the treatment group (denoted by ‘T’ in the upper-right corner) and lower part is the control group (denoted by ‘C’ in the lower-right corner). In the first scenario, one needs to conceptualize the counterfactual observation of a treated subject T_i as his observation had he participated in the control group, which is represented by a dashed circle C_i. Similarly, the counterfactual observation of a controlled subject C_j is symbolized as a dashed circle T_j which would have been this subject’s observation had he participated in the treatment group. The rectangle contains the dashed circles C_i and T_j is an unobserved sample in the first scenario, and the arrows with a label ‘1’ symbolize the conceptualization of an unobserved sample in the first scenario. The second scenario implicates that the scope of ideal population can be narrowed from both real and counterfactual observations to real observations only because of strong internal validity. Due to limited external validity in the second scenario, the same observed sample with blue-shaded circles T_i and C_j is still problematic as it can only represent the representable part of ideal population. Therefore, we need another conceptualization, symbolized by the arrows with a

label ‘2’, to envision an unobserved sample drawn from the non-representable part of ideal population. This unobserved sample is thought to be formed by first randomly drawing a sample from the non-representable part of ideal population and then randomly assigning subjects to treatment group and control group. In figure 1, a treated subject in this unobserved sample is represented by the solid unshaded circle T_k and a controlled subject in this unobserved sample is represented by the solid unshaded circle C_l.

Figure 2.3: The structure of ideal population in both scenarios



The above definitions and arguments have demonstrated that a missing unobserved sample is mainly responsible for the bias that invalidates regression-based causal inference in the first and second scenarios. To quantify the robustness of regression-based causal inference, it’s inevitable to conceptualize an unobserved sample and shape this conceptualization into a proper statistical model.

1.3-Research objectives

This research is motivated by Frank & Min (2007), which first proposed a Bayesian framework to address the concern of limited external validity, by defining a prior distribution in terms of an unobserved sample and a likelihood in terms of the observed sample. They suggested that, defining a Bayesian model in this fashion would lead to a posterior distribution that is built upon an ideal sample which is just the combination of an unobserved sample and the observed sample. Following their argument, I develop a comprehensive Bayesian framework to address both concerns of limited internal validity and limited external validity for regression-based causal inference, as have been summarized as the two scenarios. Grounded in the philosophy of robustness indices, this Bayesian framework of robustness indices considers a prior as if it is built on an unobserved sample and then purposes a posterior probability of invalidating an inference conditional on the observed sample.

This paper is structured as follows: The second section I formalize a unifying framework of robustness indices, which has a frequentist recipe and a Bayesian recipe, for regression-based causal inference. In the third section, I discuss in general and in depth how to fit the Bayesian framework of robustness indices to raw data for regression-based causal inference. In the fourth section, I demonstrate that the Bayesian framework of robustness indices can be greatly simplified for centered and standardized data. The fifth section addresses an issue of adjusting the statistical threshold $\delta^{\#}$ to the sampling perspective of adding an unobserved sample to the observed sample, by identifying the standard deviation of estimate based on an ideal sample or by considering replacing a portion of the observed sample with an unobserved sample rather than simply adding an unobserved sample to the observed sample. The sixth section provides the detailed applications of my framework to Borman et al. (2008) as well as Hong & Raudenbush (2005). The seventh section concludes this paper with a review of the Bayesian framework of robustness indices for regression-based causal inference and other comparable approaches and a conclusion.

2-The unifying framework of robustness indices for regression-based causal inference

2.1-Setting and notation

The entire discussion in this paper is restricted to the following setting: Every sample, regardless of whether it's observed or not, should contain a vector of outcomes denoted by **Y** and a matrix of predictors denoted by **X**. **X** should have p+2 columns and this means the number of predictors is p+2. Among those p+2 predictors, there is one variable containing all 1s in its data entry to represent the intercept and a treatment indicator denoted by W. All remaining p predictors are pure covariates, and I label them as $Z_1, Z_2, \ldots, Z_p$ respectively. Moreover, W=1 means a subject receives/selects the treatment, for example, the Open Court Reading curriculum or kindergarten retention. Accordingly, W=0 refers to the case that a subject receives or selects the control group. There are only two possible groups in this setting, i.e., treatment and control groups. The outcome Y should be continuous, or at least not a categorical variable. When a data of such structure is available to researchers, I assume they conduct regression-based causal inference to estimate the average treatment effect, which is parameterized as the regression coefficient of W in the regression of **Y** on **X**. For example, this regression can be written as $Y_i = \gamma_0 + \gamma_w W_i + \mathbf{Z_i}\boldsymbol{\gamma}_\mathbf{Z} + \varepsilon_i$ for every individual i in the sample and γ_w is the regression coefficient of W as well as the estimate of average treatment effect.

Under this setting, an ideal population, as well as its unobserved and observed parts, is consisted of observations which all have one value for the outcome **Y** and p+2 values for the predictors. It follows that the observed sample and an unobserved sample are both collections of observations which all have one values for the outcome value and p+2 values for the predictors as well.

The first notation rule I adhere to throughout this paper is that I use a superscript to inform readers about which sample a statistic pertains to. A statistic or a part of data has a superscript as “ob” signifies that it comes from the observed sample, while a superscript “un” indicates that it comes from an unobserved sample. A statistic or a part of data that belongs to an ideal sample, which is just an integration of the observed sample and an unobserved sample and thus can be thought of as a random sample from ideal population, will be labeled with a superscript “id”. For example, $\mathbf{Y^{ob}, Y^{un}, Y^{id}}$ refer to the outcomes **Y** in the observed sample, an unobserved sample and an ideal sample respectively. However, there are some exceptions: A true parameter will not have a superscript. For example, the true regression coefficient of W is symbolized by γ_w. The threshold of regression coefficient W for making an inference is denoted by $\gamma_w^{\#}$. The second notation rule is that a subscript of a sample statistic is used to describe the variable(s) which this sample statistic pertaining to. For an instance, the sample covariance between the treatment indicator W and the outcome Y in the observed sample is denoted by $\hat{\sigma}_{WY}^{ob}$ and its counterpart in an unobserved sample is denoted by $\hat{\sigma}_{WY}^{un}$.

2.2-The frequentist recipe

The logical flow of any robustness indices always starts at the definition of bias for causal inference. The formal definition of bias for regression-based causal inference, i.e., the regression coefficient representing an estimate of the average treatment effect, is:

$$\begin{aligned}\beta = \mathrm{E}[\hat{\gamma}_w^{ob}] - \mathrm{E}[\hat{\gamma}_w^{id}] = \{\mathrm{E}[Y \mid \mathrm{Z}, W=1]^{ob} - \mathrm{E}[Y \mid \mathrm{Z}, W=0]^{ob}\} \\ -\{\mathrm{E}[Y \mid \mathrm{Z}, W=1]^{id} - \mathrm{E}[Y \mid \mathrm{Z}, W=0]^{id}\}\end{aligned} \tag{2.1}$$

In the above definition, $\hat{\gamma}_w^{ob}$ is the estimated regression coefficient of the treatment indicator W based on the observed sample and $\hat{\gamma}_w^{id}$ is the estimated regression coefficient of the treatment indicator W based on an ideal sample. Since the observed sample is random sampled from the observed part of ideal population, $\hat{\gamma}_w^{ob}$ should be unbiased for the average treatment effect conditional on the covariates X in the observed part of ideal population, which is represented by the difference within the first curly bracket. Furthermore, as an ideal sample is conceptualized as a random sample from ideal population, $\hat{\gamma}_w^{id}$ should be unbiased for the average treatment effect conditional on the covariates X in the whole ideal population, which is represented by the difference within the second curly bracket.

The next stage is adjusting the definition (2.1) to an empirical research setting where only the observed sample is available. This implies that the observed sample should be treated as fixed. In addition, I assume an ideal sample and an unobserved sample are both known for this frequentist recipe. In light of this principle, the bias for regression-based causal inference becomes as follows:

$$\beta|\mathbf{Y^{id}},\mathbf{X^{id}} = \hat{\gamma}_w^{ob} - (\gamma_w \mid \mathbf{Y^{id}},\mathbf{X^{id}}) \tag{2.2}$$

The first term appears in the definition (2.2) is $\hat{\gamma}_w^{ob}$ which should be fixed and reported by most empirical research conducting regression-based causal inference. The second term in (2.2) is a random variable characterizing the conditional distribution of regression coefficient of W based on a known ideal sample $(\mathbf{Y^{id}},\mathbf{X^{id}})$. This random variable has taken the random sampling error associated with this known ideal sample into consideration, just like one can derive the distribution of any regression coefficient from a given observed sample to reflect the uncertainty in sampling and thereby conduct a T-test. The randomness of this ideal variable is mostly due to the imaginary nature of an unobserved sample.

According to (2.2) and Frank et al. (2013), an inference will be invalid if:

$$\beta = \hat{\gamma}_w^{ob} - (\gamma_w \mid \mathbf{Y^{id}}, \mathbf{X^{id}}) > \hat{\gamma}_w^{ob} - \gamma_w^{\#} \text{ for inferring a positive effect}$$
$$\beta = \hat{\gamma}_w^{ob} - (\gamma_w \mid \mathbf{Y^{id}}, \mathbf{X^{id}}) < \hat{\gamma}_w^{ob} - \gamma_w^{\#} \text{ for inferring a negative effect}$$

(2.3)

Or equivalently because both $\hat{\gamma}_w^{ob}$ and $\gamma_w^{\#}$ are constants:

$$\begin{aligned} &\gamma_w < \gamma_w^{\#} \mid \mathbf{Y^{id}}, \mathbf{X^{id}} \text{ for inferring a positive effect} \\ &\gamma_w > \gamma_w^{\#} \mid \mathbf{Y^{id}}, \mathbf{X^{id}} \text{ for inferring a negative effect} \end{aligned} \qquad (2.4)$$

Draw on the decision rules (2.4) and the distribution of γ_w^{id}, I propose the following probabilities of invalidating an inference as the robustness indices of regression-based causal inference:

$$\begin{aligned} &P(\gamma_w < \gamma_w^{\#} \mid \mathbf{Y^{id}}, \mathbf{X^{id}}) \text{ for inferring a positive effect} \\ &P(\gamma_w > \gamma_w^{\#} \mid \mathbf{Y^{id}}, \mathbf{X^{id}}) \text{ for inferring a negative effect} \end{aligned} \qquad (2.5)$$

To express the distribution of γ_w conditional on an ideal sample and the probability of invalidating an inference explicitly, I formulate the classical linear regression model (CLRM) for the observed sample next:

$$\begin{aligned} &\mathbf{Y^{ob}} = \mathbf{X^{ob}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^{\mathbf{ob}} \\ &\boldsymbol{\varepsilon}^{\mathbf{ob}} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n^{ob}}) \end{aligned} \qquad (2.6)$$

The CLRM in (2.6) should look familiar for most empirical researchers. For (2.6), the residuals are denoted as $\boldsymbol{\varepsilon}^{\mathbf{ob}}$ and the observed sample size is n^{ob}. Moreover, the residual variance is σ^2 and assumed to be estimated in this context. Based on the CLRM displayed in (2.6), the least square estimates of regression coefficients (i.e., $\hat{\boldsymbol{\gamma}}^{\mathbf{ob}}$) and a multivariate distribution of regression coefficients conditional on this observed sample (i.e., $\boldsymbol{\gamma} \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}}$) can be shown as follows:

$$\begin{aligned}&\hat{\boldsymbol{\gamma}}^{\mathbf{ob}} = ((\mathbf{X}^{\mathbf{ob}})^{\mathbf{T}}\mathbf{X}^{\mathbf{ob}})^{\mathbf{-1}}(\mathbf{X}^{\mathbf{ob}})^{\mathbf{T}}\mathbf{Y}^{\mathbf{ob}}\\&\boldsymbol{\gamma} \,|\, \mathbf{X}^{\mathbf{ob}}, \mathbf{Y}^{\mathbf{ob}} \sim N(((\mathbf{X}^{\mathbf{ob}})^{\mathbf{T}}\mathbf{X}^{\mathbf{ob}})^{\mathbf{-1}}(\mathbf{X}^{\mathbf{ob}})^{\mathbf{T}}\mathbf{Y}^{\mathbf{ob}}, \sigma^2((\mathbf{X}^{\mathbf{ob}})^{\mathbf{T}}\mathbf{X}^{\mathbf{ob}})^{\mathbf{-1}})\end{aligned} \tag{2.7}$$

Analogously, the CLRM for an unobserved sample, the least square estimates of regression coefficients for this unobserved sample and the distribution of regression coefficients conditional on this unobserved sample are formulated as below:

$$\begin{aligned}&\mathbf{Y}^{\mathbf{un}} = \mathbf{X}^{\mathbf{un}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^{\mathbf{un}}\\&\boldsymbol{\varepsilon}^{\mathbf{un}} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}_{n^{un}})\\&\hat{\boldsymbol{\gamma}}^{\mathbf{un}} = ((\mathbf{X}^{\mathbf{un}})^{\mathbf{T}}\mathbf{X}^{\mathbf{un}})^{\mathbf{-1}}(\mathbf{X}^{\mathbf{un}})^{\mathbf{T}}\mathbf{Y}^{\mathbf{un}}\\&\boldsymbol{\gamma} \,|\, \mathbf{X}^{\mathbf{un}}, \mathbf{Y}^{\mathbf{un}} \sim N(((\mathbf{X}^{\mathbf{un}})^{\mathbf{T}}\mathbf{X}^{\mathbf{un}})^{\mathbf{-1}}(\mathbf{X}^{\mathbf{un}})^{\mathbf{T}}\mathbf{Y}^{\mathbf{un}}, \sigma^2((\mathbf{X}^{\mathbf{un}})^{\mathbf{T}}\mathbf{X}^{\mathbf{un}})^{\mathbf{-1}})\end{aligned} \tag{2.8}$$

It’s remarkable that this unobserved sample has a sample size n^{un} which is likely to be different from the observed sample size. However, the residual variance for this CLRM is still σ^2, which hints that the residual variances in CLRMs for both observed sample and unobserved sample are equal. This is a core assumption for the derivation of robustness indices and therefore maintained throughout this paper.

Finally, the ideal sample is formed by combining the observed and unobserved samples mentioned in above CLRMs and again the CLRM with regard to this ideal sample can be defined similarly as in (2.6) through (2.8):

$$\begin{aligned}&\mathbf{Y}^{\mathbf{id}} = \mathbf{X}^{\mathbf{id}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^{\mathbf{id}}\\&\boldsymbol{\varepsilon}^{\mathbf{id}} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}_{(n^{un}+n^{ob})})\\&\hat{\boldsymbol{\gamma}}^{\mathbf{id}} = ((\mathbf{X}^{\mathbf{id}})^{\mathbf{T}}\mathbf{X}^{\mathbf{id}})^{\mathbf{-1}}(\mathbf{X}^{\mathbf{id}})^{\mathbf{T}}\mathbf{Y}^{\mathbf{id}}\\&\boldsymbol{\gamma} \,|\, \mathbf{X}^{\mathbf{id}}, \mathbf{Y}^{\mathbf{id}} \sim N(((\mathbf{X}^{\mathbf{id}})^{\mathbf{T}}\mathbf{X}^{\mathbf{id}})^{\mathbf{-1}}(\mathbf{X}^{\mathbf{id}})^{\mathbf{T}}\mathbf{Y}^{\mathbf{id}}, \sigma^2((\mathbf{X}^{\mathbf{id}})^{\mathbf{T}}\mathbf{X}^{\mathbf{id}})^{\mathbf{-1}})\end{aligned} \tag{2.9}$$

where the data matrices of $\mathbf{Y^{id}}$ and $\mathbf{X^{id}}$ can be understood as block matrices which have the following structures:

$$\mathbf{X^{id}} = \begin{bmatrix} \mathbf{X^{un}} \\ \mathbf{X^{ob}} \end{bmatrix}_{(n^{un}+n^{ob})\times(\mathrm{p}+2)}$$

$$\mathbf{Y^{id}} = \begin{bmatrix} \mathbf{Y^{un}} \\ \mathbf{Y^{ob}} \end{bmatrix}_{(n^{un}+n^{ob})\times 1} \tag{2.10}$$

Therefore, the probability of invalidating an inference can be readily computed from the distribution of $\boldsymbol{\gamma} \mid \mathbf{Y^{id}}, \mathbf{X^{id}}$ in (2.9), provided one can randomly draw an unobserved sample from the unobserved part of ideal population. Unfortunately, a robustness index is useful only when an unobserved sample is unapproachable, and this contradicts the premise of the frequentist recipe. For this reason, I turn to the Bayesian recipe next.

2.3-The Bayesian recipe

The Bayesian recipe starts with a modified version of the definition of bias, which defines the bias for regression-based causal inference as the difference between least square estimate of regression coefficient of treatment indicator W and the random variable that follows the posterior distribution of the regression coefficient of W:

$$\beta = \hat{\gamma}_w^{ob} - (\gamma_w \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}) \tag{2.11}$$

The decision rules in the Bayesian recipe for deciding whether an inference is invalid are almost the same as their parallels in the frequentist recipe. An inference will be invalid if one of the following conditions is true:

$$\begin{aligned} \beta &= \hat{\gamma}_w^{ob} - (\gamma_w \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}) > \hat{\gamma}_w^{ob} - \gamma_w^{\#} \text{ for inferring a positive effect} \\ \beta &= \hat{\gamma}_w^{ob} - (\gamma_w \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}) < \hat{\gamma}_w^{ob} - \gamma_w^{\#} \text{ for inferring a negative effect} \end{aligned} \quad (2.12)$$

Consequently, the probability of invalidating an inference is accessible through the posterior distribution $\gamma_w \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}$ and generally they should be:

$$\begin{aligned} &P(\gamma_w < \gamma_w^{\#} \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}) \text{ for inferring a positive effect} \\ &P(\gamma_w > \gamma_w^{\#} \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}) \text{ for inferring a negative effect} \end{aligned} \quad (2.13)$$

In the Bayesian recipe, the bias for regression-based causal inference is built on the observed sample solely. This doesn't result in a discrepancy between the frequentist recipe and Bayesian recipe, so long as one choose to model the prior as the distribution of a focal parameter based on an unobserved sample. To demonstrate this relationship between the two recipes as well as formalize the posterior distribution, the Bayesian model of regression-based causal inference is provided as follows:

$$\begin{aligned} &\boldsymbol{\gamma} \sim N(((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{X^{un}})^{\mathbf{-1}}(\mathbf{X^{un}})^{\mathbf{T}}\mathbf{Y^{un}}, \sigma^2((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{X^{un}})^{\mathbf{-1}}) \\ &Y_i \mid \boldsymbol{\gamma}, \mathbf{X_i} \sim N(\mathbf{X_i}\boldsymbol{\gamma}, \sigma^2) \\ &\boldsymbol{\gamma} \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}} \sim N(\boldsymbol{\theta}_{\gamma}, \boldsymbol{\Phi}_{\gamma}) \end{aligned} \quad (2.14)$$

where:

$$\begin{aligned} \boldsymbol{\theta}_{\gamma} &= ((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{X^{un}} + (\mathbf{X^{ob}})^{\mathbf{T}}\mathbf{X^{ob}})^{\mathbf{-1}}((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{Y^{un}} + (\mathbf{X^{ob}})^{\mathbf{T}}\mathbf{Y^{ob}}) \\ \boldsymbol{\Phi}_{\gamma} &= \sigma^2((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{X^{un}} + (\mathbf{X^{ob}})^{\mathbf{T}}\mathbf{X^{ob}})^{\mathbf{-1}} \end{aligned} \quad (2.15)$$

There is nothing special about the formulation of this Bayesian model except the parameterization of the prior distribution. The prior distribution in (2.14) is identical to the distribution of regression coefficients conditional on an unobserved sample, as specified in (2.8).

This is in accordance with the Bayesian framework propounded by Frank & Min (2007), given that the prior is defined as a distribution of regression coefficients conditional on an unobserved sample and the likelihood function will only be fit to the observed sample. The term σ^2 which denotes residual variance appears in both prior and likelihood function, which reflects the assumption that the classical linear regression models underlying the prior and likelihood function are restricted to have the same known residual variance. Most interestingly, the following equations are established inasmuch as (2.10) uncovers that the data matrices contained in an ideal sample can be written as block matrices:

$$\begin{aligned}(\mathbf{X^{id}})^{\mathbf{T}}\mathbf{X^{id}} &= (\mathbf{X^{un}})^{\mathbf{T}}\mathbf{X^{un}} + (\mathbf{X^{ob}})^{\mathbf{T}}\mathbf{X^{ob}}\\ (\mathbf{X^{id}})^{\mathbf{T}}\mathbf{Y^{id}} &= (\mathbf{X^{un}})^{\mathbf{T}}\mathbf{Y^{un}} + (\mathbf{X^{ob}})^{\mathbf{T}}\mathbf{Y^{ob}}\end{aligned} \qquad (2.16)$$

Once the results in (2.16) are plugged into the expressions of posterior mean and variance in (2.15), the posterior distribution becomes:

$$\boldsymbol{\gamma}\,|\,\mathbf{X^{ob}},\mathbf{Y^{ob}} \sim N(((\mathbf{X^{id}})^{\mathbf{T}}\mathbf{X^{id}})^{\mathbf{-1}}(\mathbf{X^{id}})^{\mathbf{T}}\mathbf{Y^{id}},\sigma^2((\mathbf{X^{id}})^{\mathbf{T}}\mathbf{X^{id}})^{\mathbf{-1}}) \qquad (2.17)$$

What (2.17) uncovers is that the posterior distribution will be identical to the distribution of regression coefficients conditional on an ideal sample when one parameterizes the prior distribution as if it is a distribution of regression coefficients conditional on an unobserved sample. However, I caution readers that the frequentist recipe and the Bayesian recipe are conceptually and empirically distinct even though they both arrive at the same model of robustness indices. Conceptually, the frequentist recipe requires an unobserved sample to be a real one while the Bayesian recipe only considers an unobserved sample as one's belief which is subjective and shapes this belief into a prior distribution. Empirically, the frequentist approach is unfeasible since no unobserved sample is available whereas the Bayesian approach is practical in the sense that one only needs to transform his belief about an unobserved sample into a prior distribution so as to make the corresponding posterior distribution qualified as a distribution based on an imaginary ideal sample. It’s remarkable that oftentimes a belief about an unobserved sample should be constantly changing instead of fixed, and in this case the learning goal of the Bayesian recipe is to determine the relationship between the probability of invalidating an inference and the prior parameters, which will be the theme of subsequent sections.

3-Bayesian models of robustness indices for raw data

3.1-Data and the sample statistics

This section primarily focuses on the derivation of posterior distribution discussed earlier in the unifying Bayesian framework as if we have collected raw data for both unobserved sample and observed sample. By saying “raw data is collected for an unobserved sample”, I point to the construction of prior distribution which is conceptualized as the distribution of regression coefficients conditional on the imaginary raw data for this unobserved sample. The target is to express the posterior mean and variance as functions of sample statistics built on either an imaginary unobserved sample or the observed sample. To be aligned with the settings of earlier discussion, the raw data for the observed sample should be in the following form:

$$\begin{aligned}
\mathbf{D^{ob}} &= [\mathbf{Y}_{n^{ob}\times 1}, \mathbf{X}_{n^{ob}\times (p+2)}] \\
\mathbf{X^{ob}} &= [\mathbf{1}_{n^{ob}\times 1}, \mathbf{V}_{n^{ob}\times (p+1)}] \\
\mathbf{V^{ob}} &= [\mathbf{Z}_{n^{ob}\times p}, \mathbf{W}_{n^{ob}\times 1}] \\
\mathbf{Z^{ob}} &= [\mathbf{Z_1}, \mathbf{Z_2}, ..., \mathbf{Z_p}]_{n^{ob}\times p}
\end{aligned} \tag{3.1}$$

Some notations need explanations in (3.1): **D** refers to the whole data and it is composed of a data vector of outcome **Y** and a data matrix of all the predictors **X**. The data matrix **X** has two parts: The first part is a constant vector **1** and the second part is the group of predictors **V**. I further decompose **V** into a data matrix **Z** which only includes the pure covariates and the vector of treatment indicators **W**. The matrix **Z** will have p columns, which symbolizes that there are p pure covariates in the raw data.

By analogy, the raw data for an imaginary unobserved sample is structured as follows:

$$\begin{aligned}
\mathbf{D^{un}} &= [\mathbf{Y}_{n^{un}\times 1}, \mathbf{X}_{n^{un}\times(p+2)}] \\
\mathbf{X^{un}} &= [\mathbf{1}_{n^{un}\times 1}, \mathbf{V}_{n^{un}\times(p+1)}] \\
\mathbf{V^{un}} &= [\mathbf{Z}_{n^{un}\times p}, \mathbf{W}_{n^{un}\times 1}] \\
\mathbf{Z^{un}} &= [\mathbf{Z_1}, \mathbf{Z_2}, ..., \mathbf{Z_p}]_{n^{un}\times p}
\end{aligned} \tag{3.2}$$

Finally, the observed sample and an imaginary unobserved sample are consolidated to create an ideal sample which is styled as below (see (2.10) for a reference):

$$\begin{aligned}
\mathbf{D^{id}} &= [\mathbf{Y}_{(n^{un}+n^{ob})\times 1}, \mathbf{X}_{(n^{un}+n^{ob})\times(p+2)}] \\
\mathbf{X^{id}} &= [\mathbf{1}_{(n^{un}+n^{ob})\times 1}, \mathbf{V}_{(n^{un}+n^{ob})\times(p+1)}] \\
\mathbf{V^{id}} &= [\mathbf{Z}_{(n^{un}+n^{ob})\times p}, \mathbf{W}_{(n^{un}+n^{ob})\times 1}] \\
\mathbf{Z^{id}} &= [\mathbf{Z_1}, \mathbf{Z_2}, ..., \mathbf{Z_p}]_{(n^{un}+n^{ob})\times p}
\end{aligned} \tag{3.3}$$

Next some key sample statistics are introduced based on aforementioned raw data forms for unobserved, observed and ideal samples. First, I define the sample mean vectors for unobserved, observed and ideal samples as:

$$
\begin{aligned}
\bar{\mathbf{X}}^{\mathbf{id}} &= [1, \bar{Z}_1^{id}, \bar{Z}_2^{id}, \cdots, \bar{Z}_p^{id}, \bar{W}^{id}]_{1\times(p+2)} \\
\bar{\mathbf{Z}}^{\mathbf{id}} &= [\bar{\mathrm{Z}}_1^{id}, \bar{\mathrm{Z}}_2^{id}, \cdots, \bar{\mathrm{Z}}_p^{id}]_{1\times p} \\
\bar{\mathbf{X}}^{\mathbf{un}} &= [1, \bar{Z}_1^{un}, \bar{Z}_2^{un}, \cdots, \bar{Z}_p^{un}, \bar{W}^{un}]_{1\times(p+2)} \\
\bar{\mathbf{Z}}^{\mathbf{un}} &= [\bar{\mathrm{Z}}_1^{un}, \bar{\mathrm{Z}}_2^{un}, \cdots, \bar{\mathrm{Z}}_p^{un}]_{1\times p} \\
\bar{\mathbf{X}}^{\mathbf{ob}} &= [1, \bar{Z}_1^{ob}, \bar{Z}_2^{ob}, \cdots, \bar{Z}_p^{ob}, \bar{W}^{ob}]_{1\times(p+2)} \\
\bar{\mathbf{Z}}^{\mathbf{ob}} &= [\bar{\mathrm{Z}}_1^{ob}, \bar{\mathrm{Z}}_2^{ob}, \cdots, \bar{\mathrm{Z}}_p^{ob}]_{1\times p}
\end{aligned}
\tag{3.4}
$$

Recall that my notation rule specifies that a superscript of a statistical term is used to denote the kind of sample which it is computed based on and a subscript of a statistical term is used to denote the variable(s) it pertains to. To abide by this notation rule, the variance-covariance matrix of all the predictors in **V** for an ideal sample will be fashioned as follows:

$$
\mathbf{S}_{\mathbf{VV}}^{\mathbf{id}} = \begin{pmatrix} \mathbf{S}_{\mathbf{ZZ}}^{\mathbf{id}} & \mathbf{S}_{\mathbf{ZW}}^{\mathbf{id}} \\ \mathbf{S}_{\mathbf{WZ}}^{\mathbf{id}} & \hat{\sigma}_{WW}^{id} \end{pmatrix}_{(p+1)\times(p+1)}
\tag{3.5}
$$

where:

$$
\begin{aligned}
\mathbf{S}_{\mathbf{ZZ}}^{\mathbf{id}} &= \begin{pmatrix} \hat{\sigma}_{Z_1Z_1}^{id} & \cdots & \hat{\sigma}_{Z_1Z_p}^{id} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{Z_pZ_1}^{id} & \cdots & \hat{\sigma}_{Z_pZ_p}^{id} \end{pmatrix}_{p\times p} \\
\mathbf{S}_{\mathbf{ZW}}^{\mathbf{id}} &= \begin{pmatrix} \hat{\sigma}_{Z_1W}^{id} \\ \vdots \\ \hat{\sigma}_{Z_pW}^{id} \end{pmatrix}_{p\times 1} \\
\mathbf{S}_{\mathbf{WZ}}^{\mathbf{id}} &= \begin{pmatrix} \hat{\sigma}_{Z_1W}^{id} & \cdots & \hat{\sigma}_{Z_pW}^{id} \end{pmatrix}_{1\times p}
\end{aligned}
\tag{3.6}
$$

Likewise, the vector of covariances between predictors in **V** and the outcome Y for an ideal sample is generically written as below:

$$\mathbf{S}_{\mathbf{VY}}^{\mathbf{id}} = \begin{pmatrix} \mathbf{S}_{\mathbf{ZY}}^{\mathbf{id}} \\ \hat{\sigma}_{WY}^{id} \end{pmatrix}_{(p+1)\times 1} \tag{3.7}$$

where:

$$\mathbf{S}_{\mathbf{ZY}}^{\mathbf{id}} = \begin{pmatrix} \hat{\sigma}_{Z_1 Y}^{id} \\ \vdots \\ \hat{\sigma}_{Z_p Y}^{id} \end{pmatrix}_{p\times 1} \tag{3.8}$$

All aforementioned sample covariances and variances are supposed to be computed according to the following formula:

$$\begin{aligned} \hat{\sigma}_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ \hat{\sigma}_{xx} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \end{aligned} \tag{3.9}$$

I emphasize here that the small x, small y and small n in (3.9) are all symbolic and their numeric values depend on the actual context. The small x could represent any variable in **D** except the constant vector **1** and the small y, when calculating a covariance, could be any variable in **D** other than the actual variable represented by small x. The small n is the size of a sample, which could possibly be an unobserved one, the observed one, or combinatively an ideal one. For example, the treatment indicator W could be the small x and the outcome Y could be the small y and consequently one would obtain the covariances between W and Y for an ideal sample, by replacing the sample size n with the actual ideal sample size $n^{ob} + n^{un}$.

In summary, the notation rule will generally guide readers about the interpretation of a statistical term, especially a sample variance or covariance, that later appears in this paper. Although only the variance-covariance matrix of **V** and covariance vector between **V** and **Y** for an ideal sample is discussed in (3.5) through (3.8), one should recognize that he can write down the variance-covariance matrix of **V** and covariances between **V** and **Y** almost identically for an unobserved sample or the observed sample and the only change he needs is to modify the superscript of every variance and covariance term.

3.2-The posterior distribution of γ_w for raw data

The posterior distribution of γ_w, i.e., the regression coefficient of the treatment indicator W, is of paramount value for regression-based causal inference since it is the ground on which the inference of a true average treatment effect is carried out. The following theorem will bridge this posterior distribution and the sample statistics by recasting its mean and variance as functions of sample variances and covariances for unobserved and observed samples:

Theorem 1. Suppose the CLRMs for unobserved, observed and ideal samples as presented in (2.6) through (2.9) are true and the raw data is in the same format as what I have outlined in (3.1) through (3.3), the posterior distribution of γ_w within the Bayesian framework proposed in (2.14) through (2.17) will be as follows:

$$\gamma_w \mid \mathbf{X}^{\mathbf{ob}}, \mathbf{Y}^{\mathbf{ob}} \sim N\left(\frac{\hat{\sigma}_{WY}^{id} - \mathbf{S}_{\mathbf{WZ}}^{\mathbf{id}}(\mathbf{S}_{\mathbf{ZZ}}^{\mathbf{id}})^{\mathbf{-1}}\mathbf{S}_{\mathbf{ZY}}^{\mathbf{id}}}{\hat{\sigma}_{WW}^{id} - \mathbf{S}_{\mathbf{WZ}}^{\mathbf{id}}(\mathbf{S}_{\mathbf{ZZ}}^{\mathbf{id}})^{\mathbf{-1}}\mathbf{S}_{\mathbf{ZW}}^{\mathbf{id}}}, \frac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S}_{\mathbf{WZ}}^{\mathbf{id}}(\mathbf{S}_{\mathbf{ZZ}}^{\mathbf{id}})^{\mathbf{-1}}\mathbf{S}_{\mathbf{ZW}}^{\mathbf{id}})^{\mathbf{-1}}\right) \quad (3.10)$$

where:

$$
\begin{aligned}
\hat{\sigma}^{id}_{Z_iZ_j} &= \frac{n^{un}\hat{\sigma}^{un}_{Z_iZ_j} + n^{ob}\hat{\sigma}^{ob}_{Z_iZ_j}}{n^{un}+n^{ob}} + \frac{n^{un}\overline{Z}^{un}_i\overline{Z}^{un}_j + n^{ob}\overline{Z}^{ob}_i\overline{Z}^{ob}_j}{n^{un}+n^{ob}} - \overline{Z}^{id}_i\overline{Z}^{id}_j \quad \text{for } i \neq j = 1,2,\ldots,p \\
\hat{\sigma}^{id}_{WZ_i} &= \frac{n^{un}\hat{\sigma}^{un}_{WZ_i} + n^{ob}\hat{\sigma}^{ob}_{WZ_i}}{n^{un}+n^{ob}} + \frac{n^{un}\overline{W}^{un}\overline{Z}^{un}_i + n^{ob}\overline{W}^{ob}\overline{Z}^{ob}_i}{n^{un}+n^{ob}} - \overline{W}^{id}\overline{Z}^{id}_i \quad \text{for } i = 1,2,\ldots,p \\
\hat{\sigma}^{id}_{WW} &= \frac{n^{un}\hat{\sigma}^{un}_{WW} + n^{ob}\hat{\sigma}^{ob}_{WW}}{n^{un}+n^{ob}} + \frac{n^{un}(\overline{W}^{un})^2 + n^{ob}(\overline{W}^{ob})^2}{n^{un}+n^{ob}} - (\overline{W}^{id})^2 \\
\hat{\sigma}^{id}_{Z_iY} &= \frac{n^{un}\hat{\sigma}^{un}_{Z_iY} + n^{ob}\hat{\sigma}^{ob}_{Z_iY}}{n^{un}+n^{ob}} + \frac{n^{un}\overline{Z}^{un}_i\overline{Y}^{un} + n^{ob}\overline{Z}^{ob}_i\overline{Y}^{ob}}{n^{un}+n^{ob}} - \overline{Z}^{id}_i\overline{Y}^{id} \quad \text{for } i = 1,2,\ldots,p \\
\hat{\sigma}^{id}_{WY} &= \frac{n^{un}\hat{\sigma}^{un}_{WY} + n^{ob}\hat{\sigma}^{ob}_{WY}}{n^{un}+n^{ob}} + \frac{n^{un}\overline{W}^{un}\overline{Y}^{un} + n^{ob}\overline{W}^{ob}\overline{Y}^{ob}}{n^{un}+n^{ob}} - \overline{W}^{id}\overline{Y}^{id}
\end{aligned}
\tag{3.11}
$$

(Proof in Appendix A; Additional proof of the equivalence between theorem 1 and some common expressions of regression coefficients is provided in Appendix B, to demonstrate how the Bayesian paradigm of robustness indices is connected to regression coefficients, semi-correlations and partial correlations).

The equation below will serve as the instruction of computing the ideal sample means appear in (3.11):

$$
\begin{aligned}
\overline{Z}^{id}_i &= \frac{n^{un}\overline{Z}^{un}_i + n^{ob}\overline{Z}^{ob}_i}{n^{un}+n^{ob}} \quad \text{for } i = 1,\ 2,\ldots,p \\
\overline{W}^{id} &= \frac{n^{un}\overline{W}^{un} + n^{ob}\overline{W}^{ob}}{n^{un}+n^{ob}} \\
\overline{Y}^{id} &= \frac{n^{un}\overline{Y}^{un} + n^{ob}\overline{Y}^{ob}}{n^{un}+n^{ob}}
\end{aligned}
\tag{3.12}
$$

The formula in (3.11) will have a tidier form, as described next (see details in Frank & Min (2007) Appendix):

$$
\begin{aligned}
&\hat{\sigma}^{id}_{Z_iZ_j} = \lambda\hat{\sigma}^{un}_{Z_iZ_j} + (1-\lambda)\hat{\sigma}^{ob}_{Z_iZ_j} + (1-\lambda)\lambda(\bar{Z}^{ob}_i - \bar{Z}^{un}_i)(\bar{Z}^{ob}_j - \bar{Z}^{un}_j) \text{ for } i \neq j = 1,2,\ldots,p \\
&\hat{\sigma}^{id}_{WZ_i} = \lambda\hat{\sigma}^{un}_{WZ_i} + (1-\lambda)\hat{\sigma}^{ob}_{WZ_i} + (1-\lambda)\lambda(\bar{W}^{ob} - \bar{W}^{un})(\bar{Z}^{ob}_i - \bar{Z}^{un}_i) \text{ for } i = 1,2,\ldots,p \\
&\hat{\sigma}^{id}_{WW} = \lambda\hat{\sigma}^{un}_{WW} + (1-\lambda)\hat{\sigma}^{ob}_{WW} + (1-\lambda)\lambda(\bar{W}^{ob} - \bar{W}^{un})^2 \\
&\hat{\sigma}^{id}_{Z_iY} = \lambda\hat{\sigma}^{un}_{Z_iY} + (1-\lambda)\hat{\sigma}^{ob}_{Z_iY} + (1-\lambda)\lambda(\bar{Z}^{ob}_i - \bar{Z}^{un}_i)(\bar{Y}^{ob} - \bar{Y}^{un}) \text{ for } i = 1,2,\ldots,p \\
&\hat{\sigma}^{id}_{WY} = \lambda\hat{\sigma}^{un}_{WY} + (1-\lambda)\hat{\sigma}^{ob}_{WY} + (1-\lambda)\lambda(\bar{W}^{ob} - \bar{W}^{un})(\bar{Y}^{ob} - \bar{Y}^{un})
\end{aligned}
$$

(3.13)

where:

$$\lambda = \frac{n^{un}}{n^{un} + n^{ob}} \tag{3.14}$$

The above formula, equations and distribution constitute the entire theorem 1, which unveils a cardinal perspective on the evaluation of robustness of regression-based causal inference: First, as suggested by (2.17), the posterior distribution of γ_w could be conceptualized as a distribution of γ_w when a whole ideal sample is available. However, such conceptualization is hinged on two assumptions. The first one is that a CLRM assumption could be made for observed, unobserved and ideal samples. The second one is the residual variances in the CLRMs for observed, unobserved and ideal samples are all equal and known. Theorem 1 shows that, the mean and variance of the posterior distribution of γ_w are functions of sample variances and covariances for an ideal sample, which can be further expressed as functions of sample means, variances and covariances for observed and unobserved samples. Ultimately, by fixing the observed sample statistics (means, variances, covariances and sample size) as well as the residual variance, the posterior mean and variance of γ_w should be functions of unobserved sample statistics, such as the size, means, variances and covariances for an unobserved sample. Essentially, those unobserved sample statistics are all parameters of the prior distribution of the unifying Bayesian framework proposed in the last section, and changing values of the unobserved sample statistics will result in variations in the posterior distribution. Such logic of the analysis of robustness is in line with Bayesian sensitivity analysis which purposes checking the influence of prior distribution on posterior distribution.

3.3-Probit models for the probability of invalidating an inference

I propound the probability of invalidating an inference, which is based on the posterior distribution of γ_w offered by theorem 1, as the robustness index for regression-based causal inference. Recall that the probability of invalidating an inference is either the posterior probability of $\gamma_w < \gamma_w^{\#}$ when inferring a positive effect or the posterior probability of $\gamma_w > \gamma_w^{\#}$ when inferring a negative effect. Given the posterior distribution of γ_w is normal with mean and variance as definitive functions of sample statistics, the probability of invalidating an inference is expected to be a probit function of the sample statistics that shows in (3.10). For this reason, I turn to the next theorem:

Theorem 2. Assume the CLRMs for unobserved, observed and ideal samples that are shown in (2.6) through (2.9) are true and the raw data conforms to the structure defined in (3.1) through (3.3). Moreover, I assume both the threshold of making a decision $\gamma_w^{\#}$ and the common residual variance σ^2 shared by all CLRMs are given. Then the following probit models are true for the probability of invalidating an inference (denoted by p):

For inferring a positive effect:

$$probit(p) = \frac{\sqrt{n^{un}+n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{-1}\mathbf{S_{ZW}^{id}}}}[\gamma_w^{\#}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{-1}\mathbf{S_{ZW}^{id}}) - (\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{-1}\mathbf{S_{ZY}^{id}})]$$

(3.15)

For inferring a negative effect:

$$probit(p) = \frac{\sqrt{n^{un}+n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{-1}\mathbf{S_{ZW}^{id}}}}[(\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{-1}\mathbf{S_{ZY}^{id}}) - \gamma_w^{\#}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{-1}\mathbf{S_{ZW}^{id}})]$$

(3.16)

(Proof in Appendix).

Although the sample variances and covariances in the probit models above are all based on an ideal sample, they are in fact functions of sample means, variances and covariances based on unobserved and observed samples, as manifested by (3.13). The analytical strategy for the probit models above is that I only isolate a small number of unobserved sample statistics as focal parameters in the analysis of robustness while holding all other observed and unobserved sample statistics as fixed. The probit models above may turn out to be a linear or nonlinear function of the focal parameters, depending on the choice of focal parameters. For example, the probit models are linear functions of $\hat{\sigma}_{WY}^{un}$ if all other unobserved and observed sample statistics are held constant. An exemplary question to ask in this case would be "what does $\hat{\sigma}_{WY}^{un}$ need to be in order to make the probability of invalidating an inference smaller than a certain number (say 0.5), holding all other unobserved and observed sample statistics as fixed?". The ensuing subsection will detail this analytical strategy and discuss its implications for the two scenarios described at the beginning.

3.4-External validity and internal validity

It’s important to choose a small but appropriate subset of parameters (as focal parameters) from all the terms of sample statistics in the probit models. There are three main reasons for doing so: First, in some cases, some unobserved sample statistics are more meaningful and therefore more suitable choices as focal parameters than other unobserved sample statistics. For example, we might be more interested in the covariance between pretest scores and Open Court Reading (OCR) curriculum as well as the covariance between posttest scores and OCR in a possibly sample of classrooms randomly drawn from the non-volunteered schools, as pretest is the main covariate in the model and posttest is the outcome. We might, instead, focus on the parameter λ, i.e., the proportion of unobserved classrooms in an ideal sample of classrooms, while assuming the covariances between posttest scores and OCR as well as between pretest scores and OCR are both 0, which mimics a research question addressed by Frank et al. (2013). Second, one should recognize that the number of possible parameters (unobserved sample statistics) will quadratically increase as the number of variables increase. I must point out, that my approach requires one to analyze one focal parameter at a time while holding all others as fixed and to subsequently report the thresholds of invalidating an inference at some levels of probabilities for this focal parameter. Therefore, it’s impossible to learn all possible unobserved sample statistics as focal parameters as it will make this analysis too complex to conduct and understand.

The last reason is about the preferences and constraints on parameters in the two scenarios where regression-based causal inference likely fails. In the first scenario, a targeted research is usually a quasi-experiment or observational study which lacks randomized assignment to groups. Recall that an unobserved sample in this scenario is defined as the collection of counterfactual observations of all the subjects in the observed sample. A counterfactual observation, according to definition 1.2, should be an observation in which everything is the same as the original observation except its values of the treatment indicator and the outcome. Therefore, an unobserved sample has a distinctive structure as shown below:

If the observed sample in the first scenario has the following structure:

$$\begin{aligned}
\mathbf{D^{ob}} &= [\mathbf{Y}_{n^{ob}\times 1}, \mathbf{X}_{n^{ob}\times(p+2)}] \\
\mathbf{X^{ob}} &= [\mathbf{1}_{n^{ob}\times 1}, \mathbf{V}_{n^{ob}\times(p+1)}] \\
\mathbf{V^{ob}} &= [\mathbf{Z}^{ob}_{n^{ob}\times p}, \mathbf{W}^{ob}_{n^{ob}\times 1}] \\
\mathbf{Z^{ob}} &= [\mathbf{Z_1}, \mathbf{Z_2}, ..., \mathbf{Z_p}]_{n^{ob}\times p}
\end{aligned} \tag{3.17}$$

An unobserved sample will be in the following form:

$$\begin{aligned}
\mathbf{D^{un}} &= [\mathbf{Y}_{n^{ob}\times 1}, \mathbf{X}_{n^{ob}\times(p+2)}] \\
\mathbf{X^{un}} &= [\mathbf{1}_{n^{ob}\times 1}, \mathbf{V}_{n^{ob}\times(p+1)}] \\
\mathbf{V^{un}} &= [\mathbf{Z}^{ob}_{n^{ob}\times p}, \mathbf{1}_{n^{ob}\times 1} - \mathbf{W}^{ob}_{n^{ob}\times 1}] \\
\mathbf{Z^{ob}} &= [\mathbf{Z_1}, \mathbf{Z_2}, ..., \mathbf{Z_p}]_{n^{ob}\times p}
\end{aligned} \tag{3.18}$$

As a result, an ideal sample is formed by stacking the data matrix of unobserved sample over the data matrix of observed sample:

$$\begin{aligned}
\mathbf{D^{id}} &= [\mathbf{Y}_{2n^{ob}\times 1}, \mathbf{X}_{2n^{ob}\times(p+2)}] \\
\mathbf{X^{id}} &= [\mathbf{1}_{2n^{ob}\times 1}, \mathbf{V}_{2n^{ob}\times(p+1)}] \\
\mathbf{V^{id}} &= \begin{bmatrix} \mathbf{Z}^{ob}_{n^{ob}\times p}, \mathbf{1}_{n^{ob}\times 1} - \mathbf{W}^{ob}_{n^{ob}\times 1} \\ \mathbf{Z}^{ob}_{n^{ob}\times p}, \mathbf{W}^{ob}_{n^{ob}\times 1} \end{bmatrix} \\
\mathbf{Z^{ob}} &= [\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_p}]_{n^{ob}\times p}
\end{aligned} \tag{3.19}$$

The data format defined in (3.17) through (3.19) has a notable difference from the format defined earlier: The data matrix of **Z** and vector of **W** now have a superscript “ob”, which signals both are now fixed as portions of the observed sample. For example, the values of covariates in **Z** for an unobserved sample in this case will be identical to $\mathbf{Z}^{ob}_{n^{ob}\times p}$, i.e., the observed values of covariates in **Z**. What $\mathbf{1}_{n^{ob}\times 1} - \mathbf{W}^{ob}_{n^{ob}\times 1}$ indicates is that every treated subject in the observed sample will choose the control group and every control subject in the observed sample will choose the treatment group, in an unobserved sample. The raw data defined by this fashion will have the following properties:

$$\begin{aligned}
&\mathbf{S^{id}_{ZZ}} = \mathbf{S^{un}_{ZZ}} = \mathbf{S^{ob}_{ZZ}} \\
&\sigma^{id}_{WW} = \sigma^{un}_{WW} = \sigma^{ob}_{WW} \\
&n^{un} = n^{ob} \\
&\mathbf{S^{id}_{WZ}} = \mathbf{0} \\
&\mathbf{\bar{Z}^{id}} = \mathbf{\bar{Z}^{un}} = \mathbf{\bar{Z}^{ob}} \\
&\bar{W}^{un} = 1 - \bar{W}^{ob} \\
&\bar{W}^{id} = 0.5
\end{aligned} \tag{3.20}$$

The above constraints on the parameters in the probit models impart an insight that the probit models will be greatly simplified when the analysis of robustness is applied to a research with limited internal validity, due to the nature of an unobserved sample that is composed of counterfactuals of actual observations. Specifically, the probit models (3.15) and (3.16) will be reduced in the first scenario as follows:

For inferring a positive effect:

$$probit(p)=\frac{\sqrt{2n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{ob}}}[\gamma_{w}^{\#}\hat{\sigma}_{WW}^{ob}-0.5\hat{\sigma}_{WY}^{un}-0.5\hat{\sigma}_{WY}^{ob}-(0.5\bar{W}^{ob}-0.25)(\bar{Y}^{ob}-\bar{Y}^{un})] \quad (3.21)$$

For inferring a negative effect:

$$probit(p)=\frac{\sqrt{2n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{ob}}}[0.5\hat{\sigma}_{WY}^{un}+0.5\hat{\sigma}_{WY}^{ob}+(0.5\bar{W}^{ob}-0.25)(\bar{Y}^{ob}-\bar{Y}^{un})-\gamma_{w}^{\#}\hat{\sigma}_{WW}^{ob}] \quad (3.22)$$

The above probit models are intended for research with limited internal validity and it’s remarkable that the only existing parameters are the covariance between the outcome Y and the treatment indicator W in an unobserved sample (i.e., $\hat{\sigma}_{WY}^{un}$) and the unobserved sample mean for the outcome Y (i.e., $\bar{Y}^{un}$). The probit link of the probability of invalidating an inference is now a linear function of $\bar{Y}^{un}$ and $\hat{\sigma}_{WY}^{un}$. The analytical strategy regarding those probit models then becomes straightforward: one can either assume $\bar{Y}^{un}$ is fixed and identify the threshold of $\hat{\sigma}_{WY}^{un}$ that makes the probability of invalidating an inference smaller than a certain number (say 0.5), or locate the threshold of $\bar{Y}^{un}$ that makes the probability of invalidating an inference smaller than a certain number (say 0.5) conditional on a value of $\hat{\sigma}_{WY}^{un}$.

In contrast to the first scenario where probit models have been greatly simplified comparing to (3.15) and (3.16), probit models will remain the same as presented in (3.15) and (3.16) in the second scenario where a research design has strong internal validity but limited external validity. This means the pool of the candidates of focal parameters in the subsequent analysis of robustness could be large and therefore the choice of focal parameters is important and necessary. It's worthy of stressing here that, a reasonable conceptualization of a possible unobserved sample is the preliminary for an analysis of robustness. In the example of Borman et al. (2008), an enlightening model of conceptualization would be asking questions such as "what would the covariance between the posttest scores and OCR have been had they drawn a random sample of classrooms from the non-volunteered schools" or "what would the mean pretest and posttest scores have been had they drawn a random sample of classrooms from the non-volunteered schools". Ideally, one should extrapolate the unobserved sample size and means, variances and covariances for all relevant variables in an unobserved sample. In practice, as illustrated later, an unobserved sample statistic will be restricted to be equal to its counterpart in the observed sample except that it is of particular interest to a researcher in his analysis of robustness or some strong evidence has suggested that it should be different from its counterpart in the observed sample.

4-Bayesian models of robustness indices for centered and standardized data

4.1-For centered data

The purpose of this section is to demonstrate that, although the Bayesian models of robustness indices may seem complicated and difficult to work with, they can be simplified by considering data in observed and unobserved samples as centered (or standardized), rather as raw. The centered data in observed, unobserved and ideal samples will be in the identical form as presented in (3.1) through (3.3) except that the means of all variables will be zero now.

Consequently, the posterior distribution of γ_w will remain unchanged for centered data with the following instructions for calculating the sample variances and covariances in an ideal sample:

$$\begin{aligned}
&\hat{\sigma}^{id}_{Z_iZ_j} = \lambda\hat{\sigma}^{un}_{Z_iZ_j} + (1-\lambda)\hat{\sigma}^{ob}_{Z_iZ_j} \quad \text{for } i \neq j = 1,2,\dots,p \\
&\hat{\sigma}^{id}_{WZ_i} = \lambda\hat{\sigma}^{un}_{WZ_i} + (1-\lambda)\hat{\sigma}^{ob}_{WZ_i} \quad \text{for } i = 1,2,\dots,p \\
&\hat{\sigma}^{id}_{WW} = \lambda\hat{\sigma}^{un}_{WW} + (1-\lambda)\hat{\sigma}^{ob}_{WW} \\
&\hat{\sigma}^{id}_{Z_iY} = \lambda\hat{\sigma}^{un}_{Z_iY} + (1-\lambda)\hat{\sigma}^{ob}_{Z_iY} \quad \text{for } i = 1,2,\dots,p \\
&\hat{\sigma}^{id}_{WY} = \lambda\hat{\sigma}^{un}_{WY} + (1-\lambda)\hat{\sigma}^{ob}_{WY}
\end{aligned} \tag{4.1}$$

Obviously, (4.1) is a reduced form of its counterpart for raw data as displayed in (3.13). The rest of analysis of robustness for centered data is the same as the analysis of robustness for raw data: Built on the probit models derived in (3.15) and (3.16), one can pinpoint the threshold of a focal parameter for the probability of invalidating an inference to be smaller than a certain value (say 0.5), conditional on some values of all other unobserved and observed sample statistics. I comment here that, as a special case of the probit models in (3.15) and (3.16), the probit models for research that lacks internal validity can be further reduced as follows:

For inferring a positive effect:

$$probit(p) = \frac{\sqrt{2n^{ob}}}{\sigma\sqrt{\hat{\sigma}^{ob}_{WW}}}[\gamma^{\#}_{w}\hat{\sigma}^{ob}_{WW} - 0.5\hat{\sigma}^{un}_{WY} - 0.5\hat{\sigma}^{ob}_{WY}] \tag{4.2}$$

For inferring a negative effect:

$$probit(p) = \frac{\sqrt{2n^{ob}}}{\sigma\sqrt{\hat{\sigma}^{ob}_{WW}}}[0.5\hat{\sigma}^{un}_{WY} + 0.5\hat{\sigma}^{ob}_{WY} - \gamma^{\#}_{w}\hat{\sigma}^{ob}_{WW}] \tag{4.3}$$

4.2-For standardized data

If one has the raw data and chooses to standardize all the variables, or if he can assume or infer the standardized coefficients from the results of a research, the data in observed, unobserved and ideal samples might be thought of as being standardized and resultantly the posterior mean and variance of γ_w as well as the probit models will be functions of sample correlation matrices defined below:

$$\mathbf{R_{ZZ}^{id}} = \begin{pmatrix} r_{Z_1Z_1}^{id} & \cdots & r_{Z_1Z_p}^{id} \\ \vdots & \ddots & \vdots \\ r_{Z_pZ_1}^{id} & \cdots & r_{Z_pZ_p}^{id} \end{pmatrix}_{p\times p}$$

$$\mathbf{R_{ZW}^{id}} = \begin{pmatrix} r_{Z_1W}^{id} \\ \vdots \\ r_{Z_pW}^{id} \end{pmatrix}_{p\times 1}$$

$$\mathbf{R_{WZ}^{id}} = \begin{pmatrix} r_{Z_1W}^{id} & \cdots & r_{Z_pW}^{id} \end{pmatrix}_{1\times p} \tag{4.4}$$

$$\mathbf{R_{ZY}^{id}} = \begin{pmatrix} r_{Z_1Y}^{id} \\ \vdots \\ r_{Z_pY}^{id} \end{pmatrix}_{p\times 1}$$

where r is used to denote the sample correlation between two variables. For example, r_{WY}^{id} is the sample correlation between the treatment indicator W and the outcome Y in an ideal sample.

Theorem 3. The posterior distribution of γ_w could be rewritten as follows, providing all assumptions of theorem 1 can be upheld, for standardized data:

$$\gamma_w \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}} \sim N\left(\frac{r_{WY}^{id} - \mathbf{R_{WZ}^{id}}(\mathbf{R_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{R_{ZY}^{id}}}{1 - \mathbf{R_{WZ}^{id}}(\mathbf{R_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{R_{ZW}^{id}}}, \frac{\sigma^2}{n^{un} + n^{ob}}(1 - \mathbf{R_{WZ}^{id}}(\mathbf{R_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{R_{ZW}^{id}})^{\mathbf{-1}}\right) \tag{4.5}$$

where:

$$
\begin{aligned}
\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{id}} &= \frac{n^{un} * \mathbf{R}_{\mathbf{ZZ}}^{\mathbf{un}} + n^{ob} * \mathbf{R}_{\mathbf{ZZ}}^{\mathbf{ob}}}{n^{un} + n^{ob}} = \lambda \mathbf{R}_{\mathbf{ZZ}}^{\mathbf{un}} + (1-\lambda)\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{ob}} \\
\mathbf{R}_{\mathbf{WZ}}^{\mathbf{id}} &= \frac{n^{un} * \mathbf{R}_{\mathbf{WZ}}^{\mathbf{un}} + n^{ob} * \mathbf{R}_{\mathbf{WZ}}^{\mathbf{ob}}}{n^{un} + n^{ob}} = \lambda \mathbf{R}_{\mathbf{WZ}}^{\mathbf{un}} + (1-\lambda)\mathbf{R}_{\mathbf{WZ}}^{\mathbf{ob}} \\
\mathbf{R}_{\mathbf{ZY}}^{\mathbf{id}} &= \frac{n^{un} * \mathbf{R}_{\mathbf{ZY}}^{\mathbf{un}} + n^{ob} * \mathbf{R}_{\mathbf{ZY}}^{\mathbf{ob}}}{n^{un} + n^{ob}} = \lambda \mathbf{R}_{\mathbf{ZY}}^{\mathbf{un}} + (1-\lambda)\mathbf{R}_{\mathbf{ZY}}^{\mathbf{ob}} \\
r_{WY}^{id} &= \frac{n^{un} * r_{WY}^{un} + n^{ob} * r_{WY}^{ob}}{n^{un} + n^{ob}} = \lambda r_{WY}^{un} + (1-\lambda) r_{WY}^{ob}
\end{aligned}
\tag{4.6}
$$

Proof:

The proof of theorem 3 should be straightforward once $\hat{\sigma}_{WW}^{id}$ is plugged in as 1 and every covariance term as its corresponding correlation term (i.e., a covariance matrix equal to its corresponding correlation matrix, a covariance equal to its corresponding correlation) into the distribution proposed by theorem 1.

Theorem 4. The probit models for the probability of invalidating an inference (p) could be rewritten as follows, providing all assumptions of theorem 2 are met, for standardized data:

For inferring a positive effect:

$$
probit(p) = \frac{\sqrt{n^{un} + n^{ob}}}{\sigma\sqrt{1 - \mathbf{R}_{\mathbf{WZ}}^{\mathbf{id}} (\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{id}})^{\mathbf{-1}} \mathbf{R}_{\mathbf{ZW}}^{\mathbf{id}}}} [\gamma_w^{\#}(1 - \mathbf{R}_{\mathbf{WZ}}^{\mathbf{id}} (\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{id}})^{\mathbf{-1}} \mathbf{R}_{\mathbf{ZW}}^{\mathbf{id}}) - (r_{WY}^{id} - \mathbf{R}_{\mathbf{WZ}}^{\mathbf{id}} (\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{id}})^{\mathbf{-1}} \mathbf{R}_{\mathbf{ZY}}^{\mathbf{id}})]
$$

(4.7)

For inferring a negative effect:

$$
probit(p) = \frac{\sqrt{n^{un} + n^{ob}}}{\sigma\sqrt{1 - \mathbf{R}_{\mathbf{WZ}}^{\mathbf{id}} (\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{id}})^{\mathbf{-1}} \mathbf{R}_{\mathbf{ZW}}^{\mathbf{id}}}} [(r_{WY}^{id} - \mathbf{R}_{\mathbf{WZ}}^{\mathbf{id}} (\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{id}})^{\mathbf{-1}} \mathbf{R}_{\mathbf{ZY}}^{\mathbf{id}}) - \gamma_w^{\#}(1 - \mathbf{R}_{\mathbf{WZ}}^{\mathbf{id}} (\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{id}})^{\mathbf{-1}} \mathbf{R}_{\mathbf{ZW}}^{\mathbf{id}})]
$$

(4.8)

Proof:

Just as the proof of theorem 3, I plug $\hat{\sigma}_{WW}^{id}$ as 1 and every covariance term as its corresponding correlation term (i.e., a covariance matrix equal to its corresponding correlation matrix, a covariance equal to its corresponding correlation) into the probit models proposed by theorem 2. Finally, the probit models for a research with limited internal validity have quite simple and tidy forms as below:

For inferring a positive effect:

$$probit(p) = \frac{\sqrt{2n^{ob}}}{\sigma}[\gamma_w^{\#} - 0.5r_{WY}^{un} - 0.5r_{WY}^{ob}] \tag{4.9}$$

For inferring a negative effect:

$$probit(p) = \frac{\sqrt{2n^{ob}}}{\sigma}[0.5r_{WY}^{un} + 0.5r_{WY}^{ob} - \gamma_w^{\#}] \tag{4.10}$$

5-Appropriate statistical threshold and Bayesian models for replacing observed cases

5.1-Appropriate statistical threshold

The analysis of robustness of causal inferences I have discussed so far can be perceived as a procedure of retesting the hypothesis for a conceptualized ideal sample. Such hypothesis testing procedure entails a statistical threshold which is a product of the chosen critical value (traditionally it’s 1.96) and the standard error of the estimate of treatment effect based on an ideal sample (i.e., $\hat{\gamma}_w^{id}$). I emphasize that, unlike the standard error of $\hat{\gamma}_w^{ob}$ (the estimate of treatment effect based on the observed sample) which only considers the observed sample, the standard error of $\hat{\gamma}_w^{id}$ accounts for observations of both observed sample and unobserved sample and thus becomes the appropriate choice.

The standard error of $\hat{\gamma}_w^{id}$ has been given by theorem 1 as follows:

$$se(\hat{\gamma}_w^{id}) = \sqrt{\frac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}} \tag{5.1}$$

where the computations of $\hat{\sigma}_{WW}^{id}, \mathbf{S_{WZ}^{id}}, \mathbf{S_{ZZ}^{id}}$ are guided by (3.11) through (3.14). Furthermore, the statistical threshold of $\hat{\gamma}_w^{id}$ is calculated as below, for testing the null hypothesis of $\gamma_w = 0$ with level of significance as 0.05:

$$\gamma_w^{\#} = 1.96 * \sqrt{\frac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}} \text{ for inferring a positive effect}$$

$$\gamma_w^{\#} = -1.96 * \sqrt{\frac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}} \text{ for inferring a negative effect}$$

(5.2)

With the threshold $\gamma_w^{\#}$ given in (5.2), the probit models of the probabilities of invalidating an inference turn out to be as follows:

For inferring a positive effect:

$$probit(p) = 1.96 - \frac{\sqrt{n^{un}+n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}}(\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}) \tag{5.3}$$

For inferring a negative effect:

$$probit(p) = 1.96 + \frac{\sqrt{n^{un}+n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}}(\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}) \tag{5.4}$$

The thresholds offered in (5.2) will remain instructive as long as the decision about threshold $\gamma_w^{\#}$ is a pure statistical one. However, other factors such as transaction cost of proposed action may come into play in determining $\gamma_w^{\#}$ and a relevant discussion about non-statistical threshold has been offered by Frank et al. (2013).

5.2-The Bayesian models of robustness indices for replacing observed cases

Frank & Min (2007) has proposed two mechanisms of forming an ideal sample: The first one is neutralization by addition, which creates an ideal sample by adding an unobserved sample to the existent observed sample. Until now the Bayesian models of robustness indices of causal inferences have exclusively centered on this mechanism. The other one is neutralization by replacement, which generates an ideal sample by replacing a part of the observed sample with an unobserved sample. In this case, an ideal sample will have the same size as the observed sample and it has both observations inherited from the observed sample and observations introduced by an unobserved sample. In this subsection, a new set of Bayesian models of robustness indices will be devoted to neutralization by replacement, as they are supplementary to the existing Bayesian models and provides alternative conceptualizations and interpretations to the robustness indices.

To parameterize the mechanism of neutralization by replacement, the following notations are defined: I_i is the binary indicator of whether i^th observed case (say his name is Tom) is kept in an ideal sample (equivalently, this means Tom is not replaced with an unobserved case). **s** represents the collection of all I_i i=1, 2, …, n^{ob} . Therefore, every **s** is a subsample of the observed sample and it will be kept in an ideal sample.

The Bayesian models of robustness indices of causal inferences for replacing observed cases are defined as follows:

$$\begin{aligned}&\boldsymbol{\gamma} \sim N(((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{X^{un}})^{\mathbf{-1}}(\mathbf{X^{un}})^{\mathbf{T}}\mathbf{Y^{un}}, \sigma^2((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{X^{un}})^{\mathbf{-1}})\\ &Y_i \mid \boldsymbol{\gamma}, \mathbf{X_i}, I_i \sim N(\mathbf{X_i}\boldsymbol{\gamma}, \sigma^2)\\ &\boldsymbol{\gamma} \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}, \mathbf{s} \sim N(\boldsymbol{\theta}^s_{\gamma}, \boldsymbol{\Phi}^s_{\gamma})\end{aligned} \qquad (5.5)$$

where:

$$\begin{aligned}&\boldsymbol{\theta}^s_{\gamma} = ((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{X^{un}} + (\mathbf{X^{ob|s}})^{\mathbf{T}}\mathbf{X^{ob|s}})^{\mathbf{-1}}((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{Y^{un}} + (\mathbf{X^{ob|s}})^{\mathbf{T}}\mathbf{Y^{ob|s}})\\ &\boldsymbol{\Phi}^s_{\gamma} = \sigma^2((\mathbf{X^{un}})^{\mathbf{T}}\mathbf{X^{un}} + (\mathbf{X^{ob|s}})^{\mathbf{T}}\mathbf{X^{ob|s}})^{\mathbf{-1}}\end{aligned} \qquad (5.6)$$

In (5.6), $\mathbf{X^{ob|s}}$ and $\mathbf{Y^{ob|s}}$ refer to the matrix of covariates and vector of outcomes respectively for the observed cases which are not replaced with unobserved ones (so they are retained in an ideal sample). To obtain the target posterior distribution $\boldsymbol{\gamma} \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}$, one should take the expectation of the probability density function of $\boldsymbol{\gamma} \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}, \mathbf{s}$ over the distribution of subsample **s**. This could be practically done through a Monte Carlo simulation.

A closed theoretical form of $\boldsymbol{\gamma} \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}$ may not be straightforward for most cases. Fortunately, for a random sampling procedure, any sample statistics of **s** should center around the same sample statistic of the whole observed sample. Consequently, the distribution of $\boldsymbol{\gamma} \mid \mathbf{Y^{ob}}, \mathbf{X^{ob}}$ can be approximated by assuming the sample statistics of the retained observed cases are identical to the sample statistics of the whole observed sample.

Resultantly, based on theorem 1, I have:

$$\gamma_w \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}} \sim N\left(\frac{\hat{\sigma}^{id}_{WY} - \mathbf{S^{id}_{WZ}}(\mathbf{S^{id}_{ZZ}})^{\mathbf{-1}}\mathbf{S^{id}_{ZY}}}{\hat{\sigma}^{id}_{WW} - \mathbf{S^{id}_{WZ}}(\mathbf{S^{id}_{ZZ}})^{\mathbf{-1}}\mathbf{S^{id}_{ZW}}}, \frac{\sigma^2}{n^{ob}}(\hat{\sigma}^{id}_{WW} - \mathbf{S^{id}_{WZ}}(\mathbf{S^{id}_{ZZ}})^{\mathbf{-1}}\mathbf{S^{id}_{ZW}})^{\mathbf{-1}}\right) \qquad (5.7)$$

where:

$$\begin{aligned}
\hat{\sigma}^{id}_{Z_iZ_j} &= \lambda\hat{\sigma}^{un}_{Z_iZ_j} + (1-\lambda)\hat{\sigma}^{ob}_{Z_iZ_j} + (1-\lambda)\lambda(\bar{Z}^{ob}_i - \bar{Z}^{un}_i)(\bar{Z}^{ob}_j - \bar{Z}^{un}_j) \text{ for } i \neq j = 1,2,\ldots,p \\
\hat{\sigma}^{id}_{WZ_i} &= \lambda\hat{\sigma}^{un}_{WZ_i} + (1-\lambda)\hat{\sigma}^{ob}_{WZ_i} + (1-\lambda)\lambda(\bar{W}^{ob} - \bar{W}^{un})(\bar{Z}^{ob}_i - \bar{Z}^{un}_i) \text{ for } i = 1,2,\ldots,p \\
\hat{\sigma}^{id}_{WW} &= \lambda\hat{\sigma}^{un}_{WW} + (1-\lambda)\hat{\sigma}^{ob}_{WW} + (1-\lambda)\lambda(\bar{W}^{ob} - \bar{W}^{un})^2 \\
\hat{\sigma}^{id}_{Z_iY} &= \lambda\hat{\sigma}^{un}_{Z_iY} + (1-\lambda)\hat{\sigma}^{ob}_{Z_iY} + (1-\lambda)\lambda(\bar{Z}^{ob}_i - \bar{Z}^{un}_i)(\bar{Y}^{ob} - \bar{Y}^{un}) \text{ for } i = 1,2,\ldots,p \\
\hat{\sigma}^{id}_{WY} &= \lambda\hat{\sigma}^{un}_{WY} + (1-\lambda)\hat{\sigma}^{ob}_{WY} + (1-\lambda)\lambda(\bar{W}^{ob} - \bar{W}^{un})(\bar{Y}^{ob} - \bar{Y}^{un})
\end{aligned}$$

(5.8)

and:

$$\lambda = \frac{n^{un}}{n^{ob}} \tag{5.9}$$

λ symbolizes the proportion of observed sample to be replaced with an unobserved sample. Draw on the Bayesian models of robustness indices for replacing observed cases, as formalized in (5.5) through (5.9), we can quantify the probabilities of invalidating an inference and formulate their corresponding probit models which are identical to (3.15) and (3.16) except the expression of λ.

6-Illustrative examples

6.1-The Bayesian robustness indices of the effect of Open Court Reading on reading achievement

Open Court Reading (OCR) program is a curriculum that is rooted in research-based practices and has been in the market for a long time and widely adopted by many districts and schools. Although OCR is potentially a beneficial program because it responds to recommendations from research that focused on developing early reading skills, its effect has never been assessed and confirmed by a randomized experiment. Seeing this, Borman et al. (2008) designed a multisite cluster randomized experiment and randomly drew six schools from those who volunteered in their study. Subsequently, they define a block as a single grade of one sampled school and within each block classrooms were randomly assigned to the OCR group or the control group. Controlling for the pretest scores and block membership, Borman et al. (2008) estimated the effect of OCR as 7.95 (on reading composite scores) which was statistically significant and went on to conclude that “the outcomes from these analyses provided not only evidence of the promising 1-year effects of OCR on students’ reading outcomes but also suggest that these effects may be replicated across varying contexts with rather consistent and positive results”. Ideally, the findings of Borman et al. (2008) implicate that, the estimated effect of OCR would be around 7.95 if one were to conduct a large-scale completely randomized experiment, controlling for the pretest scores. In other words, regression-based causal inference based on a design where a large random sample of classrooms from the entire U.S. is available and all sampled classrooms are randomly assigned to the OCR group and the control group, would lead to an estimate of the effect of OCR as nearly 7.95 with the mean posttest scores of sampled classrooms as the outcome and the mean pretest scores of sampled classrooms as the covariate. Comparing Borman et al. (2008) to the regression-based causal inference built on this imaginary large-scale completely randomized experiment is a necessary starting point of the analysis of robustness in this paper. Nevertheless, such comparison isn’t necessarily plausible as Borman et al. (2008) only had a random sample of classrooms from the volunteered schools instead of a nationwide random sample of classrooms and thus Borman et al. (2008) actually had a nonrandom sample from its target population, i.e., the classrooms in the entire U.S.. Therefore, Borman et al. (2008) fits the description of the second scenario and we can proceed with its analysis of robustness.

Conforming to my notation rules, data structure and formulations proposed earlier, the pretest score, OCR and the posttest score are the covariate Z, the treatment indicator W and the outcome Y respectively. Furthermore, the sample statistics of the observed sample in Borman et al. (2008) should be fixed as follows:

Means of Z, W and Y and observed sample size:

$$\bar{Z}^{ob} = 576.62,\ \bar{W}^{ob} = 0.55,\ \ \bar{Y}^{ob} = 609.96,\ n^{ob} = 49 \tag{6.1}$$

Covariance matrix of [Z, W, Y]:

$$\begin{pmatrix} 2079.36 & 0.39 & 1832.2 \\ 0.39 & 0.25 & 2.33 \\ 1832.2 & 2.33 & 2401 \end{pmatrix} \tag{6.2}$$

Next step is crucial: we need to present our research questions and make assumptions about an unobserved sample. The purpose of presenting our research questions in analysis of robustness is to isolate the focal parameters that are required by the answer of research questions. In this example, I decide to follow the logic of Frank et al. (2013) and wonder the number of classrooms randomly drawn from non-volunteered schools we need to add to the observed sample of Borman et al. (2008) such that the probability of invalidating their inference is smaller than a prespecified benchmark (say 0.5), assuming the covariance between OCR and posttest is 0 in those classrooms sampled from non-volunteered schools. The meaning of this research question is three-fold: First, an unobserved sample which in this context is a random sample of classrooms drawn from non-volunteered schools is needed and to be added to the observed sample so that an ideal sample is constructed. My robustness index, i.e., the probability of invalidating an inference is computed based on this imaginary ideal sample. Two, this research question involves an assumption that the sample covariance between OCR (W) and the posttest scores (Y) is zero in this unobserved sample. Three, the focal parameter for this research question is the unobserved sample size n^{un}. Particularly, I am interested in the relationship between n^{un} and the probability of invalidating the inference of Borman et al. (2008), holding all other unobserved and observed sample statistics fixed.

To fix every unobserved sample statistic other than the focal parameters, it's inevitable to impose some constraints on them. As discussed earlier, an unobserved sample statistic can be quantified as a number whenever it's defensible to do so. More often, an unobserved sample statistic is constrained to be its observed counterpart. In this case, every unobserved statistic other than the focal parameter n^{un} and $\hat{\sigma}_{WY}^{un}$ (which is assumed to be 0) is thought to be identical to its observed counterpart. The constraints imposed on all unobserved sample statistics as well as the assumptions about the observed sample statistics and residual variance (i.e., σ^2) are summarized below:

$$\begin{aligned}
&\hat{\sigma}_{WY}^{un} = 0,\ \hat{\sigma}_{WY}^{ob} = 2.33 \\
&n^{ob} = 49, \sigma^2 = 32 \\
&\hat{\sigma}_{ZZ}^{un} = \hat{\sigma}_{ZZ}^{ob} = 2079.36 \\
&\hat{\sigma}_{ZY}^{un} = \hat{\sigma}_{ZY}^{ob} = 1832.2 \\
&\hat{\sigma}_{WW}^{un} = \hat{\sigma}_{WW}^{ob} = 0.25 \\
&\hat{\sigma}_{ZW}^{un} = \hat{\sigma}_{ZW}^{ob} = 0.39 \\
&\bar{Z}^{un} = \bar{Z}^{ob} = 576.62 \\
&\bar{Y}^{un} = \bar{Y}^{ob} = 609.96 \\
&\bar{W}^{un} = \bar{W}^{ob} = 0.55
\end{aligned} \tag{6.3}$$

Assuming the statistical threshold in (5.2) is adopted, the parametric values provided in (6.3) lead to the following probit model of the probability of invalidating the inference for Borman et al. (2008):

$$probit(p) = 1.96 - \frac{40.36}{\sqrt{n^{un}+49}} + 0.12\sqrt{n^{un}+49} \tag{6.4}$$

Drawing on the probit model in (6.4), the main analytical strategy is to pinpoint the threshold of n^{un} that makes the probability of invalidating the inference of Borman et al. (2008) smaller than a desired value. For an example, if we would like to find the threshold of n^{un} corresponding to a probability smaller than 0.5, the probit model (6.4) needs to be transformed as an inequality with regard to n^{un} as follows:

$$1.96 - \frac{40.36}{\sqrt{n^{un}+49}} + 0.12\sqrt{n^{un}+49} < 0 \tag{6.5}$$

The above inequality suggests that the probability of invalidating the inference of Borman et al. (2008) would be smaller than 0.5 as long as n^{un} is not greater than 91, which means the probability of invalidating the inference of Borman et al. (2008) is smaller than 0.5 when 91 or less classrooms are randomly sampled from the non-volunteered schools assuming the covariance between OCR and the posttest scores is 0 among those sampled classrooms and all other parameters are fixed as in (6.3). Furthermore, the estimated regression coefficient of the treatment indicator W based on an ideal sample is calculable with the parametric values in (6.3) and the threshold value of n^{un} as follows:

$$\hat{\gamma}_w^{id} = \frac{456.68}{49+n^{un}} - 1.375 \tag{6.6}$$

whose general form is the following:

$$\hat{\gamma}_w^{id} = \frac{\hat{\sigma}_{WY}^{id} - \mathbf{S}_{\mathbf{WZ}}^{\mathbf{id}}(\mathbf{S}_{\mathbf{ZZ}}^{\mathbf{id}})^{-1}\mathbf{S}_{\mathbf{ZY}}^{\mathbf{id}}}{\hat{\sigma}_{WW}^{id} - \mathbf{S}_{\mathbf{WZ}}^{\mathbf{id}}(\mathbf{S}_{\mathbf{ZZ}}^{\mathbf{id}})^{-1}\mathbf{S}_{\mathbf{ZW}}^{\mathbf{id}}} \tag{6.7}$$

It turns out that the threshold of $n^{un} = 91$ corresponds to the estimated regression coefficient of W based on an ideal sample as 1.89, which further corresponds to the probability of invalidating the inference of Borman et al. (2008) as 0.5. To gain comprehensive knowledge about the one-to-one relationships among the probability of invalidating Borman et al. (2008)'s inference, the threshold of n^{un} and the estimated regression coefficient of W in an ideal sample, it's strongly recommended that the threshold of n^{un} and the estimated regression coefficient of W are repeatedly calculated regarding various desired values of the probability of invalidating the inference of Borman et al. (2008). Below are a table and a graph illustrate those relationships. Table 2.1 tabulates the thresholds of n^{un} and associated estimated regression coefficients of W that make the probability of invalidating the inference of Borman et al. (2008) lower than 0.1, 0.2, …, 0.9 respectively. For example, at most 64 classrooms which are randomly drawn from the non-volunteered schools and have a zero sample correlation between OCR and posttest reading scores can be added to the observed sample so as to keep the probability of invalidating the inference of Borman et al. (2008) under 0.3, given the parametric values in (6.3). The relationship between n^{un} and the probability of invalidating the inference of Borman et al. (2008) is further delineated in figure 2.4, and as expected they are positively correlated, which means adding more sampled classrooms with zero correlation between OCR and posttest scores in the non-volunteered schools to the observed sample will weaken the inference of Borman et al. (2008).

In figure 2.5, I intend to present the posterior probability of invalidating the inference of Borman et al. (2008) in a context of testing null hypothesis. The black curves in figure 5 are distributions corresponding to null hypothesis $\gamma_w = 0$ and the red curves are distributions of γ_w conditional on a given ideal sample. Figure 2.5 depicts the same pattern as manifested in Table 2.1: as the unobserved sample size becomes larger the distribution corresponding to null hypothesis and the distribution of γ_w will get closer, and therefore the probability of invalidating the inference of Borman et al. (2008) will be larger as well. The appropriate statistical threshold will, in this case, keep dropping because the ideal sample size keeps growing. The knowledge of paramount importance imparted by figure 2.5 is that the posterior probability of invalidating the inference of Borman et al. (2008) can be conceptualized as type II error with regard to retesting the null hypothesis: $\gamma_w = 0$ versus the alternative hypothesis when an unobserved sample can be randomly drawn from the non-volunteered schools and merged to the observed sample of Borman et al. (2008).

Table 2.1: Thresholds of n^{un} assuming $\hat{\sigma}_{WY}^{un} = 0$

Level of probability	Threshold of n^{un}	The estimated regression coefficient of W based on an ideal sample
0.1	36	4.00
0.2	51	3.19
0.3	64	2.67
0.4	77	2.25
0.5	91	1.89
0.6	107	1.55
0.7	126	1.23
0.8	152	0.90
0.9	195	0.50

Figure 2.4: The relationship between n^{un} and the probability of invalidating the inference of Borman et al. (2008)

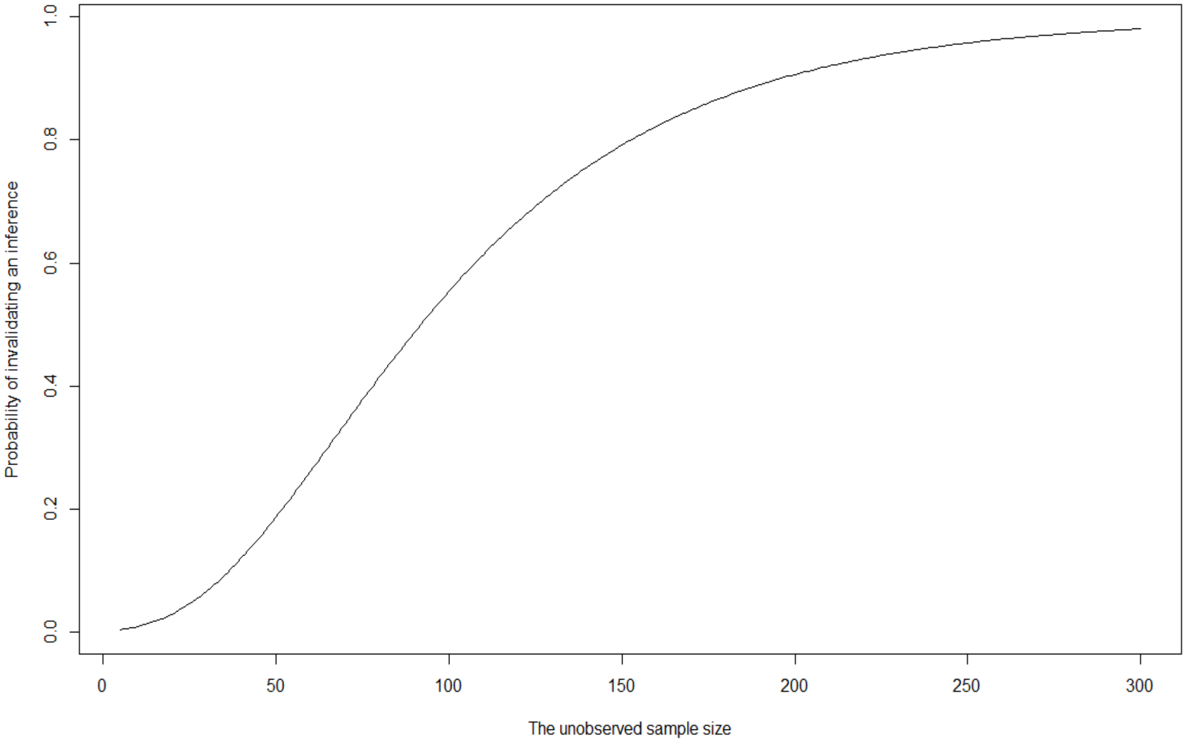
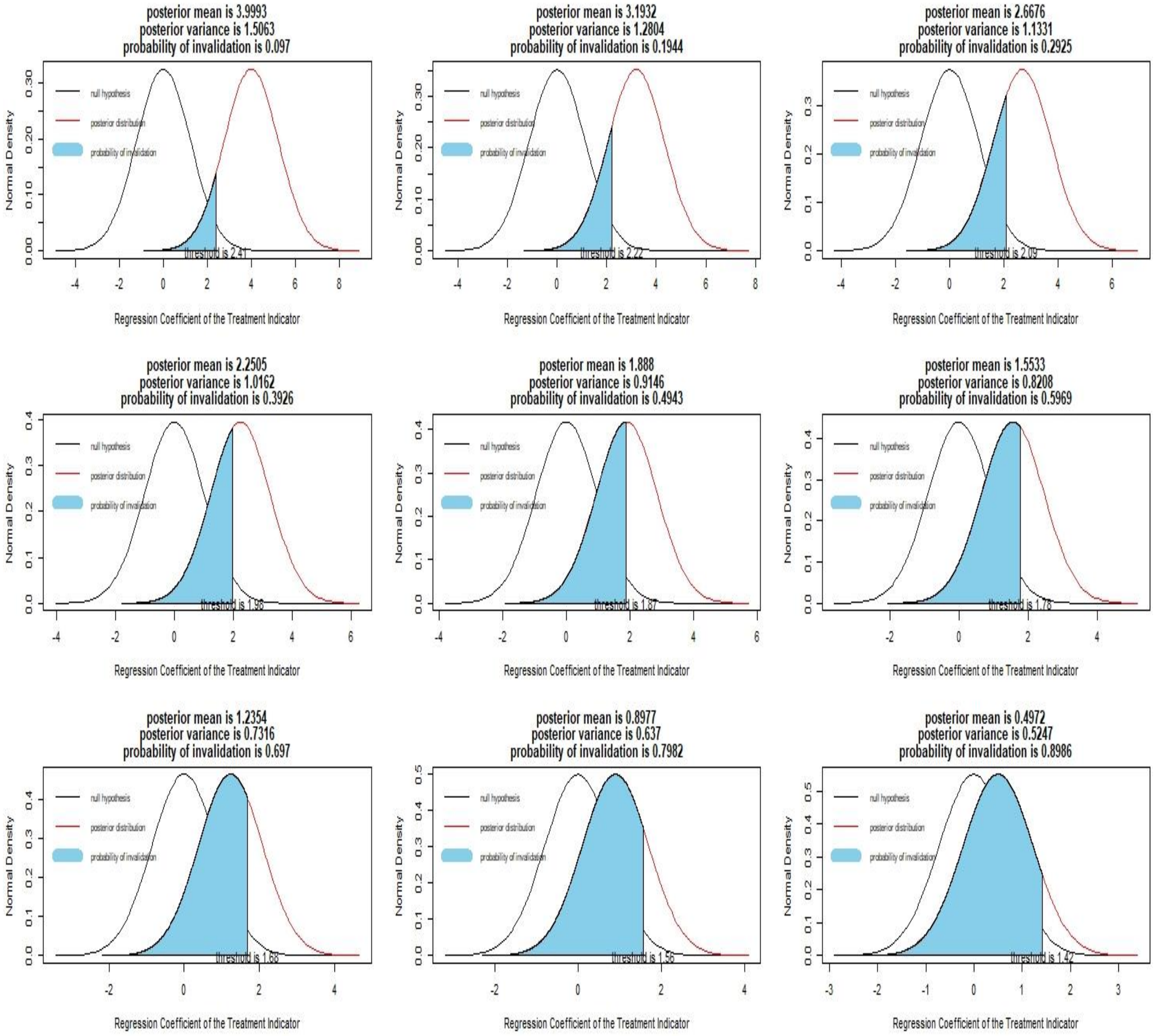


Figure 2.5: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Borman et al. (2008)



Alternatively, the inference of Borman et al. (2008) could be conceived as it's built on standardized data instead of raw data as discussed earlier. For standardized data, the correlation matrix needs to be specified for the observed sample, and in this case it is:

$$\begin{pmatrix} 1 & 0.017 & 0.82 \\ 0.017 & 1 & 0.095 \\ 0.82 & 0.095 & 1 \end{pmatrix} \tag{6.8}$$

for variables [Z, W, Y], where Z refers to the pretest scores, W refers to the OCR curriculum and Y refers to the posttest scores.

The same research question raised for raw data could now be asked again for standardized data, i.e., how many sampled classrooms do we need from the non-volunteered schools such that the probability of invalidating the inference of Borman et al. (2008) is lower than 0.5 assuming the correlation between OCR and posttest scores is 0 for those sampled classrooms? Again, I assume all unobserved sample correlations are equal to their observed counterparts except the unobserved sample correlation between OCR and posttest scores. The assumptions about the parameters are formally written as follows:

$$\begin{aligned} & r_{zy}^{ob} = r_{zy}^{un} = 0.82 \\ & r_{wy}^{ob} = 0.095,\ r_{wy}^{un} = 0 \\ & r_{zw}^{ob} = r_{zw}^{un} = 0.017 \\ & n^{ob} = 49,\ \sigma^2 = 0.0133 \end{aligned} \tag{6.9}$$

The probit model for the probability of invalidating the inference of Borman et al. (2008) then becomes explicit:

$$probit(p) = 1.96 - \frac{40.36}{\sqrt{n^{un} + 49}} + 0.12\sqrt{n^{un} + 49} \tag{6.10}$$

It shouldn’t be surprising to observe the probit model for standardized data (as in (6.10)) is exactly same as the probit model for raw data (as in (6.4)), just as standardizing variables in a regression model won’t change the t-ratio and p-value of every estimated regression coefficient. The equality between (6.4) and (6.10) is not a coincidence: For any given set of parametric values, research question and focal parameter(s) the probit model for the probability of invalidating an inference will remain the same regardless of whether the data is raw, centered or standardized. This further implies that, for any given set of parametric values, research question and focal parameter(s), the analysis and results will be identical as well for raw, centered or standardized data. For this reason, I omit the results pertaining to the probit model for standardized data as they have already been generated and presented as a product of the probit model for raw data.

6.2-The Bayesian robustness indices of the effect of kindergarten retention on reading achievement

Kindergarten retention is an educational issue which has been long and vehemently debated. As an attempt to settle this issue, Hong & Raudenbush (2005) analyzed a nationally representative sample which contained about 7639 students and 1070 schools and conducted a multilevel modeling with additional controls of the logits of estimated propensity scores and the propensity score strata. Their estimate of the effect of kindergarten retention on reading achievement was about -9, which was negatively significant. Such estimate, according to Hong & Raudenbush (2005), evidenced that “children who were retained would have learned more had they been promoted”.

Suppose the finding of Hong & Raudenbush is indeed the truth, we would have expected a regression-based causal inference to generate a similar result if we had been able to randomly assign those sampled students in Hong & Raudenbush (2005) to retention group and promotion group. This means the estimated regression coefficient of kindergarten retention should be around -9 in a regression where the outcome, the treatment and the covariate are the reading scores, kindergarten retention and the logits of estimated propensity scores respectively. However, this is not the case as such random assignment to retention and promotion groups is unrealizable. Therefore, a regression-based causal inference corresponding to Hong & Raudenbush (2005) would be problematic as it is based on a quasi-experiment with a representative sample, and clearly it falls into the first scenario I introduced in the beginning. To recognize the potential of the bias associated with the analysis of Hong & Raudenbush (2005) and profile the robustness of their inference, we need to first define a potential unobserved sample and its data form for Hong & Raudenbush (2005). As explained in the section 3.4, an unobserved sample in the first scenario should be the counterfactuals of the observed sample. Therefore, a potential unobserved sample for Hong & Raudenbush (2005) should be the counterfactuals of all the observed students. By definition, a counterfactual of a retained student would be an observation where the outcome is his/her reading score had he/she been promoted instead and his/her treatment status is promotion instead of retention. Likewise, a counterfactual of a promoted student would be an observation where the outcome is his/her reading score had he/she been retained instead and his/her treatment status is retention instead of promotion. For simplicity, I further assume that the data has been standardized for both the observed sample and an unobserved sample (counterfactuals) since standardized data will produce the same results as raw data does.

Due to the special data structures of unobserved sample and ideal sample for research with limited internal validity, as I have covered in (3.18) and (3.19), the following constraints are automatically imposed on the sample sizes and correlations:

$$\begin{aligned} & n^{un} = n^{ob} = 7639 \\ & r_{WZ}^{id} = 0 \end{aligned} \tag{6.11}$$

My research question for the analysis of robustness of Hong & Raudenbush (2005) is "what would the sample correlation between kindergarten retention and the reading scores have to be in the counterfactuals in order to make the probability of invalidating the inference of Hong & Raudenbush smaller than a desired value (say 0.5)". Apparently, the focal parameter suggested by this research question is the unobserved sample correlation between kindergarten retention and the reading scores. Constructing the probit model between the probability of invalidating the inference of Hong & Raudenbush (2005) and this focal parameter will be especially simple, as manifested by (4.10), and all relevant parametric values and assumptions are listed here:

$$\begin{aligned} & n^{un} = n^{ob} = 7639 \\ & r_{WY}^{ob} = -0.37 \\ & \sigma = 0.8 \end{aligned} \tag{6.12}$$

Furthermore, I choose the threshold $\gamma_w^{\#}$ purely based on statistical significance, and in this case $\gamma_w^{\#}$ is proven to be $-1.96\frac{\sigma}{\sqrt{2n^{ob}}}$, which equals -0.0127.

The probit model for Hong & Raudenbush (2005) should then become straightforward upon (6.12) is given:

$$probit(p) = 77.25 r_{wy}^{un} - 26.62 \tag{6.13}$$

This indicates that, for example, the threshold of r_{wy}^{un} to make the probability of invalidating the inference of Hong & Raudenbush (2005) lower than 0.5 would be identified by the following inequality:

$$77.25r_{WY}^{un} - 26.62 < 0 \tag{6.14}$$

which pinpoints this threshold as 0.3446. The proper interpretation of this threshold would be the unobserved sample correlation between kindergarten retention and the reading scores in the counterfactuals need to be smaller than 0.3446 such that the probability of invalidating the inference of Hong & Raudenbush (2005) stays below 0.5. This threshold in the meantime corresponds to an ideal sample correlation between kindergarten retention and the reading scores as -0.0127, which is exactly the threshold of statistical significance calculated based on an ideal sample. In general, for a research with questionable internal validity, the ideal sample correlation between the treatment indicator W and the outcome Y symbolizes the regression coefficient estimate of treatment indicator W if data is standardized, since the correlation between any covariate and W in an ideal sample is 0. The computation of the ideal sample correlation between W and Y is as follows:

$$r_{WY}^{id} = \frac{r_{WY}^{un} + r_{WY}^{ob}}{2} \tag{6.15}$$

A scrutiny of the relationship between the probability of invalidating the inference of Hong & Raudenbush (2005) and the unobserved sample correlation between kindergarten retention and reading scores entails repeat calculations of thresholds of r_{WY}^{un} as well as r_{WY}^{id} for other selected desired values. Table 2.2 lists those thresholds for a desired value of this probability ranging from 0.1 to 0.9. For an instance, the correlation between kindergarten retention and reading scores in the counterfactuals needs to be smaller than 0.3378 in order to keep the probability of invalidating the inference of Hong & Raudenbush (2005) under 0.3, conditional on the parametric values provided in (6.12). Figure 2.6 unearths that this probability will ascend from 0 to 1 abruptly when r_{WY}^{un} is in the range between 0.32 and 0.36. Why would this probability be so sensitive to a minute change of r_{WY}^{un} in the range [0.32, 0.36]? It is most likely due to the large sample size and effect size in Hong & Raudenbush (2005). The large sample size of Hong & Raudenbush (2005) amplifies the slope of r_{WY}^{un} in the probit model, and a strong positive correlation (stronger than 0.32) is needed in an unobserved sample so as to mitigate the large negative effect of kindergarten retention found in their observed sample.

Figure 2.7 depicts the relationship between posterior probability of invalidating Hong & Raudenbush (2005)'s inference and null hypothesis testing. As the correlation between kindergarten retention (W) and reading scores (Y) stay increasing, the gap between the distribution corresponding to null hypothesis (black curve) and the distribution of γ_w based on an ideal sample (red curve) will stay shrinking and consequently the posterior probability of invalidating the inference of Hong & Raudenbush (2005) will stay growing. The appropriate statistical threshold in this case is fixed as 0.0127 since the ideal sample size is constant (7639*2=15278). Most importantly, figure 2.7 uncovers that the posterior probability of invalidating the inference of Hong & Raudenbush (2005) can be interpreted as type II error in the context of retesting null hypothesis: $\gamma_w = 0$ against the alternative hypothesis when an unobserved sample (i.e., the set of counterfactual observations of all sampled students) is realized and added to their observed sample.

Table 2.2: Thresholds of r_{WY}^{un}

Level of probability	Threshold of r_{WY}^{un}	Threshold of r_{WY}^{id}
0.1	0.328	-0.021
0.2	0.3337	-0.0182
0.3	0.3378	-0.0161
0.4	0.3413	-0.0144
0.5	0.3446	-0.0127
0.6	0.3479	-0.0111
0.7	0.3514	-0.0093
0.8	0.3555	-0.0073
0.9	0.3612	-0.0044

Figure 2.6: The relationship between r_{WY}^{un} and the probability of invalidating the inference of Hong & Raudenbush (2005)

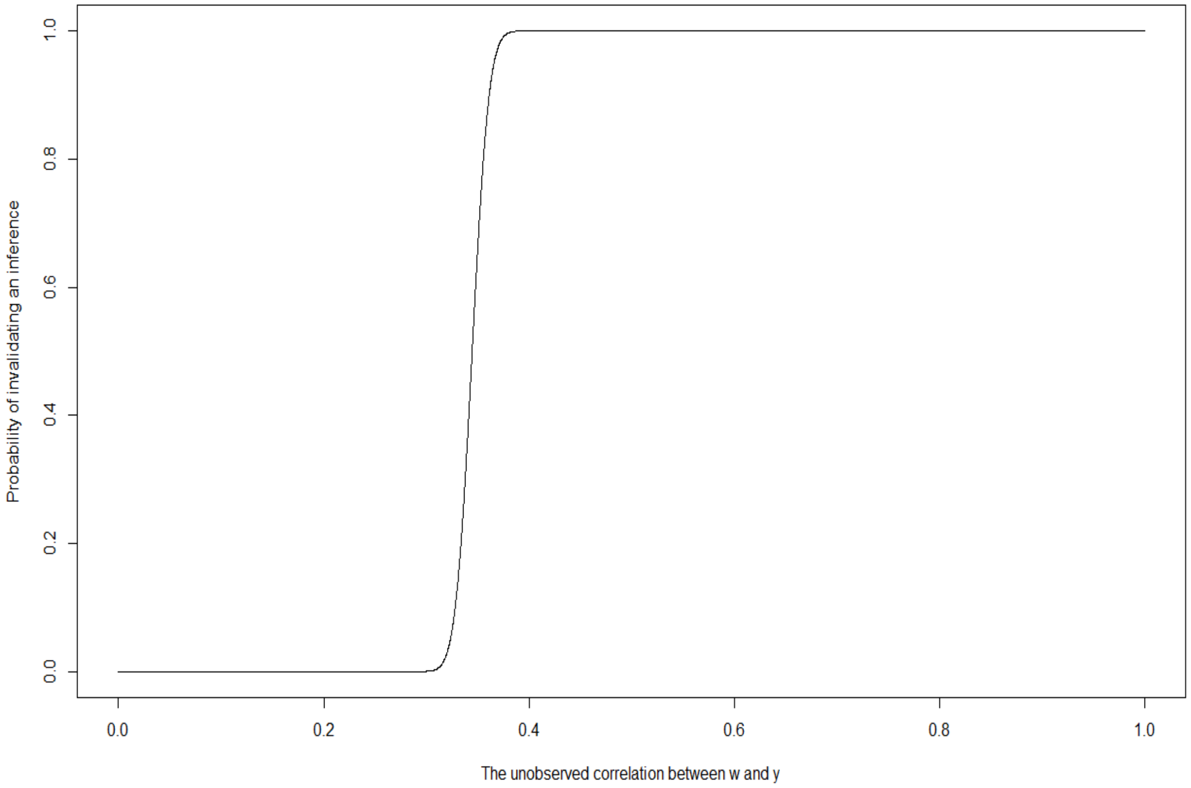
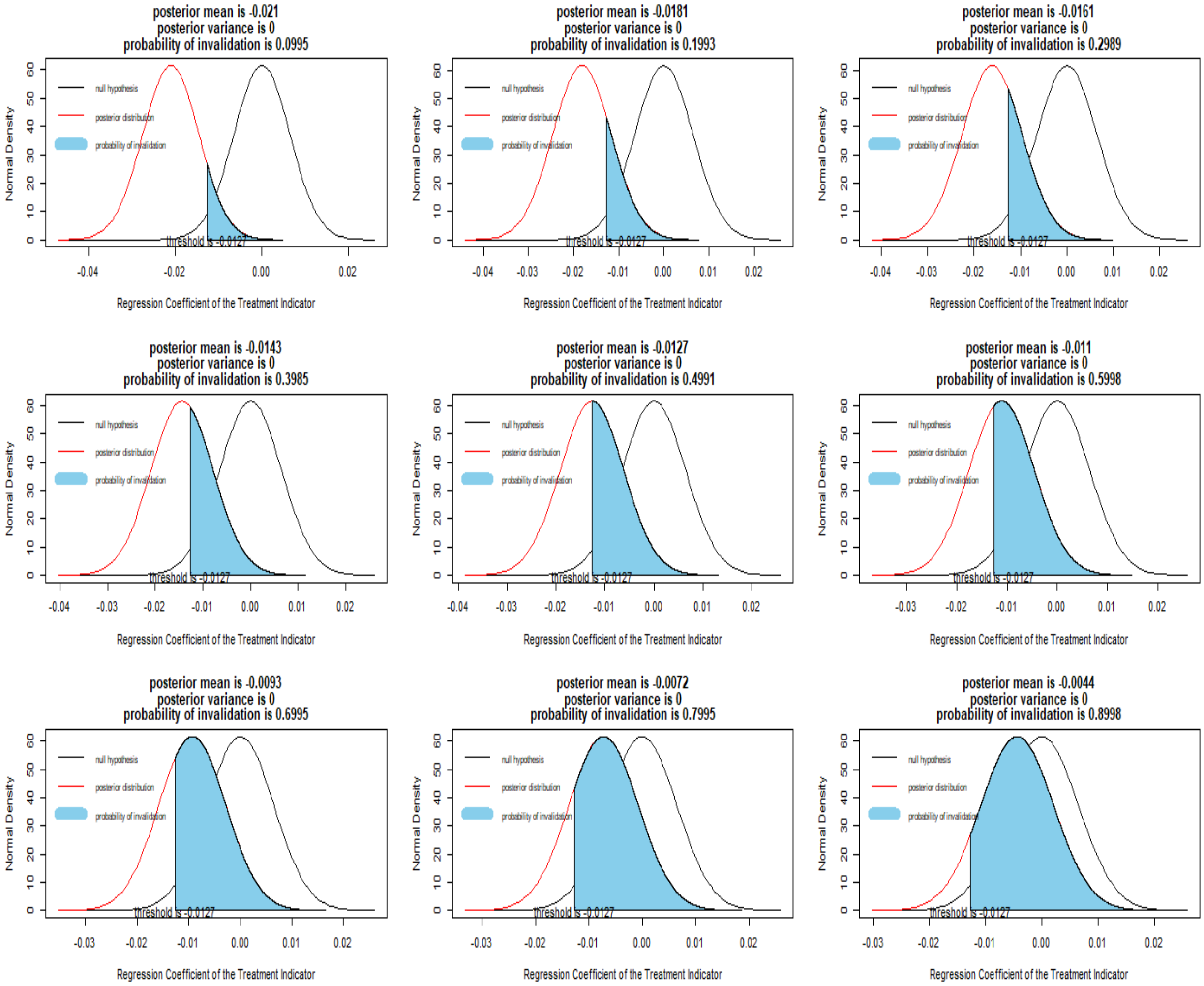


Figure 2.7: The relationship between testing null hypothesis and the posterior probability of invalidating the inference of Hong & Raudenbush (2005)



7-Discussion

7.1-A summary of the Bayesian paradigm of robustness indices for regression-based causal inference

To summarize, the Bayesian paradigm of robustness indices for regression-based causal inference centers on a probit model of the probability of invalidating an inference. It is a thought experiment built on the observed data. The observed sample and sample statistics pertaining to it are fixed throughout the analysis. Doing so is logical because the object of analysis of robustness is supposed to be a single analysis and its observed sample ought to be fixed, and this is consistent with the Bayesian reasoning that the same observed sample is fed to likelihood function irrespective of prior distribution. The analysis of robustness is a thought experiment as it impels a thorough and detailed conceptualization of an unobserved sample, which is thought of as a random sample from the unobserved part of ideal population and to be expressed by a prior. To accomplish this, a clear definition about the unobserved part of ideal population is firstly needed based on the research context. Moreover, possibly with a good knowledge about what a random sample of this unobserved part of ideal population would look like, assumptions about sample statistics of an unobserved sample are made and whenever plausible they are assumed to be equal to their observed counterparts. Most importantly, a few unobserved sample statistics are chosen as focal parameters based on a research question, and the probability of invalidating an inference is in an explicit probit relationship with focal parameters given all assumed parametric values and observed sample statistics. The learning goal of the analysis of robustness is to identify the thresholds of focal parameters that make the probability of invalidating an inference just below a desired value. It's worth emphasizing here that a comprehensive knowledge about the robustness of any single research cannot be gained without repeatedly computing the thresholds of focal parameters for a series of desired values and describing the probit curve between the probability of invalidating an inference and a focal parameter.

The Bayesian paradigm of robustness indices is consistent with the argument made by Frank & Min (2007), which proposed to treat prior as a distribution of parameter based on an unobserved sample. This indeed is how I frame the Bayesian paradigm of robustness indices for regression-based causal inference in this paper: The prior is defined as a distribution of regression coefficients based on an imaginary unobserved sample whose data structure has been formalized. The likelihood function is defined as a parametric distribution for the outcomes of target population and fit to the observed sample. Consequently, the posterior distribution of regression coefficients generated by this fashion has a form that is identical to the distribution of regression coefficients based on an ideal sample, which is just the consolidation of an unobserved sample and the observed sample. Built on such posterior distribution, the probit link of the probability of invalidating an inference is a function of prior parameters such as unobserved sample size, unobserved sample means and elements in unobserved sample variance-covariance matrix. Intrinsically, the analysis of robustness is an exploratory Bayesian sensitivity analysis where prior parameters are manipulated and thus their impacts on the probability of invalidating an inference can be learned.

Just as the frequentist recipe, the Bayesian paradigm of robustness indices could be interpreted as a two-phase sampling approach: The first phase refers to the analysis where the observed sample is randomly drawn from the observed part of ideal population and regression is carried out for the observed sample. The second phase refers to the analysis where an unobserved sample is randomly drawn from the unobserved part of ideal population and subsequently regression is performed for this unobserved sample. The distribution of regression coefficients in the second phase is equivalent to the prior in the Bayesian recipe and the distribution of regression coefficients for an ideal sample produced by this two-phase sampling is equivalent to the posterior in the Bayesian recipe.

7.2-Comparisons with other similar approaches

7.2.1-The impact thresholds in Frank (2000)

The impact of an unmeasured variable, defined as the product of the correlation between this variable and the focal predictor and the correlation between this variable and the outcome, is often the subject of a discussion about the robustness of a causal research. Frank (2000) derived the impact threshold for an unmeasured confounder or suppressor in a multiple regression. The logic is that, given the observed correlation between a focal variable (like the treatment indicator W) and the outcome, sample size and level of significance, the impact threshold can inform researchers that how large the impact of an unmeasured confounder/suppressor needs to be so that it can make an inference invalid. By definition, the impact threshold of an unmeasured confounder defines the boundary beyond which an original significant regression coefficient becomes insignificant. Moreover, the impact threshold of an unmeasured suppressor defines the boundary beyond which either an original significant regression coefficient becomes significant in the opposite direction or an original insignificant regression coefficient turns to be significant in either direction. The impact threshold helps conceptualization of the robustness of a causal research since it can be naturally extended to cases where a regression model has multiple covariates and multiple unmeasured confounders/suppressors and can be evaluated through a reference distribution (see Pan & Frank 2003, 2004 as well).

Logically, I approach the problem of causal inference and its robustness essentially the same way as Frank (2000) did. I perceive disputable causal inference as an inference based on insufficient information. In Frank (2000), the missing piece is the uncontrolled confounders/suppressors which have the potential to invalidate a regression-based causal inference. In the Bayesian paradigm of robustness indices, the missing piece is actually the missing data, which could be either a potential random sample from the unobserved part of ideal population or counterfactuals defined in Rubin Causal Model. Both approaches ask the same question “what would this missing piece have to be such that the current inference is no longer established?” By this logic, the threshold of a sufficient statistic or a parameter of main interest characterizing the missing piece will be pursued.

Even though both the impact threshold and the Bayesian paradigm of robustness indices reside in the context of regression, there a key difference between their perspectives: Bayesian paradigm of robustness indices emphasizes a sampling or missing data perspective, like discussed earlier. By defining and differentiating the observed and unobserved parts of ideal population, an unobserved sample can be conceptualized as a random sample from the unobserved part of ideal population. My robustness index, the probability of invalidating an inference, is built on this unobserved sample. Frank (2000) accentuates the variable selection problem for observational studies and quasi-experiments when the assumption of unconfoundedness is questioned. This is exactly the theme of the first scenario and Bayesian paradigm of robustness indices has offered a solution for it, though from a different perspective. Statistically speaking, Frank (2000) is a pure frequentist framework while Bayesian paradigm of robustness indices is a Bayesian framework with a supplementary frequentist viewpoint.

7.2.2-The robustness indices in Frank & Min (2007)

The Bayesian paradigm of robustness indices is a detailed and comprehensive expansion of the Bayesian framework proposed by Frank & Min (2007). The main principle of Frank & Min (2007) has been maintained throughout this paper: I treat prior as it is based on an unobserved sample and likelihood as it is shaped by the observed sample. The posterior distribution in Bayesian paradigm of robustness indices is proven to be a distribution based on an ideal sample, just as theorized in Frank & Min (2007). The Bayesian paradigm of robustness indices has a broader scope than Frank & Min (2007): It appeals to both research with limited internal validity and research with limited external validity, whereas the robustness indices of Frank & Min (2007) is designated for the research with limited external validity only.

7.2.3-The robustness indices in Frank et al. (2013)

The Bayesian paradigm of robustness indices could be deemed as a Bayesian version of Frank et al. (2013) as they share the same goal of assessing the robustness of research with strong internal validity but weak external validity as well as research with strong external validity but weak internal validity. Some key concepts of the Bayesian paradigm of robustness indices, such as the threshold for making an inference and the decision rule of invalidating an inference, are inherited from Frank et al. (2013). However, the Bayesian paradigm of robustness indices is more probabilistically oriented and requires a more precise and detailed modeling of an unobserved sample than Frank et al. (2013). The robustness index in Frank et al. (2013) is the proportion of the observed sample a research can afford to be replaced with an unobserved sample where the treatment effect is zero, without nullifying an inference. On the contrary, the robustness index in this paper is the probability of invalidating an inference, which is built on an imaginary ideal sample. This ideal sample is not formed by replacing a portion of the observed sample with an unobserved sample but by adding this unobserved sample to the existing observed sample.

7.3-Limitations

Contributory insights about the robustness of regression-based causal inference can be elicited by the proper application of Bayesian paradigm of robustness indices. Conversely, Bayesian paradigm of robustness indices can lead to misguiding results and baffling conclusions if researchers are unaware of its pitfalls and limitations. I warn readers of two major limitations of the Bayesian paradigm of robustness indices: First, the Bayesian paradigm of robustness indices demonstrated throughout this paper is well situated in the regression-based causal inference, which by definition is an approach of treating the estimated regression coefficient of the treatment indicator W as the estimate of average treatment effect, in a multiple regression where the outcome Y should be continuous (or at least not categorical). This makes the Bayesian paradigm of robustness indices inappropriate for statistical methods such as logistic regression, multinomial logistic regression or any other non-regression methods. It is also counterproductive to apply the Bayesian paradigm of robustness indices to research questions which cannot be answered by the regression coefficient of W. For example, a research seeking answers about the treatment effect for the treated or for the control cannot be simply satisfied with the regression coefficient of W. In general, the Bayesian paradigm of robustness indices is best applied to those two research scenarios, i.e., research with strong internal validity but weak external validity and research with weak internal validity but strong external validity, as long as they intend to find out the average treatment effect only.

Another limitation of the Bayesian paradigm of robustness indices is that it has no power assessing the robustness of a causal research that could be biased by factors other than weak internal validity or weak external validity. Factors such as measurement error and violation of the SUTVA assumption can and often jeopardize a causal inference. Nevertheless, they are beyond the scope of the Bayesian paradigm of robustness indices.

7.4-Conclusion

A causal relationship can never be established by merely one research. Rather, to confirm a causal relationship and accept it as gained scientific knowledge, much more assessments need to be done by experts in the substantive field and those assessments are typically “more demanding and meaningful than that of a one-time, stand-alone test of scientific value” (Sohn, 1998). It’s my wish that the Bayesian paradigm of robustness indices can equip causal researchers a framework which allows the assessments of the robustness of a causal inference to be done in a systematic, informative and organized fashion. I believe that the Bayesian paradigm of robustness indices has reflected an important and frequently mentioned recommendation emerged from the discussion of replicability/reproducibility, i.e., the consideration of the prior probabilities of hypotheses. This is exactly the spirit of analysis of robustness. By treating the prior distribution as a distribution built on an unobserved sample, the regression estimate of average treatment effect can be repeatedly evaluated and thereby the belief about a causal inference is updated, conditional on different hypotheses about an unobserved sample. As many researchers pointed out, assigning prior probabilities to all possible hypotheses is likely to be unavoidable in the journey from the long-criticized p-value to a meaningful index of replicability.

APPENDICES

Appendix A: Proofs of Theorem 1 and Theorem 2

Proof of Theorem 1:

My goal is to derive the formula for least square estimate of regression coefficient for W (i.e., $\hat{\gamma}_w$) based on an ideal sample since it is the posterior mean and the variance of $\hat{\gamma}_w$ given it is identical to the posterior variance, as manifested by (2.17).

First, I need to define the following ordered data matrices for an ideal sample:

$$\mathbf{D} = [\mathbf{Y}_{(n^{un}+n^{ob})\times 1}, \mathbf{X}_{(n^{un}+n^{ob})\times(p+2)}]$$
$$\mathbf{X} = [\mathbf{1}_{(n^{un}+n^{ob})\times 1}, \mathbf{V}_{(n^{un}+n^{ob})\times(p+1)}]$$
$$\mathbf{V} = [\mathbf{Z}_{(n^{un}+n^{ob})\times p}, \mathbf{W}_{(n^{un}+n^{ob})\times 1}]$$
$$\mathbf{Z} = [\mathbf{Z_1}, \mathbf{Z_2}, ..., \mathbf{Z_p}]_{(n^{un}+n^{ob})\times p}$$

and ordered mean vectors:

$$\overline{\mathbf{V}}^{\mathbf{id}} = [\overline{\mathbf{Z}}^{\mathbf{id}}, \overline{W}^{id}]_{1\times(p+1)}$$
$$\overline{\mathbf{Z}}^{\mathbf{id}} = [\overline{\mathrm{Z}}_1^{id}, \overline{\mathrm{Z}}_2^{id}, \cdots, \overline{\mathrm{Z}}_p^{id}]_{1\times p}$$

The matrix $\mathbf{X^T X}$ for an ideal sample could then be molded as the following block matrix:

$$\mathbf{X^T X} = \begin{pmatrix} n^{un}+n^{ob} & (n^{un}+n^{ob})\overline{\mathbf{V}}^{\mathbf{id}} \\ (n^{un}+n^{ob})(\overline{\mathbf{V}}^{\mathbf{id}})^{\mathbf{T}} & \mathbf{V^T V} \end{pmatrix}$$

It turns out that, the inverse of $\mathbf{X^T X}$ has the following form:

$$\mathbf{(X^TX)^{-1}} = \begin{pmatrix} \frac{1}{n^{un}+n^{ob}} + \mathbf{\bar{V}^{id}} \frac{1}{n^{un}+n^{ob}} \mathbf{(S^{id}_{VV})^{-1}(\bar{V}^{id})^T} & -\frac{1}{n^{un}+n^{ob}} \mathbf{\bar{V}^{id}(S^{id}_{VV})^{-1}} \\ -\frac{1}{n^{un}+n^{ob}} \mathbf{(S^{id}_{VV})^{-1}(\bar{V}^{id})^T} & \frac{1}{n^{un}+n^{ob}} \mathbf{(S^{id}_{VV})^{-1}} \end{pmatrix}$$

It should be clear now that, to determine the definite form of $\mathbf{(X^TX)^{-1}}$ I need to find out what $\mathbf{(S^{id}_{VV})^{-1}}$ is. As a variance-covariance matrix for the vector of predictors V, $\mathbf{S^{id}_{VV}}$ can be expressed as the block matrix in (3.5) whose elements is formalized in (3.6) through (3.8). Consequently, the inverse of $\mathbf{S^{id}_{VV}}$ can be formulated here:

$$\mathbf{(S^{id}_{VV})^{-1}} =$$

$$\begin{bmatrix} \mathbf{(S^{id}_{ZZ})^{-1} + (S^{id}_{ZZ})^{-1}S^{id}_{ZW}}(\hat{\sigma}^{id}_{WW} \mathbf{- S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}S^{id}_{WZ}(S^{id}_{ZZ})^{-1}} & \mathbf{-(S^{id}_{ZZ})^{-1}S^{id}_{ZW}}(\hat{\sigma}^{id}_{WW} \mathbf{- S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}} \\ \mathbf{-(}\hat{\sigma}^{id}_{WW} \mathbf{- S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}S^{id}_{WZ}(S^{id}_{ZZ})^{-1}} & (\hat{\sigma}^{id}_{WW} \mathbf{- S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}} \end{bmatrix}$$

Plugging the above matrix of $\mathbf{(S^{id}_{VV})^{-1}}$ into the block matrix of $\mathbf{(X^TX)^{-1}}$ will give us the complete definite form of matrix $\mathbf{(X^TX)^{-1}}$, whose elements are all ideal sample statistics such as ideal sample variances, ideal sample covariances and ideal sample means. To isolate the estimated regression coefficient for W, I only need to use the elements in the last row of $\mathbf{(X^TX)^{-1}}$, which are provided next:

$$\mathbf{(X^TX)^{-1}}_{(p+2)1} = \frac{1}{n^{un}+n^{ob}}[(\hat{\sigma}^{id}_{WW} \mathbf{- S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}S^{id}_{WZ}(S^{id}_{ZZ})^{-1}(\bar{Z}^{id})^T} - \bar{W}^{id}(\hat{\sigma}^{id}_{WW} \mathbf{- S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}}]$$

$$[\mathbf{(X^TX)^{-1}}_{(p+2)2},\cdots,\mathbf{(X^TX)^{-1}}_{(p+2)(p+1)}] = -\frac{1}{n^{un}+n^{ob}}(\hat{\sigma}^{id}_{WW} \mathbf{- S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}S^{id}_{WZ}(S^{id}_{ZZ})^{-1}}$$

$$\mathbf{(X^TX)^{-1}}_{(p+2)(p+2)} = \frac{1}{n^{un}+n^{ob}}(\hat{\sigma}^{id}_{WW} \mathbf{- S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}}$$

Because the estimated regression coefficient for W is the last element of $\mathbf{(X^TX)^{-1}X^TY}$ which is the dot product between the last row of $\mathbf{(X^TX)^{-1}}$ and $\mathbf{X^TY}$, the expression of $\mathbf{X^TY}$ is also needed here:

$$\mathbf{X^TY} = \begin{bmatrix} (n^{un}+n^{ob})\bar{Y}^{id} \\ \mathbf{Z^TY} \\ \mathbf{W^TY} \end{bmatrix}$$

where:

$$\mathbf{Z^TY} = (n^{un}+n^{ob})\mathbf{S^{id}_{ZY}} + (n^{un}+n^{ob})\bar{Y}^{id}\mathbf{(\bar{Z}^{id})^T}$$
$$\mathbf{W^TY} = (n^{un}+n^{ob})\hat{\sigma}^{id}_{WY} + (n^{un}+n^{ob})\bar{W}^{id}\bar{Y}^{id}$$

Now I can calculate the estimated regression coefficient for W as the dot product between the last row of $\mathbf{(X^TX)^{-1}}$ and the vector $\mathbf{X^TY}$. The result is presented below:

$$\hat{\gamma}_w = \frac{\hat{\sigma}^{id}_{WY} - \mathbf{S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZY}}}{\hat{\sigma}^{id}_{WW} - \mathbf{S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW}}}$$

The variance of $\hat{\gamma}_w$ should be straightforward: it is just the product of the known residual variance σ^2 and the element in the p+2th row and the p+2th column of $\mathbf{(X^TX)^{-1}}$:

$$Var(\hat{\gamma}_w) = \frac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}^{id}_{WW}\ \mathbf{- S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}}$$

Taken together, the posterior distribution of γ_w conditional on the observed sample is given by:

$$\gamma_w \,|\, \mathbf{X^{ob}, Y^{ob}} \sim N\left(\frac{\hat{\sigma}^{id}_{WY} - \mathbf{S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZY}}}{\hat{\sigma}^{id}_{WW} - \mathbf{S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW}}}, \frac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}^{id}_{WW} - \mathbf{S^{id}_{WZ}(S^{id}_{ZZ})^{-1}S^{id}_{ZW})^{-1}}\right)$$

The derivations of ideal variances/covariances as functions of observed and unobserved sample statistics follow the same reasoning. Here I just take the covariance between W and Y in an ideal sample as an example. First of all, I have:

$$\hat{\sigma}_{WY}^{id} = \frac{1}{n^{un} + n^{ob}} \sum_{i=1}^{n^{un}+n^{ob}} (W_i - \overline{W}^{id})(Y_i - \overline{Y}^{id})$$

The above equation can be rearranged and reexpressed as follows:

$$\sum_{i=1}^{n^{un}+n^{ob}} W_i Y_i = \left(n^{un} + n^{ob}\right)\hat{\sigma}_{WY}^{id} + \left(n^{un} + n^{ob}\right)\overline{W}^{id}\overline{Y}^{id} = \sum_{i=1}^{n^{un}} W_i Y_i + \sum_{i=1}^{n^{ob}} W_i Y_i$$

Similarly, the following equations are true for the observed sample and an unobserved sample:

$$\sum_{i=1}^{n^{un}} W_i Y_i = n^{un}\hat{\sigma}_{WY}^{un} + n^{un}\overline{W}^{un}\overline{Y}^{un}$$

$$\sum_{i=1}^{n^{ob}} W_i Y_i = n^{ob}\hat{\sigma}_{WY}^{ob} + n^{ob}\overline{W}^{ob}\overline{Y}^{ob}$$

The last three equations could be consolidated into an expanded one as follows:

$$\left(n^{un} + n^{ob}\right)\hat{\sigma}_{WY}^{id} + \left(n^{un} + n^{ob}\right)\overline{W}^{id}\overline{Y}^{id} = n^{un}\hat{\sigma}_{WY}^{un} + n^{ob}\hat{\sigma}_{WY}^{ob} + n^{un}\overline{W}^{un}\overline{Y}^{un} + n^{ob}\overline{W}^{ob}\overline{Y}^{ob}$$

Finally, the expression of $\hat{\sigma}_{WY}^{id}$ as a function of unobserved and observed sample statistics could be deduced from the equation above:

$$\begin{aligned}
\hat{\sigma}_{WY}^{id} &= \frac{n^{un}\hat{\sigma}_{WY}^{un} + n^{ob}\hat{\sigma}_{WY}^{ob}}{n^{un}+n^{ob}} + \frac{n^{un}\overline{W}^{un}\overline{Y}^{un} + n^{ob}\overline{W}^{ob}\overline{Y}^{ob}}{n^{un}+n^{ob}} - \overline{W}^{id}\overline{Y}^{id} \\
&= \frac{n^{un}}{n^{un}+n^{ob}}\hat{\sigma}_{WY}^{un} + \frac{n^{ob}}{n^{un}+n^{ob}}\hat{\sigma}_{WY}^{ob} + \frac{n^{un}}{n^{un}+n^{ob}}\overline{W}^{un}\overline{Y}^{un} + \frac{n^{ob}}{n^{un}+n^{ob}}\overline{W}^{ob}\overline{Y}^{ob} \\
&\quad - \left(\frac{n^{un}\overline{W}^{un} + n^{ob}\overline{W}^{ob}}{n^{un}+n^{ob}}\right)\left(\frac{n^{un}\overline{Y}^{un} + n^{ob}\overline{Y}^{ob}}{n^{un}+n^{ob}}\right) \\
&= \lambda\hat{\sigma}_{WY}^{un} + (1-\lambda)\hat{\sigma}_{WY}^{ob} + \lambda\overline{W}^{un}\overline{Y}^{un} + (1-\lambda)\overline{W}^{ob}\overline{Y}^{ob} \\
&\quad - \left[\lambda^2\overline{W}^{un}\overline{Y}^{un} + \lambda(1-\lambda)\overline{W}^{un}\overline{Y}^{ob} + \lambda(1-\lambda)\overline{W}^{ob}\overline{Y}^{un} + (1-\lambda)^2\overline{W}^{ob}\overline{Y}^{ob}\right] \\
&= \lambda\hat{\sigma}_{WY}^{un} + (1-\lambda)\hat{\sigma}_{WY}^{ob} + \lambda(1-\lambda)[\overline{W}^{un}\overline{Y}^{un} - \overline{W}^{un}\overline{Y}^{ob} - \overline{W}^{ob}\overline{Y}^{un} + \overline{W}^{ob}\overline{Y}^{ob}] \\
&= \lambda\hat{\sigma}_{WY}^{un} + (1-\lambda)\hat{\sigma}_{WY}^{ob} + (1-\lambda)\lambda(\overline{W}^{ob} - \overline{W}^{un})(\overline{Y}^{ob} - \overline{Y}^{un})
\end{aligned}$$

where:

$$\lambda = \frac{n^{un}}{n^{un}+n^{ob}}$$

Proof of Theorem 2:

For inferring a positive effect, the probability of invalidating an inference is:

$$P(\gamma_w < \gamma_w^{\#} \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}})$$

To recast this probability of invalidating an inference in terms of the cumulative distribution function of the standard normal distribution, I need to standardize the random variable $\gamma_w \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}}$:

$$P(\gamma_w < \gamma_w^{\#} \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}}) =$$

$$P\left(\frac{\gamma_w - \dfrac{\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}}{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}}{\sqrt{\dfrac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}}} < \frac{\gamma_w^{\#} - \dfrac{\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}}{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}}{\sqrt{\dfrac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}}} \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}} \right)$$

$$= \Phi\left(\frac{\gamma_w^{\#} - \dfrac{\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}}{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}}{\sqrt{\dfrac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}}} \right)$$

$$= \Phi\left(\frac{\sqrt{n^{un}+n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}} [\gamma_w^{\#}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}) - (\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}})] \right)$$

Given the probit function is just the inverse of the cumulative distribution function of the standard normal distribution, plugging either side of the above equation into the probit function will lead to the following result:

$$probit(p) = \frac{\sqrt{n^{un}+n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}} [\gamma_w^{\#}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}) - (\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}})]$$

For inferring a negative effect, the probability of invalidating an inference generally becomes:

$$P(\gamma_w > \gamma_w^{\#} \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}})$$

By the same logic, the probit function of this probability is approached by deriving its corresponding cumulative function of the standard normal distribution first:

$$P(\gamma_w > \gamma_w^{\#} \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}}) =$$

$$P\left(\frac{\gamma_w - \dfrac{\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}}{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}}{\sqrt{\dfrac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}}} > \frac{\gamma_w^{\#} - \dfrac{\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}}{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}}{\sqrt{\dfrac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}}} \mid \mathbf{X^{ob}}, \mathbf{Y^{ob}} \right)$$

$$= 1 - \Phi\left(\frac{\gamma_w^{\#} - \dfrac{\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}}{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}}{\sqrt{\dfrac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}}} \right) = \Phi\left(\frac{\dfrac{\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}}{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}} - \gamma_w^{\#}}{\sqrt{\dfrac{\sigma^2}{n^{un}+n^{ob}}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})^{\mathbf{-1}}}} \right)$$

$$= \Phi\left(\frac{\sqrt{n^{un}+n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}} [(\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}) - \gamma_w^{\#}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})] \right)$$

Apparently, taking probit operation on both sides of the equation above will generate the following probit function of the probability of invalidating an inference when inferring a negative effect:

$$probit(p) = \frac{\sqrt{n^{un}+n^{ob}}}{\sigma\sqrt{\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}}}} [(\hat{\sigma}_{WY}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}) - \gamma_w^{\#}(\hat{\sigma}_{WW}^{id} - \mathbf{S_{WZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZW}^{id}})]$$

This completes the proof of theorem 2.

Appendix B: The Algebraic Equivalence Between Theorem 1 and Common Expressions of Regression Coefficients

In this appendix, I will demonstrate the algebraic equivalence between theorem 1 and the common expressions of ordinary least square (OLS) estimates of simple regression coefficient as well as standardized multiple regression coefficients (for two covariates).

The common expression of simple regression coefficient is provided below:

$$\hat{\gamma}_x = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Now I show how theorem 1 is connected to the expression above: First, the distribution in theorem 1 can be treated as the distribution of any single regression coefficient. Therefore, for a predictor x its OLS estimate of regression coefficient is provided by theorem 1 as follows:

$$\hat{\gamma}_x = \frac{\hat{\sigma}_{xy}^{id} - \mathbf{S_{XZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZY}^{id}}}{\hat{\sigma}_{xx}^{id} - \mathbf{S_{XZ}^{id}}(\mathbf{S_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{S_{ZX}^{id}}}$$

However, there is no covariates **Z** in a simple regression model and thus any sample variance-covariance matrices (or vectors) involves **Z** will be cancelled, which means the matrices and vectors $\mathbf{S_{ZZ}^{id}},\mathbf{S_{XZ}^{id}},\mathbf{S_{ZY}^{id}},\mathbf{S_{ZX}^{id}}$ are all cancelled and the above expression of $\hat{\gamma}_x$ becomes:

$$\hat{\gamma}_x = \frac{\hat{\sigma}_{xy}^{id}}{\hat{\sigma}_{xy}^{id}}$$

Based on the formulae of computing sample variances and sample covariances in (3.9), the following equations can be derived:

$$\hat{\gamma}_x = \frac{\hat{\sigma}_{xy}^{id}}{\hat{\sigma}_{xy}^{id}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

This establishes the algebraic equivalence between theorem 1 and the common expression of OLS estimate of simple regression coefficient. I note here in this case $n = n^{un} + n^{ob}$ because all sample statistics pertain to an ideal sample.

Next, I prove the algebraic equivalence between theorem 1 and the expressions of the OLS estimates of standardized multiple regression coefficients, through an example of regressing y on x$_1$ and x$_2$. It is well known that, the OLS estimates of standardized regression coefficients of x$_1$ and x$_2$ have the following forms:

$$\hat{\gamma}_{x_1} = \frac{r_{x_1 y}^{id} - r_{x_1 x_2}^{id} r_{x_2 y}^{id}}{1 - (r_{x_1 x_2}^{id})^2}$$

$$\hat{\gamma}_{x_2} = \frac{r_{x_2 y}^{id} - r_{x_1 x_2}^{id} r_{x_1 y}^{id}}{1 - (r_{x_1 x_2}^{id})^2}$$

The OLS estimate of any single standardized multiple regression coefficient has been offered by theorem 3 (which is the standardized version of theorem 1) as follows:

$$\hat{\gamma}_x = \frac{r_{xy}^{id} - \mathbf{R_{XZ}^{id}}(\mathbf{R_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{R_{ZY}^{id}}}{1 - \mathbf{R_{XZ}^{id}}(\mathbf{R_{ZZ}^{id}})^{\mathbf{-1}}\mathbf{R_{ZX}^{id}}}$$

To isolate the expression of OLS estimate of regression coefficient of x$_1$ from the above formula, one needs to treat x$_1$, x$_2$ and y as x, Z and y in the context of theorem 1 (or theorem 3) and consequently the facts below are learned:

$$r_{xy}^{id} = r_{x_1 y}^{id}$$
$$\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{id}} = 1 = r_{x_2 x_2}^{id}$$
$$\mathbf{R}_{\mathbf{XZ}}^{\mathbf{id}} = \mathbf{R}_{\mathbf{ZX}}^{\mathbf{id}} = r_{x_1 x_2}^{id}$$
$$\mathbf{R}_{\mathbf{ZY}}^{\mathbf{id}} = r_{x_2 y}^{id}$$

With the above facts, the OLS estimate of regression coefficient of x$_1$ offered by theorem 3 can be rewritten as:

$$\hat{\gamma}_{x_1} = \frac{r_{x_1 y}^{id} - r_{x_1 x_2}^{id} r_{x_2 y}^{id}}{1-(r_{x_1 x_2}^{id})^2}$$

Similarly, one should treat x$_1$, x$_2$ and y as Z, x and y in the context of theorem 1 (or theorem 3) and subsequently acknowledge the following facts in order to derive the OLS estimate of regression coefficient of x$_2$ from theorem 3:

$$r_{xy}^{id} = r_{x_2 y}^{id}$$
$$\mathbf{R}_{\mathbf{ZZ}}^{\mathbf{id}} = 1 = r_{x_1 x_1}^{id}$$
$$\mathbf{R}_{\mathbf{XZ}}^{\mathbf{id}} = \mathbf{R}_{\mathbf{ZX}}^{\mathbf{id}} = r_{x_1 x_2}^{id}$$
$$\mathbf{R}_{\mathbf{ZY}}^{\mathbf{id}} = r_{x_1 y}^{id}$$

The OLS estimate of regression coefficient of x$_2$ is then straightforward:

$$\hat{\gamma}_{x_2} = \frac{r_{x_2 y}^{id} - r_{x_1 x_2}^{id} r_{x_1 y}^{id}}{1-(r_{x_1 x_2}^{id})^2}$$

Now I observe that the derived expressions of $\hat{\gamma}_{x_1}$ and $\hat{\gamma}_{x_2}$ based on theorem 3 are identical to their common expressions. Therefore, I can confirm the algebraic equivalence between theorem 3 (also theorem 1) and common expressions of standardized multiple regression coefficients.

REFERENCES

REFERENCES

Borman, G. D., Dowling, B. M., & Schneck, C. (2008). A multi-site cluster randomized field trial of open court reading. *Educational Evaluation and Policy Analysis,* 30, 389–407.

Diaconis, P., & Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, 7, 269–281

Diaconis, P., & Ylvisaker, D. (1985). Quantifying prior opinion. *Bayesian statistics*, 2, 133–156.
Espinosa, V., Dasgupta, T., & Rubin, D. B. (2016). A Bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes. *Technometrics*, 58(1), 62-73.

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, 29(2), 147-194.

Frank, K. A., & Min, K. (2007). Indices of robustness for sample representation. *Sociological Methodology,* 37, 349–392.

Frank, K. A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., & McCrory, R. (2008). Extended influence: National board certified teachers as help providers. *Education Evaluation and Policy Analysis,* 30, 3–30.

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. *Education Evaluation and Policy Analysis,* 35, 437–460.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Greenwald, A., Gonzalez, R., Harris, R., & Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology*, *33*(2), 175-183.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York, NY: Springer Science & Business Media.

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, 205–224.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1), 305-327.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York, NY: Cambridge University Press.

Iverson, G. J., Wagenmakers, E. J., & Lee, M. D. (2010). A model-averaging approach to replication: the case of p rep. *Psychological Methods*, *15*(2), 172.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological science*, *16*(5), 345-353.

Lee, D. S. (2009). Training, wages and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76, 1071-1102.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. The American Economic Review, 80, 319-323.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*(4), 617-640.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, UK: Cambridge University Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, Vol. 349, Issue 6251, DOI: 10.1126/science.aac4716

Psychological Science editorial board. (2005). Information for contributors. *Psychological Science*, 16(12).

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1), 34-58.

Rubin, D. B. , & Zell, E. R. (2010). Dealing with noncompliance and missing outcomes in a randomized trial using Bayesian technology: Prevention of perinatal sepsis clinical trial, Soweto, South Africa. *Statistical Methodology*, 7, 338-350.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association*, 103, 1334–1344.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational design*. American Educational & Reseach Association.

Sohn, D. (1998). Statistical significance and replicability Why the former does not presage the latter. *Theory & Psychology*, 8(3), 291-311.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15, 250-267.

Thompson, B. (1996). Research news and comment: AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educational Researcher*, 25(2), 26-30.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd edition). Cambridge, MA: MIT Press.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th edition). Mason, OH: South-Western Cengage Learning.

Zajonc, T. (2012). Bayesian inference for dynamic treatment regimes: Mobility, equity, and efficiency in student tracking. *Journal of the American Statistical Association*, 107, 80-92.