The Dissertation Committee for Megha Joshi
Certifies that this is the approved version of the following Dissertation:

Cluster Wild Bootstrapping to Handle Dependent Effect Sizes with Small Sample Sizes

Committee:

Susan N. Beretvas, Supervisor

James E. Pustejovsky, Co-Supervisor

Seung W. Choi

Tiffany A. Whittaker

Elizabeth Tipton

Cluster Wild Bootstrapping to Handle Dependent Effect Sizes with Small Sample Sizes

by

Megha Joshi

Dissertation

Presented to the Faculty of the Graduate School
of The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin

Dec 2020

This thesis is dedicated to Tweets, Toshie, Oreo and Yori.

Acknowledgments

I am thankful for my two supervisors for this dissertation, Tasha and James. Tasha, you inspire me so much! You are an amazing teacher and mentor, and also a great leader. I am so grateful for the way that you have stood up for student concerns and for always showing compassion. James, I feel so rootless now that you are not here. You have been an amazing mentor and been so supportive of my work.

I am also thankful for my committee members. Dr. Choi, thanks for teaching me psychometrics. Your lectures were so in-depth. I miss your jokes during classes. Dr. Whittaker, thanks for leading this department through a rough transition year. I know it must have been so difficult. Thanks for always being there for us. Thank you also for your teaching. I learned foundational statistical methods in your class that have helped me in my research and teaching till now. Dr. Tipton, I have never really met you but thank you for your work! It is the basis of my dissertation.

Thank you Tweets, Toshie, Yori and Oreo. You have gotten me through many rough days. Thanks for sleeping on my laptop and books when I was trying to work on my dissertation. I always appreciate your love and playfulness.

Thank you to my family. I love you.

Thank you, Danny. I cannot imagine going through graduate school without without you. I am grateful that we got to work on homework together, play videogames, work out over Skype during quarantine, and volunteer for Bernie. Thank you to all my other colleagues in the Quantitative Methods program. Kejin, Bethany, Jihyun, RAZ, Pierce, Molly, Man, Young Ri, Suhwa, Sook Hyun, Gleb, the other Danny, Chris, and Luping: this journey would not have been as fun without you all.

Thank you, SangSuk, for always encouraging me to keep going even when I send so many messages to you saying I want to quit. You were the first person to teach me how to analyze data and to code. Thank you for your mentoring and your friendship.

Thank you, Bernie! Solidarity!

Abstract

Cluster Wild Bootstrapping to Handle Dependent Effect Sizes with Small Sample Sizes

Megha Joshi, Ph.D.
The University of Texas at Austin, 2020

Supervisors: Susan N. Beretvas and James E. Pustejovsky

Indent and begin abstract here. It should be a concise statement of the nature and content of the ETD. The text must be either double-spaced or 1.5-spaced. Abstracts should be limited to 350 words.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Meta-analysis is a set of statistical tools for synthesizing results from multiple primary studies on a common topic. The three major goals of meta-analysis include summarizing the results across the studies using some form of an effect size measure, characterizing the amount of variation in effects, and explaining variation in effect sizes (Hedges et al., 2010).

Tanner-Smith and Lipsey (2015) is an example of a published meta-analysis. The study evaluated the effectiveness of brief alcohol interventions for adolescents and young adults. The authors included 185 study samples in their analysis. The results from the study reported the overall effect size estimate, which identified whether brief alcohol interventions reduced consumption and alcohol related problems. The reported estimate indicated that the interventions led to statistically significant reduction in the outcomes. The authors also tested whether the intervention effects persisted over time and whether the effects varied by demographics of participants, intervention length and format. The results showed that the effects persisted up to one year after the interventions and did not vary across demographic characteristics of the participants, or intervention length or format.

Because meta-analysis involves synthesizing evidence from multiple primary studies, the results from a meta-analysis can have meaningful implications on policy evaluation in terms of funding interventions or policies, targeting interventions towards certain demographics, or re-developing interventions that may not be effective. The magnitude and direction of the pooled effect sizes can inform whether an intervention resulted in meaningful change in the outcome. If results from a meta-analytic study show that the effect of some intervention differs for certain populations or conditions, meta-analytic tools can be used to determine how it differs and under which conditions. For example, Tanner-Smith and Lipsey (2015) showed that the brief alcohol interventions were effective, hence, providing evidence for funding such interventions. The effects of the interventions did not differ in terms of demographics indicating that the interventions need not be modified to target certain demographics.

Typical methods to conduct meta-analysis—pooling effect sizes or analyzing moderating effects with meta-regression—involve an assumption that each effect size is independent. However, primary studies often report multiple estimates of effect sizes. For example, Tanner-Smith and Lipsey (2015), included studies that reported multiple correlated measures of the outcome variable, and thus had multiple dependent effect size estimates per study. Dependence can occur through two broad structures: correlated effects and hierarchical effects. Correlated effects typically result from multiple correlated measures of an outcome, repeated measures of the outcome data or the comparison of multiple treatment groups to the same control group (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015). Hierarchical effects can occur when primary meta-analytic studies include multiple experiments conducted by the same laboratory or in the same region creating dependence between the effect size parameters (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015).

The use of meta-analysis methods that ignore dependence when either of the types of dependence is present can result in inaccurate standard errors and therefore, hypothesis tests with incorrect Type 1 error rates and confidence intervals with incorrect coverage levels (Becker, 2000). Ad-hoc methods include averaging effect sizes by study or randomly selecting an effect size for each study. These methods result in loss of information and are not suited to study within-study variation in effect sizes. A method called shifting-of-unit analysis involves running meta-analytic models for different subsets of the data (Cooper, 1998). However, this strategy is not useful if a researcher wants to summarize effects across the subsets or study differential effects (Cooper, 1998).

The ideal solution to handling dependence would be to use a multivariate model (Hedges et al., 2010). This approach explicitly models covariances among effect size estimates (Becker, 2000; Hedges et al., 2010; Tipton, 2015). However, multivariate meta-analysis requires knowledge of correlations or covariances between pairs of effect sizes within each study, which are often difficult to obtain from primary studies (Olkin & Gleser, 2009).

Hedges et al. (2010) proposed the use of robust variance estimation (RVE) to handle dependent effect sizes. To use RVE, researchers do not need to know the covariance structure between effect sizes like when using multivariate analyses. Instead, RVE involves estimating the variances for the meta-regression model's coefficients using sandwich estimators (Hedges et al., 2010; Tipton, 2015). RVE is increasingly being used in applied meta-analyses (Tipton, 2015). However, the performance characteristics of RVE are asymptotic—a large number of clusters or studies is required to provide accurate standard error estimates (Tipton, 2015). If the number of studies in a meta-analysis is small, RVE, as originally proposed by Hedges et al. (2010), can result in downwardly biased standard errors and inflation of Type 1 error rates (Hedges et al., 2010; Tipton, 2015).

Tipton (2015) introduced several small sample corrections for RVE for tests of single coefficients. The author proposed three different adjustment methods: one based on degrees of freedom correction provided by Hedges et al. (2010); one based on the jack-knife technique; and one based on a bias reduced linearization method proposed by McCaffrey et al. (2001). Simulation results from Tipton (2015) showed that the methods, when used without further corrections, resulted in Type error rate inflation. Tipton (2015) suggested using Sattherthwaite degrees of freedom along with the adjustment methods. The simulation results showed that the bias reduced linearization method with Satterthwaite degrees of freedom resulted in close to nominal Type 1 error rates. Moreover, Tipton (2015) showed that small sample size itself was not the only important factor that could influence the performance of RVE. Distribution of the covariates (imbalanced categories or outliers in covariates) can also influence the degrees of freedom, which in turn can influence the Type 1 error rate.

Tipton and Pustejovsky (2015) extended Tipton (2015) and introduced small sample corrections for multiple contrast hypotheses. In primary meta-analysis, multiple contrast hypotheses can be important parts of the research aims; analysts might want to learn whether effect sizes are the same across different research designs or for different populations. Tipton and Pustejovsky (2015) compared several methods based on Hotelling T^2 distribution and on eigen-decomposition. Based on the results of their simulation study, the authors recommended a method (AHZ) which approximates the test statistic using Hotelling's T^2 distribution with degrees of freedom proposed by Zhang (2012, 2013). The results from Tipton and Pustejovsky (2015) showed that AHZ had Type 1 error rates closest to the nominal rate of .05. However, the estimated Type 1 error rates for AHZ were below the nomial rate for tests of multiple contrast hypotheses.

In this dissertation, I will examine an alternative method, cluster wild bootstrapping, which has been examined in econometrics literature to handle dependence when the number of clusters is small (Cameron et al., 2008). General bootstrapping is a semi-parametric computational technique to calculate error in estimates without relying on distributional assumptions (Boos et al., 2003; Cameron et al., 2008). Bootstrapping involves re-sampling with replacement from the original data several times, calculating the estimate of interest on each replicate, and estimating the standard deviation across those estimates to derive the standard error and calculate confidence intervals (Boos et al., 2003).

Cluster wild bootstrapping specifically has been studied in econometrics literature as a possible solution to handle clustered data when the number of clusters is small Cameron et al. (2008). It involves re-sampling residuals which are generated from a null hypothesis model and randomly switching the signs of the residuals for each cluster from negative to positive. These new residuals are used to generate new outcome variables, which are used to derive the estimates for each bootstrap replicate (Cameron et al., 2008). Cluster wild boostrapping has been shown to adequately control Type 1 error rate for clustered data in regression analyses; however, it has not been examined methodologically in a meta-analytic framework (Cameron et al., 2008; MacKinnon & Webb, 2018).

Although it has not been examined methodologically, cluster wild bootstrapping has been used in a handful of meta-analytic studies with dependent effect sizes and small number of studies. Examples include McEwan (2015), in Review of Educational Research, examining school-based interventions on learning in developing countries; Gallet and Doucouliagos (2014), in the Annals of Tourism Research, examining income elasticity of air travel; and Oczkowski and Doucouliagos (2015), in the American Journal of Agricultural Economy, examining the relationship between price of wine and its quality. All of these studies used cluster wild bootstrapping to handle dependence for tests of single coefficients, not for multiple contrast hypotheses tests.

I will conduct simulation studies to examine whether using cluster wild bootstrapping improves upon the performance of the AHZ test in terms of Type 1 error rate and power, for tests of single meta-regression coefficients and of multiple contrast hypotheses tests. As a part of my dissertation, I also propose to build a package or contribute a function to the clubSandwich package in R that will implement the cluster wild bootstrapping algorithm especially for meta-analysis (Pustejovsky, 2020). My study will show whether cluster wild bootstrapping is an adequate alternative to the AHZ test.

Chapter 2

Literature Review

2.1 Research Synthesis

Meta-analysis is a quantitative technique to synthesize results across multiple primary studies. Examples include synthesis of the effects of some educational intervention like Head Start or the effects of brief alcohol interventions on alcohol consumption (Mann, 1994; Tanner-Smith & Lipsey, 2015). Meta-analysis can be used to synthesize effect estimates from randomized studies or quasi-experimental studies; it can also be used to synthesize correlations between variables from descriptive studies (Swanson et al., 2003).

Scientific researchers produce literature on the same topic perhaps to replicate or extend prior studies or due to lack of awareness of prior evidence (Hedges & Cooper, 2009). Results across studies tend to vary even when researchers try to replicate studies (Hedges & Cooper, 2009). Results can vary due to sample characteristics or differences in methodologies and research design (Hedges & Cooper, 2009). The three broad goals of meta-analysis include pooling or averaging effects from the primary studies, characterizing variability in the effects, and explaining the variability. Meta-analytic findings can have implications for policy-makers in education and social sciences. Meta-analysis can provide an overall impact or effect of some intervention or program and can explain whether the intervention is working for the intended population and under what conditions.

As summarized in the Introduction, Tanner-Smith and Lipsey (2015) found that brief alcohol interventions reduced alcohol consumption and alcohol related problematic behaviors, and that the effect did not vary by demographics and length or format of treatment. Results from this study can inform funding and development of these interventions. Murray et al. (2012) conducted a meta-analysis on 40 studies that examined the associations between parental incarceration and children's antisocial behavior, mental health problems, drug use and educational performance. The results showed that parental incarceration was associated with antisocial behavior but not with mental health problems, drug use or educational performance. The authors mentioned that their results should inform criminal justice system and national support services to help these children.

2.2 Types of Effect Sizes

The statistical synthesis of results involves summarizing across the effect sizes reported in or derived from primary studies. Effect sizes are quantitative measure of relationships among variables (Hedges, 2008). Hedges (2008) noted that effect sizes measure the degree to which the null hypothesis that two variables are unrelated is false. Effect sizes capture the magnitude or strength of the relationship between variables (Hedges, 2008). The p-values calculated from tests are dependent on the test statistics, which are dependent on sample sizes (Hedges, 2008). Therefore, p-values are not comparable across studies. A small p-value does not necessarily indicate a large effect (Hedges, 2008). However, effect sizes depend on population parameters and not on sample sizes and thus, are comparable across studies (Hedges, 2008). Common measures of effect size include standardized mean differences, correlations, difference in proportions and odds ratios (Hedges, 2008; Hedges & Cooper, 2009).

2.3 Pooling Effect Sizes

The first major goal of meta-analysis is to calculate a pooled or averaged effect size across multiple studies to derive the overall effects of an intervention or the overall measure of relationship between two variables. Let m denote the number of studies in a meta-analysis, each contributing one effect size, T_i for $i = 1, ..., m$. The variance estimate, $\hat{\sigma}_i^2$, denotes sampling error. One way to pool effect sizes is by weighing them by the inverse of their variance estimates. The inverse variance weights denote the imprecision of the estimated effect sizes (Viechtbauer, 2007). Below let w_i indicate the weights where $w_i = 1/\hat{\sigma_i}^2$. Effect sizes can be pooled as follows:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \tag{2.1}$$

2.3.1 Common Effects and Fixed Effects Models

Rice et al. (2018) outline several assumptions that can be made when pooling the effect sizes. The first is identical parameters assumption underlying the common effects model. This assumption states that one true effect size underlies all of the studies (Rice et al., 2018). The second assumption is the independent parameters assumption underlying the fixed effects model. This assumption treats the set of studies as all of the studies in the population of interest (Rice et al., 2018). The inferences derived from a fixed effect analysis would be valid for the specific set of studies included in the analysis (Rice et al., 2018).

2.3.2 Random Effects Model

Unlike the common effects and fixed effects model, the random effects model treats the set of effect sizes as a sample of all possible effect sizes in the population of interest (Higgins et al., 2009; Konstantopoulos & Hedges, 2019). The variance of the effect sizes between studies is denoted by τ^2 (Higgins et al., 2009; Konstantopoulos & Hedges, 2019). The pooled effect is still a weighted average of the effects. However, the weights account for between study variance as well as sampling error, $w_i = (\hat{\tau}^2 + {\hat{\sigma}_i}^2)^{-1}$ (Konstantopoulos & Hedges, 2019).

2.4 Characterizing Variability

Pooling effect sizes provides an overall estimate of the effect of the intervention or relationships between some variables. However, researchers are also interested in the existence of variation in the effect sizes (Konstantopoulos & Hedges, 2019). To test whether a common population effect size underlies effect sizes in a meta-analysis, the Cochran's Q statistics can be used. It characterizes the sum of squared differences between individual study effects and the pooled effect across studies (Konstantopoulos & Hedges, 2019). The individual effects and the pooled effect are weighted by the inverse variance weights as described above.

$$Q = \sum_{i=1}^{w} w_i T_i^2 - \frac{\left(\sum_{i=1}^{k} w_i T_i\right)^2}{\sum_{i=1}^{k} w_i} \tag{2.2}$$

The Q statistics is compared to a χ^2 distribution with $k - 1$ degrees of freedom (Konstantopoulos & Hedges, 2019).

Another statistics that characterizes variability in the effect sizes is I^2 (Higgins

& Thompson, 2002). It is a descriptive statistics that denotes the percentage or proportion of variance in observed effect size estimates that is due to variation in the true effect sizes (Borenstein et al., 2017). Borenstein et al. (2017) contended that I^2 is not an absolute measure of heterogeneity. Konstantopoulos and Hedges (2019) provided the following formula for calculating I^2:

$$I^2 = 100\% \times \frac{Q - (k-1)}{Q} \tag{2.3}$$

Under the random effects model, the total variability in effect sizes is a sum of sampling error, $\hat{\sigma_i}^2$, and between study variance, $\hat{\tau}^2$ (Konstantopoulos & Hedges, 2019). The $\hat{\tau}^2$ value is a descriptive statistic that denotes the variation in the true effects or as, Viechtbauer (2007) described it, the variance underlying the random variable producing the true effect sizes. There are various estimators suggested in the meta-analytic literature for the between-study variance component τ^2 (Viechtbauer, 2010). Estimators include the DerSimonian-Laird estimator (DerSimonian & Laird, 1986), the Hunter-Schmidt estimator (Hunter & Schmidt, 2004), the Hedges estimator (Hedges & Olkin, 2014; Raudenbush, 2009), the Sidik-Jonkman estimator (Sidik & Jonkman, 2005), the maximum-likelihood and restricted maximum likelihood estimator (Raudenbush, 2009; Viechtbauer, 2005), and the empirical Bayes estimator (Berkey et al., 1995; Morris, 1983). For a detailed comparison of the estimators, please see Veroniki et al. (2016).

2.5 Explaining Variability

In addition to pooling effect sizes and characterizing variability in effect sizes, meta-analysts often want to examine what factors explain or are associated with the variability. For example, the major questions in Tanner-Smith and Lipsey (2015) included whether the effect of the brief alcohol interventions differed for different demographic groups and whether the effect differed based on the length and format of the intervention. Such questions can clarify whether to target the intervention towards vulnerable groups, whether the intervention is effective for groups of interest, and whether to develop or change the intervention further to be more effective.

To explain variability in the effect sizes, a meta-regression model is generally used. The model, defined below, can be used to estimate an overall effect size and also to estimate moderating effects that characterize whether effect sizes vary based on the levels of another variable. Let T_i denote effect size i, $x_{i1}..x_{ip}$ denote the set of covariate values for effect size i, $\beta_1...\beta_p$ denote vector of regression coefficients, and ϵ_i denote the error term.

$$T_i = x_{i1}\beta_1 + ... + x_{ip}\beta_p + \epsilon_i \tag{2.4}$$

An intercept-only model can be used to estimate pooled effect sizes with weighted least squares estimation.

2.6 Dependence

Until this point, the models described assume one effect size per study. However, in applied meta-analysis, oftentimes each study can yield more than one effect size. For example, Tanner-Smith and Lipsey (2015) had multiple effect sizes per study because primary studies evaluated more than one measure of the two outcomes of interest, alcohol consumption and alcohol related problems. The existence of multiple effect sizes per study causes dependence in the effect sizes, which needs to be accounted for when conducting meta-analyses.

2.6.1 Meta-Regression: Multiple Effect Sizes

To account for multiple effect sizes per study, Equation 2.4 can be re-written as follows. Let $\mathbf{T}_j$ denote a $k_j \times 1$ vector of effect size estimates from study j, $\boldsymbol{\beta}$ denote a $p \times 1$ vector of coefficients, $\mathbf{X}_j$ denote a $k_j \times p$ matrix of covariates and $\boldsymbol{\epsilon}_j$ denote a $k_j \times 1$ vector of errors with mean of 0 and covariance matrix $\boldsymbol{\Psi}_j$ for studies $j = 1, .., m$. The meta-regression model is as follows:

$$\mathbf{T}_j = \mathbf{X}_j\boldsymbol{\beta} + \boldsymbol{\epsilon}_j \tag{2.5}$$

An intercept-only model can be used to estimate the overall effect size. $\mathbf{X}$ can include categorical and quantitative variables (Tipton, 2015). The covariates can be study-level variables, ones that vary between studies, or effect size level variables, ones that could vary within studies. Examples of covariates include percent of sample that was female for each study, or the outcome that was used to measure the effect.

Let $\mathbf{W}$ denote a bock diagonal matrix of weights. And let $\mathbf{M} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$. The weighted least squares estimate of $\boldsymbol{\beta}$ in Equation 2.5 can be calculated as:

$$\mathbf{b} = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j^{'}\mathbf{W}_j\mathbf{T}_j\right) \tag{2.6}$$

The exact variance of $\mathbf{b}$ is:

$$\text{Var}(\mathbf{b}) = \mathbf{M}\left[\sum_{j=1}^{m} \mathbf{X}_j{}^{'}\mathbf{W}_j\boldsymbol{\Psi}_j\mathbf{W}_j\mathbf{X}_j\right]\mathbf{M} \tag{2.7}$$

In meta-analysis models where each study only contributes one effect size and the studies and effect sizes can be assumed to be independent, the diagonal elements of $\boldsymbol{\Psi}_j$ are the variances of the effect sizes, v_j, and the off-diagonal elements are the covariances, which are assumed to be zero (Tanner-Smith et al., 2016). The fixed and random-effects model treat $\boldsymbol{\Psi}_j$ as known or partially known (Hedges et al., 2010). Using fixed or random effects model to estimate the meta-regression model without account for dependence can result in incorrect standard errors and thus, incorrect inference from hypothesis test results (Hedges et al., 2010).

When there are multiple effect sizes present per study, the meta-analytic data typically follows two broad structures: (1) correlated effects, and (2) hierarchical effects.

2.6.2 Correlated Effects

Correlated effects occur when there is dependence within cluster. This dependence structure occurs when the same study collects: (1) multiple correlated measures of outcomes; (2) repeated measures of outcomes; (3) outcome measures when multiple treatment groups are compared to the same control group; and, (4) multiple correlations from the same sample (Hedges et al., 2010; Tipton, 2015). In the correlated effects model, the dependence occurs through the error terms (Hedges, 2009). Under the correlated effects model:

$$T_{ij} = \theta_{ij} + e_{ij} \quad \text{with} \quad e_{ij} \sim N(0, \sigma_{ij}^2) \tag{2.8}$$

Here, T_{ij} is effect size i in study j, θ_{ij} is the true effect size i in study j, and e_{ij} is the error term that is normally distributed with mean of 0 and variance of σ_{ij}^2. The correlation between two error terms, h and i, in study j is assumed to be $\text{corr}(e_{hj}, e_{ij}) = \rho$. In the correlated effects model,

$$\theta_{ij} = \gamma_j \tag{2.9}$$

$$\gamma_j = \mu + vj \quad \text{with} \quad v_j \sim N(0, \tau^2) \tag{2.10}$$

Here γ_j is the overall effect size for study j, μ is the overall effect size across all the studies and v_j is the within study sampling error term that is normally distributed with mean of 0 and variance of τ^2.

Equations 2.8 and 2.9 imply that:

$$T_{ij} = \mu + v_j + e_{ij} \tag{2.11}$$

The above formulation implies that the effect sizes have the following marginal variance-covariance matrix (Hedges et al., 2010; Tipton & Pustejovsky, 2015):

$$\mathbf{\Psi}_{cj} = \tau^2 \mathbf{J}_j + \rho v_j \left(\mathbf{J}_j - \mathbf{I}_j\right) + v_j \mathbf{I}_j \tag{2.12}$$

Here,$\mathbf{\Psi}_{cj}$ indicates the covariance matrix according to the correlated effects model for study j, τ^2 denotes between-study variance in the average effect sizes, $\mathbf{J}_j$ denotes a matrix of 1's, $\mathbf{I}_j$ denotes an identity matrix, ρ is the assumed correlation between effect sizes which is assumed to be constant across studies, and v_j denotes the within study sampling error for study j assumed to be constant within studies (Hedges et al., 2010; Tipton & Pustejovsky, 2015). For the correlated effects model, weights can be calculated as (Tipton & Pustejovsky, 2015):

$$\mathbf{W}_j = \frac{1}{k_j \left(v_{\cdot j} + \tau^2\right)} \mathbf{I}_j \tag{2.13}$$

Here, k_j denotes the number of effect sizes in study j and $v_{\cdot j}$ denotes the average of the within-study sampling errors, v_{ij} (Tipton & Pustejovsky, 2015).

As discussed above, studies included in (Tanner-Smith & Lipsey, 2015) reported multiple correlated measures of the outcome variables. For example, alcohol consumption was measured by frequency of consumption, quantity consumed, and blood alcohol consumption. Alcohol-related problems were measured by risky sexual behavior, relationship problems and DUI/DWI convictions. Such structure of dependence is an example of correlated effects model due the presence of multiple correlated measures of the outcomes. Another example of correlated effects structure is the data analyzed by Sala et al. (2018). The authors conducted a meta-analysis to study whether video game training enhances cognitive ability. The meta analysis included results from randomized controlled trials that compared one treatment group (active video game players) to multiple comparison groups (non video game players, non-active video game players) yielding multiple effect sizes per study. The effect sizes are correlated because the comparison groups were compared to the same treatment group.

2.6.3 Hierarchical Effects

Hierarchical effects refers to cases where studies are nested within a larger group (Tanner-Smith et al., 2016). For example, studies can be clustered by the same lab, investigator or region, where the methods, protocols, and personnel used for conducting the studies can be similar. Therefore, studies conducted by the same lab or in the same region tend to have similar effect sizes. In the hierarchical effects, effect sizes (level 1) are nested within studies (level 2) which are nested within a larger group like a lab (level 3) (Tanner-Smith et al., 2016). The model assumes independent errors within cluster (Tanner-Smith et al., 2016). In the hierarchical effects model, the dependence occurs through the underlying effect size parameter, study level random effects (Hedges, 2009; Hedges et al., 2010). Under the hierarchical effects model (Konstantopoulos, 2011):

$$T_{ij} = \theta_{ij} + e_{ij} \quad \text{with} \quad e_{ij} \sim N(0, \sigma_{ij}^2) \tag{2.14}$$

The correlation between two error terms is assumed to be $\text{corr}(e_{hj}, e_{ij}) = 0$.

$$\theta_{ij} = \gamma_j + u_{ij} \quad \text{with} \quad u_{ij} \sim N(0, \omega^2) \tag{2.15}$$

$$\gamma_j = \mu + v_j \quad \text{with} \quad v_j \sim N(0, \tau^2) \tag{2.16}$$

Note that unlike the correlated effects model, the hierarchical effects model as an error term u_{ij} associated with the true effect size parameter. The error term has a normal distribution with a mean of 0 and the within-study variance of the effect sizes denoted by ω^2. Also note that the correlation between two error terms within a study is assumed to be 0.

Equations 2.14, 2.15, and 2.16 imply that (Konstantopoulos, 2011):

$$T_{ij} = \mu + v_j + u_{ij} + e_{ij} \tag{2.17}$$

The above formulation implies that the effect sizes have the following marginal variance-covariance matrix Hedges et al. (2010), Tipton and Pustejovsky (2015):

$$\boldsymbol{\Psi}_{hj} = \tau^2 \mathbf{J}_j + \omega^2 \mathbf{I}_j + \mathbf{V}_j \tag{2.18}$$

Here, $\boldsymbol{\Psi}_{hj}$ indicates the covariance matrix according to the hierarchical effects model for study j, ω^2 denotes the within-study variance in the true effect sizes, and $\mathbf{V}_j$ denotes the $k_j \times k_j$ diagonal matrix of sampling error variances for study j (Tanner-Smith et al., 2016; Tipton & Pustejovsky, 2015). For the hierarchical effects model, weights can be calculated as (Tipton & Pustejovsky, 2015):

$$\mathbf{W}_j = diag(w_{1j}, ..., w_{kj}) \tag{2.19}$$

where,

$$w_{ij} = \frac{1}{(v_{ij} + \tau^2 + \omega^2)} \tag{2.20}$$

Here, w_{ij} denotes weight for effect size i in study j (Tipton & Pustejovsky, 2015).

An example of hierarchical effects structure is reported in Thompson et al. (2017). The authors conducted a meta-analysis studying whether alcohol decreases experimentally induced pain. The authors noted that several studies in the meta-analysis were conducted by the same laboratory which may have used similar methodology to conduct the different experiments.

2.6.4 Comparison of the Two Models

Dependence in the correlated effects model occurs through the sampling error terms e_{ij} (Hedges et al., 2010). The terms are assumed to have correlation value of ρ. On the other hand, dependence in hierarchical effects model occurs through the true effect sizes θ_{ij} (Hedges et al., 2010). Hierarchical effects model assumes the errors terms to have a correlation of 0 (Konstantopoulos, 2011). Correlated and hierarchical effects can occur together (Tanner-Smith et al., 2016). For example, different studies conducted by the same lab can report multiple measures of the outcome. In such cases, Tanner-Smith and Tipton (2014) recommend choosing the model for the type of dependence that is most prevalent. If there are dependent effect sizes and the true model is hierarchical, the particular methods used to handle dependence do not matter as much. However, if the underlying model is the correlated effects model, methods to handle dependence become more important. The correlated effects model does not assume ρ to be 0. The estimation strategy for ρ becomes important. Strategies used to handle dependence are discussed in the following section.

2.7 Methods to Handle Dependence

2.7.1 Ad-Hoc Methods

One strategy to handle dependence is to ignore it, assume the effect sizes are independent, and proceed with running meta-regression models (Hedges et al., 2010). Hedges et al. (2010) note that this procedure might perform well if the number of studies with dependent effect sizes is small. However, generally that may not be the case and this approach will lead to incorrect standard errors and hypothesis tests.

Ad-hoc methods for handling dependence include selecting one effect size per study either randomly or based on some criteria (Tanner-Smith et al., 2016). Effect sizes can also be averaged within a study (Tanner-Smith et al., 2016). These methods can yield independent effect sizes but can be problematic when there is within study variation in effect sizes. Deleting or averaging effect sizes can cause loss of potentially necessary information (Tanner-Smith et al., 2016).

Another ad-hoc procedure is the shifting the unit-of-analysis approach, which involves creating subsets of effect sizes (Cooper, 1998). For example, the researcher can separate effect size by each outcome and then conduct univariate meta-analysis per outcome. The shifting the unit-of-analysis prevents loss of information (Scammacca et al., 2014). However, it requires multiple meta-analytic models for different outcomes (Scammacca et al., 2014). This approach can result in low power due to multiplicity, or because some outcomes may lower the number of studies available. It also ignores the correlational structures between the effect sizes. Moreover, this strategy is not useful if a researcher wants to study effect across the subsets or study differential effects. For example, the researcher may want to calculate average effect size across all the outcomes and examine whether the magnitude of the effect is bigger for certain outcomes (Hedges et al., 2010).

2.7.2 Multivariate Methods

The ideal solution to handle dependence is to run multivariate meta-analysis. The estimation of multivariate model follow Equations 2.5, 2.6 and 2.7. When there are multiple effect sizes per study, the $\mathbf{\Psi}_j$ matrix in Equation 2.5 becomes a covariance matrix for study j containing variances of the effect sizes on the diagonal and the covariances between the effect sizes on the off-diagonals (Raudenbush et al., 1988). If each study j only has one effect size, the result of the multivariate analysis is the same as a univariate inverse variance weighted linear regression analysis (Raudenbush et al., 1988). Multivariate methods require the knowledge of the exact structure of $\mathbf{\Psi}_j$, which is difficult to derive from information provided in primary studies (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015). Therefore, although the multivariate method is the most ideal solution, it is oftentimes infeasible to use in practice.

2.7.3 Robust Variance Estimation

Hedges et al. (2010) proposed another procedure to handle dependence, robust variance estimation (RVE), which does not require knowledge of the covariance structure between the effect sizes like multivariate analyses, but instead estimates the variances of the meta-regression coefficients with sandwich estimators using observed residuals (Hedges et al., 2010; Tipton, 2015). RVE does require that the clusters or studies themselves are independent of each other (Hedges et al., 2010).

The RVE estimator of the variance of $\mathbf{b}$ is:

$$\mathbf{V^R} = \mathbf{M}\left[\sum_{j=1}^{m} \mathbf{X}_j^{'} \mathbf{W}_j \mathbf{A}_j \mathbf{e}_j \mathbf{e}_j^{'} \mathbf{A}_j \mathbf{W}_j \mathbf{X}_j\right]\mathbf{M} \tag{2.21}$$

Here, $\mathbf{e}_j$ is the vector of residuals for study j, where $\mathbf{e}_j = \mathbf{T}_j - \mathbf{X}_j\mathbf{b}$. Furthermore, $\mathbf{A}_j$ denotes a $k_j \times k_j$ adjustment matrix (Tipton, 2015). In the formulation of RVE initially proposed by Hedges et al. (2010), $\mathbf{A}_j$ is an identity matrix (Hedges et al., 2010; Tipton & Pustejovsky, 2015). In RVE, the true covariance matrix, $\mathbf{\Psi}_j$, is estimated by $\mathbf{A}_j \mathbf{e}_j \mathbf{e}_j^{'} \mathbf{A}_j$. Although $\mathbf{A}_j \mathbf{e}_j \mathbf{e}_j^{'} \mathbf{A}_j$ is a poor estimate of $\mathbf{\Psi}_j$, because of the weak law of large numbers, as the number of studies increases $\mathbf{V^R}$ converges to Var($\mathbf{b}$) (Hedges et al., 2010; Tipton, 2015):

$$\mathrm{E}[\mathbf{A}_j \mathbf{e}_j \mathbf{e}_j^{'} \mathbf{A}_j] = \mathbf{\Psi}_j \tag{2.22}$$

Therefore, the performance characteristics of RVE are asymptotic in that it requires a large number of clusters or studies to provide accurate standard errors (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015). Simulation studies have shown that if the number of studies is small, RVE can result in downwardly biased standard errors and inflation of Type 1 error rates (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015).

To try to mitigate the effects of the asymptotic characteristic of RVE, Hedges et al. (2010) suggested an adjustment to calculate variances of the regression coefficients:

$$\mathbf{A}_j = \sqrt{m/(m-p)}\mathbf{I}_j \tag{2.23}$$

where m is the number of studies, p is the number of coefficients to be estimated, and $\mathbf{I}_j$ is a $k_j \times k_j$ identity matrix (Hedges et al., 2010; Tipton & Pustejovsky, 2015). The t-statistic is compared to a distribution with $m-p$ degrees of freedom (Hedges et al., 2010; Tipton & Pustejovsky, 2015). However, simulation studies have shown that even with the adjustment, Type 1 error can be inflated when the number of studies is less than 40 (Hedges et al., 2010; Tipton, 2015). Tipton (2015) and Tipton and Pustejovsky (2015) noted that over half of the meta-analytic studies in education and social sciences contain fewer than 40 studies. Due to the prevalence of small number of studies and of dependent effect sizes in applied meta-analyses, it is important to examine small sample corrections for RVE that adequately control Type 1 error inflation.

2.8 Small Sample Corrections for Tests of Single Coefficients

Tipton (2015) proposed and evaluated several methods to improve small-sample performance of RVE for single coefficient meta-regression t-tests. The null hypothesis for the test of a single meta-regression coefficient is $\mathrm{H}_0 : \beta_s = 0$. The Wald test statistics is formulated as (Tipton & Pustejovsky, 2015):

$$t_s = \frac{b_s}{\sqrt{V_{ss}^{RVE}}} \tag{2.24}$$

Here b_s denotes s^{th} item in $\mathbf{b}$, and V_{ss}^{R} denotes the s^{th} diagonal element of $\mathbf{V^R}$ The t_s statistics is compared to a t-distribution (Tipton & Pustejovsky, 2015).

The methods evaluated by Tipton (2015) included the one proposed by Hedges et al. (2010) as described above (H), a jack-knife estimator adjustment (JK), and a bias reduced linearization method (MBB) proposed by Hedges et al. (2010). All of these methods involve multiplying the robust variance matrix of the regression coefficients with an adjustment matrix. Of these, the MBB method performed best in terms of reducing Type 1 error inflation (Tipton & Pustejovsky, 2015). The adjustment matrix used by MBB is as follows (Tipton, 2015; Tipton & Pustejovsky, 2015):

$$\mathbf{A}_j = \mathbf{W}_j^{-1/2}\left[\mathbf{W}_j^{-1/2}\left(\mathbf{W}_j^{-1} - \mathbf{X}_j\mathbf{M}\mathbf{X}_j^{'}\right)\mathbf{W}_j^{-1/2}\right]^{-1/2}\mathbf{W}_j^{-1/2} \tag{2.25}$$

Here $\mathbf{W}_j^{-1/2}$ denotes the inverse of the symmetric square root of the weight matrix (Tipton & Pustejovsky, 2015). Tipton and Pustejovsky (2015) noted that when the working model is correct, accounting for the adjustment matrix when estimating $\mathbf{V^R}$ provides an exactly unbiased estimate of the variance of $\mathbf{b}$. However, Tipton (2015) showed that in many cases using an adjustment matrix like MBB by itself was not enough to reduce Type 1 error rates to acceptable levels. Tipton (2015) proposed using Satterthwaite corrections for degrees of freedom along with using the adjustment matrices. The Satterthwaite degrees of freedom are calculated as follows:

$$v_s = \frac{2\mathrm{E}(V_{ss}^R)^2}{\mathrm{Var}(V_{ss}^R)} \tag{2.26}$$

Tipton (2015) provided formulas to calculate the expected value and the variance of V_{ss}^R as:

$$\mathrm{E}(V_{ss}^R) = \sum_{j=1}^{m} \boldsymbol{\lambda}_{jk} \tag{2.27}$$

$$\mathrm{Var}(V_{ss}^R) = \sum_{j=1}^{m} \boldsymbol{\lambda}_{jk}^2 \tag{2.28}$$

(James I need help understand what is going on here).

Simulation results showed that all methods performed well when combined with Satterthwaite corrections when the degrees of freedom are greater than or equal to four (Tipton, 2015). If the degrees of freedom are less than four, Type 1 error inflation can occur (Tanner-Smith et al., 2016; Tipton, 2015). MBB combined with Satterthwaite (MBBS) and JK with Satterthwaite (JKS) adjustment showed better Type 1 error rate control. MBBS provided Type 1 error rates closer to the nominal value (.05) with a maximum rate across simulation conditions of .06, and thus was the recommended method.

Tipton (2015) noted that small sample size was not the only factor that affected Type 1 error rates while using RVE with the corrections. Certain features of the covariates could influence the degrees of freedoms even with the sample size held constant. Categorical moderators with unbalanced sample sizes for each of the groups and continuous moderators with outliers may have undue influence on the degrees of freedom.

2.9 Small Sample Corrections for Multiple Contrast Hypothesis

Tipton and Pustejovsky (2015) extended methods developed in Tipton (2015) to evaluate small sample correction methods for F tests of multiple contrast hypothesis (e.g., comparison of nested models, moderating effect of a categorical variable with multiple levels). For example, Tanner-Smith and Lipsey (2015) included studies with multiple measures of the outcomes. The authors may want to examine whether the coefficients for all of the different measures of the outcomes are the same.

Consider the null hypothesis $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{c}$. Here $\mathbf{C}$ is a $q \times p$ contrast matrix and $\mathbf{c}$ is a $q \times 1$ vector. For example, applied meta-analysts might want to examine the equality of regression coefficients $\mathrm{H}_0 : \beta_1 = \beta_2 = ... = \beta_{p-1}$ with $\mathbf{c} = 0$. To evaluate the null hypotheses, a Q statistic can be calculated as follows (Tipton, 2015):

$$Q = (\mathbf{Cb} - \mathbf{c})^{'} \left(\mathbf{C}\mathbf{V}^{\mathbf{R}}\mathbf{C}^{'}\right)^{-1} (\mathbf{Cb} - \mathbf{c}) \tag{2.29}$$

The Q statistic follows a χ^2 distribution with q degrees of freedom when the number of studies is adequately large (Tipton & Pustejovsky, 2015).

For multiple contrast hypothesis tests, Tipton and Pustejovsky (2015) suggested a degrees of freedom correction similar to the correction suggested by Hedges et al. (2010). The test statistic would be $F = Q/q$, which would be compared to an F distribution with q and $m - p$ degrees of freedom. Tipton and Pustejovsky (2015) called this test the naive F-test and noted that this test performs well only under specific conditions. The authors contended that the test performs poorly because it uses the same degrees of freedom without any regards to the contrasts being used and does not account for the design matrix.

In the development of small sample corrections for multiple contrast hypothesis tests, Tipton and Pustejovsky (2015) used the same adjustment matrix suggested by Tipton (2015), given in Equation 2.25. The matrices are derived under a working covariance model. Tipton and Pustejovsky (2015) implemented several strategies to approximate the sampling distribution of the random matrix $\left(\mathbf{C}\mathbf{V}^{\mathbf{R}}\mathbf{C}^{'}\right)$. Tipton and Pustejovsky (2015) reviewed literature on analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), multiple regression without considering heteroskedasticity, and generalized estimating equations. In each of these areas, small sample corrections for RVE had been examined.

The first strategy proposed by Tipton and Pustejovsky (2015) approximates the sampling distribution using a Wishart distribution leading to a Hotelling's T^2 based test statistics, which is a multiple of F test statistics (Tipton & Pustejovsky, 2015). The second strategy uses a spectral decomposition of the random matrix, estimating Q as a sum of independent univariate random variables (Tipton & Pustejovsky, 2015). For the first strategy, Tipton and Pustejovsky (2015) examined three approaches differing in the degrees of freedom. For the second strategy, Tipton and Pustejovsky

(2015) examined two different approaches. One uses Satterthwaite degrees of freedom, while the other involves transforming the univariate random variables so the sum follows a χ^2 distribution (Tipton & Pustejovsky, 2015). Tipton and Pustejovsky (2015) examined a total of five methods. The simulation results showed that the methods based on eigen-decomposition performed poorly with Type 1 error rates exceeding the nominal error rate. Therefore, the eigen-decomposition approach will not be discussed further here. The approaches based on Hotelling's T^2 distribution performed better. The methods are discussed below.

Let $\boldsymbol{\Omega}$ denote the true variance of $(\mathbf{Cb} - \mathbf{c})$, where $\boldsymbol{\Omega} = \mathbf{C}\mathrm{Var}(\mathbf{b})\mathbf{C}^{'}$ [Tipton and Pustejovsky (2015)]. The Q statistics can also be written as [Tipton and Pustejovsky (2015)]:

$$Q = \mathbf{z}'\mathbf{D}^{-1}\mathbf{z} \tag{2.30}$$

where, $\mathbf{z} = \boldsymbol{\Omega}^{-1/2}(\mathbf{Cb} - \mathbf{c})$ and $\mathbf{D} = \boldsymbol{\Omega}^{-1/2}\left(\mathbf{C}\mathbf{V}^{\mathbf{R}}\mathbf{C}^{'}\right)\boldsymbol{\Omega}^{-1/2}$ [Tipton and Pustejovsky (2015)]. Under the null hypothesis, $\mathbf{z}$ follows a normal distribution with mean of 0 and covariance of $\mathbf{I}_q$. Please see Tipton and Pustejovsky (2015) for theorems describing the moments of $\mathbf{D}$.

The first approach proposed by Tipton and Pustejovsky (2015) approximates the sampling distribution of $\mathbf{D}$ using a Wishart distribution. The Q statistic follows Hotelling's T^2 distribution with degrees of freedom η:

$$\frac{\eta - q + 1}{\eta q} Q \quad \sim F(q, \eta - 1 + 1) \tag{2.31}$$

To estimate η, Tipton and Pustejovsky (2015) examined three different approaches. Of the three approaches, the one (AHZ) that performed best was proposed by Zhang (2012) and Zhang (2012). This method was originally developed for heteroskedastic one way ANOVA and MANOVA. The approach matches the total variance in the random matrix to the total variation in a Wishart distribution. The degrees of freedom are estimated as (Zhang, 2012):

$$\eta_z = \frac{q(q+1)}{\sum_{s=1}^{q}\sum_{t=1}^{q}\mathrm{Var}\left(d_{st}\right)} \tag{2.32}$$

Here d_{st} denotes entry in row s and column t of $\mathbf{D}$.

The simulation results showed that the naive F test performed poorly even when the number of studies equaled 100. Methods using Hotelling's T^2 distribution adequately reduced Type 1 error inflation. The AHZ method resulted in Type 1 error rates closest to the nominal rate of .05. The Type 1 error rate of this method ranged from 0.00 to 0.04 across simulation conditions with a median of 0.0254. The other two Hotelling's based methods were very conservative with Type 1 error rates near zero. AHZ test can be used for tests of single coefficients as well as multi-contrast hypotheses. For single coefficient tests, AHZ is identical to the MBBS test recommended by Tipton (2015). Though the AHZ method recommended by Tipton and Pustejovsky (2015) controls Type 1 error rate, AHZ had below nominal Type 1 error rates across conditions, indicating that the method may possibly be conservative.

2.10 Cluster Wild Bootstrapping

An alternative method that may account for dependence and may work under small sample size is cluster wild bootstrapping. Cluster wild bootstrapping has been investigated to correct clustered heteroskedastic error terms of regular regression parameter estimates (Cameron et al., 2008). However, as noted by Tipton and Pustejovsky (2015), it has not been studied under the meta-analytic framework.

2.10.1 Bootstrapping

Bootstrapping is a purely computational technique that involves generating iterations of pseudo-samples from the original sample (Cameron et al., 2008). In this way, it allows the creation of empirical distribution of some estimate of interest, $\hat{\theta}$, for some parameter of interest, θ. The general procedure involves re-sampling with replacement from the original data-set and estimating $\hat{\theta}$ for each of the re-samples. Then inferences can be made about the distribution of the estimate of interest from the distribution of the bootstrap replicates (Cameron et al., 2008). For example, the bootstrap standard error can then be estimated by taking the standard deviation of all the estimated $\hat{\theta}$s. Confidence intervals can be estimated and hypothesis tests can be conducted based on the bootstrap standard error. Bootstrapping techniques are non-parametric or semi-parametric and thus, involve no distributional assumptions. The re-sampling process emulates underlying sampling distribution of the estimator without having to know the true parameters to generate data via simulation. Furthermore, because bootstrapping is a purely computational technique, it can be applied to any statistical method, without regard to complicated mathematical derivations.

Cameron et al. (2008) noted that there are several aspects to consider when bootstrapping: (1) which units to sample: individual cases or clusters; (2) what to sample: actual outcome and covariance matrices or residuals; (3) what statistics to calculate on each iteration; (4) how to derive error estimates from the bootstrap results; and, (5) whether to impose null hypothesis when generating replicates.

2.10.2 Cluster Bootstrapping

The common correction to handle dependence due to clustering in econometrics literature is cluster robust variance estimation (CRVE), which is the basis of RVE in meta-analysis; CRVE has asymptotic properties (Cameron et al., 2008).

As alternative to CRVE, several bootstrapping methods have been studied in the econometrics literature: pair bootstrapping, residual bootstrapping and wild bootstrapping (Cameron et al., 2008).

Pair Bootstrapping

Pair bootstrapping involves re-sampling clusters with replacement from the original data-set. The clusters contain the pair of $\mathbf{X}_j$ and $\mathbf{T}_j$ (Cameron et al., 2008). The estimate of interest is then calculated on each bootstrap replicate. The standard error is calculated by taking the standard deviation of the estimates across the replicates (Cameron et al., 2008). This approach provides only a rough approximate of the standard error of the estimate because it involves re-sampling $\mathbf{X}_j$ instead of holding it constant. The standards errors are not conditional on the empirical distribution of the covariates. If the number of clusters is small, this procedure does not control Type 1 error rate inflation (Cameron et al., 2008). In addition, when there are small number of clusters, there may be situations where certain covariates may not have any variance and therefore the estimation of the regression coefficient and standard errors may be infeasible (Cameron et al., 2008). This problem occurs because pair bootstrapping involves re-sampling the covariates (Cameron et al., 2008).

Residual Bootstrapping

Residual bootstrapping involves re-sampling residuals while holding $\mathbf{X}_j$ constant (Cameron et al., 2008). The residuals are then used to calculate new outcome values $\mathbf{T}_j^*$ (Cameron et al., 2008). The pseudo-samples ($\mathbf{T}_1^*$, $\mathbf{X}_1, ...,$ $\mathbf{T}_j^*$, $\mathbf{X}_j$) are re-sampled (Cameron et al., 2008). This method involves strong assumptions. The first assumption is that $E[\mathbf{T}_j^*|\mathbf{X}_j] = \mathbf{X}_j\boldsymbol{\beta}_j$. Another assumption underlying this method is that the errors are independently and identically distributed and hence, homoskedastic (Cameron et al., 2008). Additionally, the method assumes that all clusters have the same sample size (Cameron et al., 2008). Residual bootstrapping does not have the problem of having no variance for certain covariates like pair bootstrapping as it does not involve re-sampling covariates (Cameron et al., 2008).

Cluster Wild Bootstrapping

Cluster wild boostrapping method conditions on transformed residuals Cluster wild bootstrapping has been examined particularly when the number of clusters is small (Cameron et al., 2008; MacKinnon & Webb, 2018). The method does not assume that the regression error vectors are identically and independently distributed or that the clusters are all of equal size, and additionally makes no assumptions about asymptotic number of clusters (Cameron et al., 2008). Cluster wild bootstrapping involves re-sampling residuals (Cameron et al., 2008).

The general process of conducting a wild bootstrap without clusters is as follows: (1) estimating a regression model based on the null hypothesis; (2) generating random pulls of the values -1 and 1 each with a 0.50 probability—although multiple values and probabilities can be used in this process, the use of the values 1 and -1 and the probabilities of 0.50 are common and are referred to as Rademacher weights (Cameron et al., 2008); (3) multiplying this random variable by the residual terms from the original model; (4) then obtaining new predicted outcome scores based on these altered residual terms; and, finally (5) re-estimating the original model using the new calculated outcome scores (Cameron et al., 2008). This process is done over R bootstrap replications with the regression coefficients from each replication being the bootstrap replicate. When clusters are involved, the process is similar except the random generated values (i.e., -1 and 1) are constant within each cluster. The bootstrap p-value is calculated by comparing estimates from the bootstrap replicates to the estimate from the original data.

For each cluster the random Rademacher weights are the same (Cameron et al., 2008; MacKinnon & Webb, 2017). MacKinnon and Webb (2017) argued because of the weights being constant for each cluster, bootstrap based inferences preserve the within-cluster variance and covariances of the error terms, to the extent that the residuals preserve them. The method of bootstrapping described above uses estimates generated under the null hypothesis to calculate the residuals (Cameron et al., 2008; MacKinnon & Webb, 2017). MacKinnon and Webb (2017) noted that residuals generated under the null hypothesis are called restricted residuals and argued that it is possible to calculate either restricted or unrestricted residuals. Using restricted residuals is advantageous because the bootstrapping can generate the distribution of a statistic under the null hypothesis. Additionally, when the data contains small number of clusters (studies), the estimated errors tends to have a small downward bias and the magnitude of the bias depends on the number of parameters in the model. Imposing a null hypothesis circumvents this issue as there are only a few parameters to estimate under the null.

Cameron et al. (2008) conducted several simulation studies to examine the finite sample properties of CRVE compared to the bootstrap techniques. The authors generated data from a linear model with a single covariate and varied the number of clusters from 5 to 30 in increments of 5. The errors were generated to be homoskedastic and heteroskedastic. Simulation results showed that even with a small number of clusters, the cluster wild bootstrap method had closest to the nominal Type I error rate compared to cluster RVE methods and other bootstrapping methods for single coefficient t-tests.

Additionally, MacKinnon and Webb (2017) found that cluster wild bootstrap performed well even cluster sizes are wildly unequal.

Derivation of Cluster Wild Bootstrapping

Below is the discussion of the derivation of cluster wild bootstrap p-values. Consider the model:

$$\mathbf{T}_j = \mathbf{U}_j\boldsymbol{\alpha} + \mathbf{X}_j\boldsymbol{\beta} + \mathbf{e}_j, \quad \text{where} \quad \mathbf{e}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.33}$$

for $j = 1, ..., m$, and where $\mathbf{e}_j$ is independent of $\mathbf{e}_k$ for $j \neq k$. The model will be estimated using weighted least squares with fixed weight matrices $\mathbf{W}_1, ..., \mathbf{W}_m$.

Suppose the hypothesis to evaluate is $H_0 : \boldsymbol{\beta} = \mathbf{0}$. The $\boldsymbol{\beta}$ vector contains the coefficients evaluated in the contrast hypothesis. And, the $\boldsymbol{\alpha}$ vector contains the coefficients that are not evaluated. To calculate wild bootstrap, $\tilde{\boldsymbol{\alpha}}$ will be estimated as:

$$\tilde{\boldsymbol{\alpha}} = \mathbf{M_U U' W T} \tag{2.34}$$

where $\mathbf{M_U} = (\mathbf{U'WU})^{-1}$. The residuals under the null model are:

$$\tilde{\mathbf{e}} = \mathbf{T} - \mathbf{U}\tilde{\boldsymbol{\alpha}} = (\mathbf{I} - \mathbf{H_U})\,\mathbf{T} \tag{2.35}$$

where $\mathbf{H_U} = \mathbf{UM_U U'W}$.

To derive the weighted least squares (WLS) estimation of $\boldsymbol{\beta}$, let $\ddot{\mathbf{X}}$ be the residuals from the weighted least squares regression of $\mathbf{X}$ on $\mathbf{U}$, i.e.,

$$\ddot{\mathbf{X}} = (\mathbf{I} - \mathbf{H_U})\,\mathbf{X} \tag{2.36}$$

The estimate of $\boldsymbol{\beta}$ can be calculated by taking the WLS regression of $\tilde{\mathbf{e}}$ on $\ddot{\mathbf{X}}$:

$$\hat{\boldsymbol{\beta}} = \mathbf{M}_{\ddot{\mathbf{X}}} \ddot{\mathbf{X}}' \mathbf{W} \tilde{\mathbf{e}} \tag{2.37}$$

where $\mathbf{M}_{\ddot{\mathbf{X}}} = \left(\ddot{\mathbf{X}}'\mathbf{W}\ddot{\mathbf{X}}\right)^{-1}$. The residuals can be calculated as:

$$\hat{\mathbf{e}} = \tilde{\mathbf{e}} - \ddot{\mathbf{X}}\hat{\boldsymbol{\beta}}. \tag{2.38}$$

The robust variance-covariance matrix can then be calculated as:

$$\mathbf{V^R} = \mathbf{M}_{\ddot{\mathbf{X}}} \left(\sum_{j=1}^{m} \ddot{\mathbf{X}}_j' \mathbf{W}_j \mathbf{A}_j \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j' \mathbf{A}_j \mathbf{W}_j \ddot{\mathbf{X}}_j \right) \mathbf{M}_{\ddot{\mathbf{X}}} \tag{2.39}$$

The Wald statistics is formulated as:

$$Q = \hat{\boldsymbol{\beta}}' \mathbf{V^R}^{-1} \hat{\boldsymbol{\beta}} \tag{2.40}$$

The approximation of the null sampling distribution of Q for cluster wild bootstrapping entails repeatedly generating new data as below:

$$\mathbf{T}_j^* = \mathbf{U}_j\tilde{\boldsymbol{\alpha}} + \eta_j\mathbf{B}_j\tilde{\boldsymbol{e}}_j \tag{2.41}$$

where $\eta_1, ..., \eta_m$ are the Rademacher weights and $\mathbf{B}_1, ..., \mathbf{B}_m$ are some adjustment matrices calculated based on the null model.

The bootstrapping procedure then involves calculating $\hat{\boldsymbol{\beta}}$, $\mathbf{V}^{\mathbf{R}}$, and Q based on the bootstrapped $\mathbf{T}^*$. Let $\mathbf{E}_j' = \ddot{\mathbf{X}}_j'\mathbf{W}_j$, and $\mathbf{G}_j' = \mathbf{E}_j'\mathbf{A}_j$. The first step involves calculating:

$$\mathbf{f}_j = \mathbf{B}_j\tilde{\boldsymbol{e}}_j \tag{2.42}$$

For $r = 1, ..., R$, with R denoting the total number of bootstrap replicates, the bootstrapping involves calculating the following:

$$\mathbf{e}_j^{(r)} = \eta_j\mathbf{f}_j \tag{2.43}$$

$$\hat{\boldsymbol{\beta}}^{(r)} = \mathbf{M}_{\ddot{\mathbf{X}}}\sum_{j=1}^{m}\mathbf{E}_j\mathbf{e}_j^{(r)} \tag{2.44}$$

$$\hat{\mathbf{e}}_j^{(r)} = \mathbf{e}_j^{(r)} - \ddot{\mathbf{X}}_j\hat{\boldsymbol{\beta}}^{(r)} \tag{2.45}$$

The variance the each estimate from each bootstrap replicate can be calculated as:

$$\mathbf{V}^{(r)} = \mathbf{M}_{\ddot{\mathbf{X}}}\left(\sum_{j=1}^{m}\mathbf{G}_j'\hat{\mathbf{e}}_j^{(r)}\left(\hat{\mathbf{e}}_j^{(r)}\right)'\mathbf{G}_j\right)\mathbf{M}_{\ddot{\mathbf{X}}} \tag{2.46}$$

The Q statistic for each replicate can be calculated as:

$$Q^{(r)} = \left(\hat{\boldsymbol{\beta}}^{(r)}\right)'\left(\mathbf{V}^{(r)}\right)^{-1}\hat{\boldsymbol{\beta}}^{(r)} \tag{2.47}$$

Finally, the p-value corresponding to the null hypothesis can be calculated as:

$$p = \frac{1}{R}\sum_{r=1}^{R} I\left(Q > Q^{(r)}\right) \tag{2.48}$$

2.10.3 Cluster Wild Bootstrapping in Applied and Methodological Research

The cluster wild bootstrapping method has been used in a handful of meta-analytic studies with dependent effect sizes and small number of studies. Examples include McEwan (2015), in Review of Educational Research, examining school-based interventions on learning in developing countries; Gallet and Doucouliagos (2014), in the Annals of Tourism Research, examining income elasticity of air travel; and Oczkowski and Doucouliagos (2015), in the American Journal of Agricultural Economy, examining the relationship between price of wine and its quality. The number of studies included varied between 36 to 77. All of the articles used bootstrapping to correct for single coefficient t-tests. None used it for omnibus tests.

Cluster wild bootstrapping has been used in a few applied studies, but its performance against small sample corrections proposed by Tipton (2015) and Tipton and Pustejovsky (2015) has not been evaluated in methodological studies under the meta-analytic framework.

2.11 Statement of Purpose

Although cluster wild bootstrapping offers a promising alternative to small sample corrections proposed by Tipton (2015) and Tipton and Pustejovsky (2015), its performance has not been evaluated in any methodological studies. Thus, the goal of this study is to examine whether using cluster wild bootstrapping improves upon the performance of the AHZ test evaluated in Tipton and Pustejovsky (2015), in terms of Type 1 error rate, for F tests of single meta-regression coefficients and of multiple contrast hypotheses.

Chapter 3

Methods

3.1 Data Generation

The data generation method will follow that of Tipton and Pustejovsky (2015). The design focuses on correlated standardized mean differences (SMDs) for each study. Tipton and Pustejovsky (2015) examined designs resembling intervention studies with multiple continuous measures outcome variables. Each analysis will be comprised of m studies. A given study j will contain k_j effect sizes.

Let $\bar{\mathbf{y}}_{Tj}$ and $\bar{\mathbf{y}}_{Cj}$ be the $k_j \times 1$ vectors of sample means for the treatment and control groups, respectively. Let $\mathbf{S}_j$ be the $k_j \times k_j$ sample covariance matrix of the outcomes, pooled across the treatment and control groups. Assuming multivariate normality:

$$\bar{\mathbf{y}}_{Cj} \sim N\left(\mathbf{0}, \frac{2}{N_j}\boldsymbol{\Psi}_j\right), \qquad \bar{\mathbf{y}}_{Tj} \sim N\left(\boldsymbol{\delta}_j, \frac{2}{N_j}\boldsymbol{\Psi}_j\right), \tag{3.1}$$

and,

$$(\bar{\mathbf{y}}_{Tj} - \bar{\mathbf{y}}_{Cj}) \sim N\left(\boldsymbol{\delta}_j, \frac{4}{N_j}\boldsymbol{\Psi}_j\right). \tag{3.2}$$

The pooled sample covariance matrix follows a multiple of a Wishart distribution with $N_j - 2$ degrees of freedom and scale matrix Ψ_j:

$$(N_j - 2)\mathbf{S}_j \sim Wishart\left(N_j - 2, \Psi_j\right). \tag{3.3}$$

Thus, to simulate the denominators of the SMD estimates, we can simulate a single Wishart matrix, pull out the diagonal entries, divide by $N_j - 2$, and take the square root. In all, we draw a single $k_j \times 1$ observation from a multi-variate normal distribution and a single $k_j \times k_j$ observation from a Wishart distribution.

$$\delta_j \sim N(\mu, \tau^2), \tag{3.4}$$

$$k_j \sim 2 + Poisson(3), \tag{3.5}$$

$$N_j \sim 20 + 2 \times Poisson(10), \tag{3.6}$$

and

$$r_j \sim Beta\left(\rho\nu, (1-\rho)\nu\right), \tag{3.7}$$

where $\rho = \mathrm{E}(r_j)$ and $\nu > 0$ controls the variability of r_j across studies, with smaller ν corresponding to more variable correlations. Specifically, $\mathrm{Var}(r_j) = \rho(1-\rho)/(1+\nu)$.

3.1.1 Covariates

To examine the Type 1 error rates, we will simulate SMDs that are entirely unrelated to the covariates. To examine power, Tipton and Pustejovsky (2015) generated five covariates, two binary and three continuous (the exact covariates are available in the online supplement). The first binary covariate, $X1$, is a study level covariate with large imbalance, equaling 1 in 15 percent of the studies. The second binary covariate, $X2$, a within-study covariate, is equal to 1 in 10 percent of the effect sizes overall and 0 to 20 percent of effect sizes within a study. $X3$ is a normally distributed study level covariate, $X4$ is a normally distributed continuous covariate that varies within study, and $X5$ is a continuous, highly skewed effect-size level covariate. Tipton and Pustejovsky (2015) noted that these types of variables are common in applied meta-analyses, with the large imbalances and skewness representing worst cases. The generated data contain 200 rows, representing the design matrix for 200 effect sizes, with 10 rows per study, totaling 20 studies. Following the procedures of Tipton and Pustejovsky (2015), in conditions where there will be more than 20 studies, we will repeat the rows of the design matrix. For conditions when the number of effect sizes is less than 10, we will select the first k_j rows from the design matrix for each study.

3.2 Simulation Design

We will generally follow the design of Tipton and Pustejovsky (2015), but with fewer conditions, as the bootstrapping procedure requires more computation time. The parameters that we will examine will be ρ, the within study correlation between outcomes; I^2, the between study variance in outcomes; and, m, the number of studies. In Tipton and Pustejovsky (2015), I^2 values were used as a measure of between study variance and the values were transformed to τ^2 to generate the effect sizes.

Tipton and Pustejovsky (2015) used the correlated effects fixed weights model assuming ρ to be 1 and I^2 to be 0. They used fixed weights to reduce computational time and generated data with different values for ρ and I^2 to examine the robustness of the small sample correct methods to model-misspecification. They found that there was no relationship between Type 1 error rate and ρ. However, as I^2 values increased, Type 1 error rate tended to increase slightly. The AHZ test contained Type 1 error rate better than the other methods examined even when the I^2 values were greater than .50 indicating high model-misspecification. We will test the performance of the methods under a small range of realistic values for true I^2 and ρ. For I^2, we will use the following values: .00, .33, and .75 to cover no to high between-study variance. The ρ values will be set to .20 and .50 to indicate small and moderate within study correlation between effect sizes.

The number of independent studies, m will be set to 20, 40 and 80 to cover realistically small and moderate and high sample sizes. Tipton and Pustejovsky (2015) found the Type 1 error rate of AHZ to be lower when number of studies equaled 20 than when it equaled 40. The number of effect sizes per study will vary between studies ranging from 1 to 10 effect sizes to create imbalanced clusters and to capture the range of effect sizes within studies seen in real meta-analytic data. The per group sample size will range from 32 to 130 following Tipton and Pustejovsky (2015) to cover realistic sample sizes of applied studies. The factors in our study include 2 correlation values x 3 I^2 values x 3 values for the number of studies, totaling 18 conditions.

References

Becker, B. J. (2000). Multivariate meta-analysis, In *Handbook of applied multivariate statistics and mathematical modeling*. Elsevier.

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

Boos, D. D. Et al. (2003). Introduction to the bootstrap world. *Statistical science*, *18*(2), 168–174.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research synthesis methods*, *8*(1), 5–18.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 47.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (Vol. 2). Sage.

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

Gallet, C. A., & Doucouliagos, H. (2014). The income elasticity of air travel: A meta-analysis. *Annals of Tourism Research*, *49*, 141–155. https://doi.org/10.1016/j.annals.2014.09.006

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171.

Hedges, L. V. (2009). Statistical considerations. *The handbook of research synthesis and meta-analysis*, 37–47.

Hedges, L. V., & Cooper, H. (2009). Research synthesis as a scientific process. *The handbook of research synthesis and meta-analysis*, *1*.

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, *21*(11), 1539–1558.

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 137–159.

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

Konstantopoulos, S, & Hedges, L. V. (2019). Statistically analyzing effect sizes: Fixed- and random-effects models. *The Handbook of Research Synthesis and Meta-Analysis,* 245–279.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76.

MacKinnon, J. G., & Webb, M. D. (2017). Wild Bootstrap Inference for Wildly Different Cluster Sizes. *Journal of Applied Econometrics*, *32*(2), 233–254. https://doi.org/10.1002/jae.2508

MacKinnon, J. G., & Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, *21*(2), 114–135. https://doi.org/10.1111/ectj.12107

Mann, C. C. (1994). Can meta-analysis make policy? JSTOR.

McCaffrey, D. F., Bell, R. M., & Botts, C. H. (2001). Generalizations of biased reduced linearization, In *Proceedings of the annual meeting of the american statistical association*.

McEwan, P. J. (2015). Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Review of Educational Research*, *85*(3), 353–394. https://doi.org/10.3102/0034654314553127

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

Murray, J., Farrington, D. P., & Sekol, I. (2012). Children's antisocial behavior, mental health, drug use, and educational performance after parental incarceration: A systematic review and meta-analysis. *Psychological bulletin*, *138*(2), 175.

Oczkowski, E., & Doucouliagos, H. (2015). Wine Prices and Quality Ratings: A Meta-regression Analysis. *American Journal of Agricultural Economics*, *97*(1), 103–121. https://doi.org/10.1093/ajae/aau057

Olkin, I, & Gleser, L. (2009). Stochastically dependent effect sizes. *The handbook of research synthesis and meta-analysis*, 357–376.

Pustejovsky, J. (2020). *Clubsandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* [R package version 0.4.1]. R package version 0.4.1. https://CRAN.R-project.org/package=clubSandwich

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

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*(1), 111.

Rice, K., Higgins, J. P., & Lumley, T. (2018). A re-evaluation of fixed effect (s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(1), 205–227.

Sala, G., Tatlidil, K. S., & Gobet, F. (2018). Video game training does not enhance cognitive ability: A comprehensive meta-analytic investigation. *Psychological bulletin*, *144*(2), 111.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of educational research*, *84*(3), 328–364.

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

Swanson, H. L., Trainin, G., Necoechea, D. M., & Hammill, D. D. (2003). Rapid naming, phonological awareness, and reading: A meta-analysis of the correlation evidence. *Review of Educational Research*, *73*(4), 407–440.

Tanner-Smith, E. E., & Lipsey, M. W. (2015). Brief alcohol interventions for adolescents and young adults: A systematic review and meta-analysis. *Journal of substance abuse treatment*, *51*, 1–18.

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in stata and spss. *Research synthesis methods*, *5*(1), 13–30.

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in r. *Journal of Developmental and Life-Course Criminology*, *2*(1), 85–112.

Thompson, T., Oram, C., Correll, C. U., Tsermentseli, S., & Stubbs, B. (2017). Analgesic effects of alcohol: A systematic review and meta-analysis of controlled experimental studies in healthy participants. *The Journal of Pain*, *18*(5), 499–510.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

Tipton, E., & Pustejovsky, J. E. (2015). Small-Sample Adjustments for Tests of Moderators and Model Fit Using Robust Variance Estimation in Meta-Regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, *7*(1), 55–79.

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

Viechtbauer, W. (2007). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie/Journal of Psychology*, *215*(2), 104–121.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of statistical software*, *36*(3), 1–48.

Zhang, J.-T. (2012). An approximate hotelling t2-test for heteroscedastic one-way manova. *Open Journal of Statistics*, *2*(1), 1–11.

Zhang, J.-T. (2013). Tests of linear hypotheses in the anova under heteroscedasticity. *International Journal of Advanced Statistics and Probability*, *1*(2), 9–24.