HHS Public Access

Author manuscript

Behav Genet. Author manuscript; available in PMC 2021 March 01.

Published in final edited form as:

Behav Genet. 2020 March ; 50(2): 127–138. doi:10.1007/s10519-020-09993-9.

A Simulation Study of Bootstrap Approaches to Estimate Confidence Intervals in DeFries-Fulker Regression Models (with Application to the Heritability of BMI Changes in the NLSY)

Patrick O'Keefe[1], **Joseph Lee Rodgers**[1]

[1]Department of Psychology and Human Development, Peabody College, Vanderbilt University, Nasvhille, Tennessee.

Abstract

The univariate bootstrap is a relatively recently developed version of the bootstrap (Lee & Rodgers, 1998). DeFries-Fulker (DF) analysis is a regression model used to estimate parameters in behavioral genetic models (DeFries & Fulker, 1985). It is appealing for its simplicity; however, it violates certain regression assumptions such as homogeneity of variance and independence of errors that make calculation of standard errors and confidence intervals problematic. Methods have been developed to account for these issues (Kohler & Rodgers, 2001), however the univariate bootstrap represents a unique means of doing so that is presaged by suggestions from previous DF research (e.g., Cherny, Cardon, Fulker, & DeFries, 1992). In the present study we use simulations to examine the performance of the univariate bootstrap in the context of DF analysis. We compare a number of possible bootstrap schemes as well as more traditional confidence interval methods. We follow up with an empirical demonstration, applying results of the simulation to models estimated to investigate changes in body mass index in adults from the National Longitudinal Survey of Youth 1979 data.

Keywords

Bootstrapping; DeFries-Fulker Regression; BMI; Confidence Intervals; NLSY

Introduction

Defries-Fulker regression (DF Analysis) is a biometrical estimation routine based on standard regression models. Though the method is around 35 years old, and there exists substantial interest in and use of the method, there are still outstanding questions about best-practices in terms of implementing DF Analysis. The purpose of this project is to develop and evaluate several confidence interval methods that can be used with this biometrical procedure. Specifically, using simulation, we evaluate the application of several bootstrap

Corresponding Author: Patrick O'Keefe, Address: 1503 Clay St., Nashville TN, 37208, Phone: (615)-525-0848, Patrick.okeefe@vanderbilt.edu.

resampling approaches, including the standard and the univariate bootstrap, to DeFries-Fulker behavioral genetic models. The goal is to provide a straightforward and accurate way to get confidence intervals for DF model parameter estimates, particularly in DF models applied to non-normal data. Current options for DF model confidence interval (CI) creation are limited to CI's formed using a sandwich estimator, which may have software limitations, or the standard regression CI's, which are based on parametric statistical assumptions and likely to be inaccurate.

In contrast to current CI options, the standard bootstrap is widely available and the univariate bootstrap is relatively simple to implement. Currently there are few applications of the univariate bootstrap in the literature, despite some advantages over the standard bootstrap. Thus, this project may provide methodological innovation, as a stepping off point for the application of univariate bootstrapping to broader multiple regression and more advanced models. An empirical application of the different CI methods to DF analysis of BMI data from the National Longitudinal Survey of Youth 1979 dataset is presented to illustrate.

Despite increasing popularity of other statistical methods (e.g., structural equation models), DF analysis remains a widely used tool in behavioral genetics. A google scholar search for "DeFries Fulker Regression" identified 14,300 results in August, 2019; limiting results to the past year (August 2018 to August 2019) returned 3,040 results. A wide variety of subjects are being studied using DF analysis. Roos & Neilson (2019) used DF analysis with the Add Health data to study status achievement. Schwartz et al. (2017) used the Add Health data to examine the heritability of self-control. Maczulskij and Böckerman (2019) used DF analysis with Finnish twin data to "account for shared environmental and genetic confounders" in a study of stresses related to the labor market. Other recent DF analysis applications include Dominque et al. (2016), who found evidence for increased heritability of smoking behavior over time; Jackson (2015), who examined the link between nutritional quality and antisocial behavior; Meldrum and Barnes (2017), who found that unstructured socializing with peers positively predicted delinquent behavior after controlling for genetic and environmental influences; and York (2017), who found that social media had a heritable component. This broad array of examples only scratches the surface. Since its original development, DF analysis has become increasingly valuable to estimate biometrical models. SEM methods obviously have greater flexibility, but DF analysis has broad accessibility. Many DF analysis studies report both SEM and DF analysis results (an early example can be found in Rodgers, Kohler, Kyvik, & Christensen, 2001, who studied fertility using DF Analysis in the Danish twin data).

DF Analysis

DeFries-Fulker (DF) analysis is a regression method to estimate biometrical parameters from behavioral genetic/kinship data (DeFries & Fulker, 1985). The typical goal in any biometric ACE analysis is to partition the total phenotypic variance of a given outcome into the proportion that is genetic variance (A, or h^2), the proportion that is shared environmental variance (C, or c^2), and the proportion that is nonshared environmental variance (E, or e^2). The DF model does this using a regression model that is easy to use and that can allow for

the inclusion of additional explanatory variables (e.g., to further partition environmental variance into known and unknown environmental factors; see Rodgers, Rowe, & Li, 1994). Dominance models can be estimated using DF analysis (Waller, 1994), and the method is useful in both selected (DeFries and Fulker, 1985; Purcell & Sham, 2003) and unselected (Rodgers & McGue, 1994) settings.

In their original formulation DeFries and Fulker (1985) intended their model to be used in cases where one member of a kinship pair had a selected condition (e.g., a reading disability, or schizophrenia). This member of the kinship pair would be the focal member, or proband. The proband's score on the outcome variable would be the IV in the regression model, call that score K_2. Using the score of the other kin pair member (call that score K_1), and the known average genetic relatedness of the kin pair (R; 1 for monozygotic twins, .5 for full siblings and dizygotic twins, etc.) DeFries and Fulker's regression model can estimate how the overall phenotypic variance partitions into a genetic component of variance, a shared environment component of variance, and a non-shared environment and error component (the error component and the non-shared environmental component are combined or fully confounded in the residual term in many DF models, in particular in those with two levels of genetic relatedness such as MZ and DZ twins). Table I shows an example of the kind of data that might be used for a DF model. This data set has four MZ twin pairs (R = 1) and two full sibling or dizygotic twin pairs (R = .5). The K_2 scores correspond to the proband for the regression analysis.

The original formulation of the DF model follows:

$$K_1 = B_0 + B_1 K_2 + B_2 R + B_3 K_2 R + e$$

In this formulation K_2 is the proband outcome score, K_1 is the co-kin outcome score, and R is the proportion of (segregating) genes shared on average. The coefficients can be directly interpreted in behavioral genetic terms (see Rodgers & McGue, 1994): B_1 is the proportion of variance caused by shared environmental effects (genetic effects have been controlled by the other regression terms, and non-shared environmental effects lead to differences, not similarities, and go into the residual term). B_3 is a direct estimate of the proportion of variability associated with genetic processes. Rodgers and Kohler (2005) pointed out that the model contains a second (hidden) estimate of these two biometrical parameters, because $E(B_0) = (1 - c^2) * \bar{K}$ and $E(B_2) = -h^2 * \bar{K}$ (each of which can be used to easily solve for a second estimate of h^2 and c^2). Note that the two estimates of h^2 and of c^2 are not in general equal to one another. Although the original DF model is simple to implement, it has two notable shortcomings. First, when kinship pairs are not selected such that one member is clearly the proband and the other is not, the decision about which member provides the IV and which provides the DV is arbitrary. Second, the ability to estimate h^2 and c^2 in more than one way causes some ambiguity about which estimate to use.

In order to address the first shortcoming of the DF model, i.e., arbitrariness in unselected settings of which score is the IV, double entry of data was introduced. Double entry of the data allows each member of a kinship pair to take a turn as both the IV and the DV, resolving

this issue (e.g., Kohler & Rodgers, 2001; Rodgers & McGue, 1994). Double entry also results in the kinship correlations being equivalent to the intraclass correlation coefficient. Table II shows the data from Table I in double entered form, which necessarily doubles the sample size.

Finally, the DF model has been simplified to provide an equivalent, but easier to interpret, model (Rodgers & Kohler, 2005). The simplified model, which will be used in this project, resolves the ambiguity of which estimate to use for h^2 and c^2 by providing a single estimate of each. The model follows.

$$(K_1 - K_m) = b_1(K_2 - K_m) + b_2(R * (K_2 - K_m)) + e$$

In the simplified DF model, K_m is the mean of K_1 and K_2 (it is identical in double entry settings), b_1 estimates the proportion of variation attributable to shared environment factors (when model assumptions are met), b_2 estimates the proportion of variation due to shared genetic factors (i.e., heritability), and e is the residual of the model; the proportion of variation due to non-shared environment can be estimated from the identity $h^2 + c^2 + e^2 = 1.0$. The model has no intercept because both sides of the equation are mean centered using K_m, which ensures that the intercept of the model is 0.

The DF model is conceptually similar to an analysis of covariance model. The outcome/predictor variable, K can (and arguably should) be a quantitative variable. R is theoretically quantitative as a ratio scale variable. In using kinship pairs, we assume an outcome variable has been measured for both members of a kinship pair. The level of genetic relatedness needs to be known, and there need to be at least two kinship groups (e.g., monozygotic twins and dizygotic twins; full siblings and half siblings) to identify estimation of both c^2 and h^2.

It makes statistical and logical sense to double enter the data (Kohler & Rodgers, 2001; Rodgers & McGue, 1994). Without double entry the decision about which member of the kin pair is the predictor and which is the predicted is entirely arbitrary in unselected settings. Furthermore, without double entry the centering of the variables and fitting of a no-intercept model is a questionable practice when using unselected samples given that the "proband" group will in general have a different mean from the co-kin group (although because there is not a true "proband" in unselected cases this difference would be due to random chance). However, double entry artificially doubles the sample size, meaning standard errors that are produced by typical regression output are too small (Kohler & Rodgers, 2001). A sandwich estimator approach has been proposed to correct this deflation (Kohler & Rodgers); however some have suggested that the sandwich estimator may not be entirely appropriate in this case because the sandwich estimator is for model misspecification, not incorrect sample sizes. Other authors have suggested a permutation technique for estimating the standard errors although they did not actually utilize that method (Cherny et al., 1992). Interestingly, both the Kohler and Rodgers (2001) paper and the Cherny et al. (1992) paper foreshadow the current study. Kohler and Rodgers (2001) used bootstrapping as a test for their estimator, and Cherny et al.'s (1992) suggestion of permutation is immediately relevant to univariate bootstrapping.

In addition to the issue of double entry, heteroscedasticity across the groups may also be a concern (Kohler & Rodgers, 2001). If there is a genetic effect it implies that more genetically related individuals will be more similar to each other (presuming that the equal environments assumption holds). Although the effects of genetics and the environment will be equal across groups (e.g., we are not estimating different environmental effects for half-siblings vs. full siblings), our ability to accurately predict an individual's score based on their co-kin score will increase as their genetic relatedness increases. This increase in predictive ability will decrease the residual term for genetically more highly related groups, particularly as genetic effects increase. This increase has the effect of guaranteeing heteroscedasticity in the model, an obvious violation of regression assumptions involved in hypothesis testing. Resampling procedures can help us account for such violations, although results from bootstrapping techniques can vary depending on how heteroscedasticity is managed (e.g., Stine, 1989; Wu 1986).

At a minimum, a well conducted DF analysis has violated two of the fundamental assumptions of regression. The errors will not be independent because of double entry, and the errors will be heteroskedastic (assuming that any genetic heritability is present). Furthermore, double entry results in a doubling of the n term in any equations used, which will result in overly narrow confidence intervals. Determining what methods (if any) are appropriate for correcting for these regression violations is the purpose of this project.

Bootstrap Resampling

Bootstrapping is a resampling procedure for obtaining accurate standard errors and confidence intervals for model parameters (e.g., Efron & Tibshirani, 1986). The bootstrap procedure can also be used to create a sampling distribution to support standard hypothesis testing. The basic bootstrap takes a data set and samples observations of that data set with replacement to create another resampled data set. In the resampling taxonomy of Rodgers (1999), bootstrapping is sampling with replacement to form a full data set. A model is then fit to the new resampled data set and the model parameters are recorded. This process is repeated thousands of times, each time with a new resampled data set. Other versions of the bootstrap exist, but each is based on the core idea of resampling with replacement from some given set of observations or distributions to create new resample data sets to refit the model being tested. If the original data being bootstrapped are approximately representative of the population, bootstrapping provides a way to approximate the results researchers would get if they replicated their study thousands of times in the population. This approach allows the researcher to create a confidence interval around the observed parameter estimates. To calculate the number of possible permutations we need to use the "multichoose" formula, which gives the number of possible combinations when k elements are chosen from n options with replacement. The formula is $\binom{n+k-1}{k}$. In bootstrapping k and n are equal and so for a given sample there are $\binom{2n-1}{n}$ possible bootstrap samples (where n is the number of unique observations), from which to calculate the parameter estimate of interest. Several sources will use $\binom{2n-1}{n-1}$ as the formula instead. While this looks

as though it ought to produce a different result, in fact these two formulas are algebraically exactly equal. Because 2n-1 is always an odd number, and because n and n-1 are the numbers just greater than and just less than exactly half of 2n-1, the number of possible choices is equal regardless of whether n or n-1 observations are chosen. Alternatively, consider that choosing n-1 observations to keep from 2n-1 possibilities is equivalent to choosing n observations to omit (since n-1+n=2n-1), so the number of combinations must be equal. Skeptical readers may also ascertain this for themselves using any n in widely available software that can calculate the binomial coefficient (e.g., the "choose()" function in R). This formula results in a large number of possible combinations, even at small samples. For example, for ten observations there are 92,378 unique samples.

Permutation Resampling

The permutation resampling procedure is similar to bootstrapping in that it takes an original sample and creates thousands of new data sets, estimates parameters in each one, and creates a distribution of parameter estimates. In a permutation resampling procedure (also known as a randomization test; see Edgington, 1987) the researcher permutes the data thousands of different ways (potentially all possible ways if the number of possible permutations is low enough) by resampling observations without replacement from individual variables of the data set. In essence each variable is shuffled like a deck of cards, randomizing the relationship between all the variables (hence a randomization test). In the taxonomy of Rodgers (1999) the permutation (randomization) test is resampling without replacement to form a full sample. This resampling framework allows researchers to create intervals around the null hypothesis of no relationship to use for null hypothesis significance testing. For a given sample there are $n!^{k-1}$ possible combinations (where k is the number of variables). Each variable can be reordered $n!$ different ways. The number of combinations of reordered variables increases exponentially when new variables are added (although some of these combinations are simply unordered duplicates of others). There are $n!^{k}$, where k is the number of variables, combinations of permuted variables. To account for duplicates the number of combinations should be divided by $n!$, which is equivalent to $n!^{k-1}$. For example, for ten observations with two variables there are 3,628,800 unique samples, which is equal to $10!^{2-1} = 10!$.

Univariate bootstrapping

The univariate bootstrap resamples with replacement, like the traditional bootstrap, but from each variable independently as in permutation analysis (Lee & Rodgers, 1998). This procedure gives a distribution of parameter estimates under the null hypothesis, as in the permutation analysis. Alternatively, the univariate bootstrap dataset can have a correlation imposed on it using a diagonalization technique (Beasley et al., 2007). The correlation imposed can either be a hypothesis imposed (HI) or observed imposed (OI) null hypothesis about the correlation. When a correlation is imposed the resulting bootstrap provides a distribution of parameter estimates that would occur if the imposed correlation were the population correlation (Rodgers & Beasley, 2012). When the observed correlations are imposed on the data set this will result in a confidence interval around the observed parameter estimates, as in standard bootstrapping. The number of unique samples in a

univariate bootstrap is $\binom{n^k + n - 1}{n}$ unique samples (as before, n is sample size and k is the number of measured variables). For example, for ten observations and two variables there are 42.6 trillion possible unique data sets.

The diagonalization technique used in past research on the univariate bootstrap was first developed by Kaiser and Dickman (1962) as an approach to generate data with a specified correlation. In the original Kaiser-Dickman method the data to be diagonalized are standardized (and are assumed to be uncorrelated in the population), and then matrix multiplied by a matrix square-root decomposition of the desired correlation matrix; the Cholesky decomposition has performed best in univariate bootstrapping applications. For the univariate bootstrap the Kaiser-Dickman procedure needs to be slightly altered. Because the goal is a sampling frame with a given correlation structure, the sampling frame needs to be standardized, not the raw data. If the raw data were standardized it would involve dividing the data by the sample standard deviation, however the sampling frame standard deviation of that variable is based on many repetitions of that variable (n^{k-1}, where k is the number of variables and n is the number of observations). This means that the sampling frame estimate of the variance is equal to $\frac{n^{(k-1)}\sum(x-\bar{x})^2}{n^k - 1}$, instead of the original sample estimate of the variance, $\frac{\sum(x-\bar{x})^2}{n-1}$. If the correct variance is used in the standardization, the normal Kaiser-Dickman procedure can then be followed and will result in a sampling frame with the correct correlation structure. If the original sample estimate of the variance is used, the correlation structure will not match the desired correlation structure. The discrepancy between the two variances will go to 0 as n gets large, because both equations will, in the limit, be equivalent to the sum of squares over n. For relatively small n cases, the difference can be quite important.

Bias Correction, Acceleration and Invalid Bootstraps

Although the typical bootstrap is conceptually simple, in practice some bias is present in where the interval is centered; the estimate needs to be corrected for this bias, because the interval may not be wide enough. Bias corrected and accelerated intervals (BCa) were created to help manage bias and width issues in standard bootstrapping (Efron, 1982). In contrast, the univariate bootstrap has generally low bias, both in its null and HI and OI forms, although the OI form performs somewhat better than the HI form with regards to alpha control in testing hypotheses about the correlations (Beasley et al., 2007). In addition, a bootstrap can (with low probability) return a data set that has a single constant resampled value for one of the variables (because the samples are with replacement the same observation could be selected every time). In the typical bootstrap that will happen with probability $\frac{n}{\binom{2n-1}{n}}$. For ten unique observations that problematic outcome will occur about .01% of the time. In the univariate bootstrap it occurs with probability $\frac{\binom{2n-1}{n} \times k}{\binom{n^k + n - 1}{n}}$. For

ten observations with two variables that occurs approximately .00000004% of the time. Although these probabilities are low, when sampling from a data set with ten bivariate observations, for 10,000 bootstrap samples there is a 66.13% chance of at least one invalid sample in the typical bootstrap procedure; with the univariate bootstrap there is only a .0004% chance of an invalid sample.

The univariate bootstrap has some limitations (as currently implemented). The univariate bootstrap eliminates heteroscedasticity in regression residuals entirely, similar to residual bootstrapping (Stine, 1989). Heteroscedasticity occurs when the variance of the residuals changes across levels of the independent variable. Some authors have suggested that failure to use a bootstrapping method that replicates heteroscedasticity can result in an unrepresentative bootstrap parameter distribution and potential bias (e.g., Stine, 1989; Wu, 1986). The univariate bootstrap creates a grid of points that is uniformly variable across the whole length of every axis, and as a result there is no heteroscedasticity in the base univariate sampling frame. Diagonalization reintroduces linear relationships, but it does not reintroduce heteroscedasticity. In addition to heteroscedasticity, if higher order relationships are of substantive importance the univariate bootstrap is problematic. Finally, the univariate bootstrap has not been adequately extended beyond bivariate correlations (but see Rodgers & Beasley, 2012, for an introductory effort at using the univariate bootstrap for multiple regression). Although these are weaknesses that need to be addressed, they are not the focus of the present study.

Current application

The DF model is an excellent case for the application of the univariate bootstrap, despite heteroscedasticity. There are only two variables of interest, which is a case where the univariate bootstrap is known to work well (e.g., Beasley et al., 2007; Beasley & Rodgers, 2012; Lee & Rodgers, 1998). The calculation of standard errors in DF analysis is not straightforward, which is a case where bootstrapping methods generally are advised. Lastly, it is nearly impossible to conceive of a case with non-selected twins where the DF analysis would contain a nonlinear effect (because items are double entered it is unlikely to make sense to say that someone's score would be a quadratic or other nonlinear function of their co-kin's score). Nonlinear effects are currently difficult to model using univariate bootstrapping, so their expected absence is a good safeguard. Heteroscedasticity would typically be an issue for the univariate bootstrap, however DF analysis is a special case where the logical resampling framework helps to obviate the issue.

The procedure to be used in this study is to simulate a setting in which we calculate sample correlations for each of the kinship groups in our sample (e.g., monozygotic twins, siblings and half siblings). A sampling frame using all possible pairs of the observed outcomes is created for each group, with the observed correlation for each group imposed on their sampling frame (this is a new feature of the univariate bootstrap, with certain advantages). Because diagonalization is imposed for a different sampling frame for each group, the natural heteroscedasticity is retained. This occurs because each kinship group (e.g., all identical twins as a group) is diagonalized separately from the other kinship groups. For each group, the number of pairs of data, equal to the number of original pairs in the group,

are then randomly selected with replacement from each sampling frame. The DF analysis is conducted on this sample. Unlike traditional DF analysis, double entry is unnecessary in this case because the repeated sampling of the bootstrap gives each co-kin an equal probability of being the predictor or the predicted. The standard errors are then formed using bootstrap confidence intervals. These intervals will typically be wider than those normally achieved using single-entry DF analysis because of the reduced sample sizes used in the bootstrap analysis, but narrower than those from double-entry settings. This method should also provide all of the typical advantages of bootstrapping (e.g., minimal distributional assumptions) that are not specific to the DF case.

Methods

All analyses were conducted in the R software package (R Core Team, 2019). Bootstrapping was and DF analyses were conducted using the Omisc package (O'Keefe, 2019). Source files are available in the code appendix. A simulation study was designed with 96 potential conditions. These were formed by crossing four factors: distribution, sample size, MZ:DZ balance and effect sizes. The distributions selected were normal, χ^2_1 and χ^2_{10}. The rationale behind these three distributions is that they provided a scenario consistent with parametric test distributional assumptions (the normal distribution), a moderately skewed case (χ^2_{10}), and a highly skewed distribution (χ^2_1). There were 2 sample sizes, 48 and 498 twin pairs, split between MZ and DZ twins. Forty eight twin pairs was chosen as being what might be expected from a convenience sample of twins, whereas 498 was chosen as what might be expected from a larger, more focused, twin study. A balanced and unbalanced twin design was used, with the unbalanced twin design having an exactly 2:1 DZ:MZ ratio. The 2:1 ratio was chosen as being approximately equal to the ratio of MZ to DZ twins in the general population. Finally, 0, 0.3 and 0.69 were used as the effect sizes for a^2 and c^2, representing no effect, a medium effect and a large effect. There were 8 allowable a^2 and c^2 effect size combinations (0, 0; 0, 0.3; 0, 0.69; 0.3, 0; 0.3, 0.3; 0.3, 0.69; 0.69, 0; 0.69, 0.3; note that 0.69, 0.69 cannot occur, because that combination sums to greater than 1).

Code for the univariate bootstrap was written in R. To test that it was performing as expected, full univariate sampling frames were created using the software and checked against what would be expected (i.e., variable means, variances and correlations were as expected), and a brief simulation study examining the univariate bootstrap CI properties was conducted. To examine the CI properties 10,000 simulations were run. For each simulation, 100 bivariate normal observations were selected with a population correlation of .3, and 1,000 bootstrap samples were taken and a CI created. The proportion of CI's that contained the true population value of .3 could not reject the nominal rate of .95, consistent with expectations that the software was behaving as expected.

Running in parallel, using the 'parallel' package in R (R Core Team, 2019), the 96 conditions took approximately eight days to run on a desktop computer. For each condition all of the confidence interval methods under consideration were conducted 10,000 times. For bootstrap methods 1,000 bootstrap resamples were used.

There were multiple plausible ways to conduct the bootstrap analyses. For the standard bootstrap it was possible to bootstrap prior to double entry, or after double entry. For cases after double entry it seemed worthwhile to examine the effects of taking a bootstrap sample equal to the double entered sample size (twice the number of twin pairs) versus taking a sample equal to the original number of twin pairs. For univariate bootstrapping all the possibilities for the standard bootstrap also existed. Additionally there was the possibility of using the entire sample (both MZ and DZ twins) as the source for each group and then diagonalizing afterwards (i.e., sampling within groups, and sampling ignoring group membership). However, because we are already assuming the same mean and variance for MZ and DZ twins on the focal variable, and the correlation is imposed after resampling, it should make little difference if observations were actually from an MZ or DZ twin. These considerations resulted in three standard bootstrapping schemes (double entry after bootstrapping, double entry before bootstrapping with a bootstrap resample the size of the double entered dataset, and double entry before bootstrapping with the bootstrap resample half the size of the double entered dataset) and six univariate bootstrapping schemes (the same three conditions as for standard bootstrapping crossed with sampling either within kinship groups or sampling using pooled kinship data). For all bootstrapping schemes, both a standard 95% CI was created as well as a BCa 95% CI using a jackknife estimate for the acceleration parameter (DiCiccio & Efron, 1996).

Ultimately there were 21 different confidence interval methods tested. The standard regression CI, the standard regression CI but with the interval width multiplied by the square root of two (to account for the doubling of the sample size due to double entry), the Kohler-Rodgers sandwich CI, six standard bootstrap CI's (half were standard intervals, half BCa intervals), and 12 univariate bootstrap CI's (half were standard intervals, half BCa intervals). Not all of these conditions will be described in detail within the Results section, though all results can be obtained from the first author. We will present the conditions of greatest utility for DF analysis researchers in the next section.

Results

The results are organized into sections as follows. First coverage rates of the various methods are presented and tested for deviations from the nominal coverage. After considering the general coverage rates the proportions of Type I errors that occur due to the confidence interval being too high or being too low are considered. Next the power of each method is presented and compared with other methods. Finally, a follow up simulation that helps illuminate some of the main results, and further confirms the reliability of the programming, is presented.

As a rule, the behavior of the CIs was quite similar within certain classes of CIs. For example, the bias corrected and accelerated CIs were not markedly different from the uncorrected versions. The kind of grouping used for the univariate bootstrap made no difference in this application. The univariate and the standard bootstrap performed similarly. The primary dividing line between methods that had adequate coverage and those that did not was whether or not the method attempted to account for double entry of the data in some way. Thus, bootstrap methods that used bootstrap resamples only half the size of the double

entered dataset had better coverage than those that did not. The square root of two correction also behaved better than methods that did not explicitly correct for sample size. Because the results were relatively similar within various classes of CIs we elected to present exemplars from each class, and particularly those which are of special interest.

Thus we present results from seven of the CI methods: the standard parametric CIs, the square root of two correction to those CIs, the Kohler-Rodgers robust correction to those CIs, univariate and standard bootstraps of the data prior to double entry, and standard and univariate bootstraps that take bootstrap resamples half the size of the previously double entered data. Methods not reported here functioned similarly to those reported. For example, the bias corrected and accelerated versions of all CIs performed nearly identically. The within group sampling and ungrouped sampling methods for the univariate bootstrap performed very similarly as well.

First, coverage will be addressed. A binomial distribution with $p = .95$ and an n of 10,000 produces cutoffs of .9457 and .9542 as the lower and upper bounds of a 95% CI. If a confidence interval method is used we would expect on average 95% of simulations using that method to capture the true population value at least 94.57% of the time and no more than 95.42% of the time. Given that researchers might allow for conservatism but not liberalism in a confidence interval we also evaluated the confidence interval methods using a cutoff of 94.64%, which is equivalent to a 1-tailed cutoff (i.e., if coverage was less than 94.64% we rejected the null hypothesis that the method had adequate coverage). Overall there were 96 conditions, each with 2 parameters, resulting in 192 tests for each confidence interval. Table III summarizes how often each confidence interval either properly captured the true population value or was not overly liberal (i.e., either proper or conservative).

Table III shows what proportion of intervals, averaged across all conditions within an interval, had expected (or non-liberal) coverage rates (i.e., a coverage rate of 95% or less). The first column shows what proportion had expected coverage rates (an ideal score would be 100 in this column). The second column shows what proportion had non-liberal coverage (coverage was not significantly less than 95%, but could be significantly lower). Based on table 3, it appears that no confidence interval method had ideal coverage; however, if conservatism is allowed there were several promising methods. In particular, the univariate bootstrap method that double entered prior to bootstrapping and then used bootstrap resamples half the size of the double entered data set had adequate or conservative coverage in all cases, although not shown in this table, standard bootstrapping methods that did the same behaved similarly. Table IV shows marginal coverage across conditions (i.e., the average coverage for each interval across parameter type, parameter value, MZ sample size, and population distribution separately). A version of this table that shows the results of all crossings of all conditions (i.e., the 96 simulated conditions and both regression parameters) can be found in Appendix A. Most confidence interval methods were overly liberal, and significantly so. Only bootstrapping methods where double entry occurred before bootstrapping and the bootstrap sample size was half the size of the double entered data set performed well by this metric across all conditions.

Next, we examined the probability of missing to the left or right. We were primarily concerned with too many misses to the left or right. We used a 1-tailed test for left and right misses (and treated missing left and right as separate events with separate tests). With 10,000 simulations and an expected miss rate of .025 for both left and right, it gave a cutoff of 276, that is, simulations in which there were more than 276 misses left or more than 276; misses right were considered statistically significantly different from expected. Note that a method could have adequate coverage (i.e., 95%) yet have more misses in one direction than expected (e.g., if misses were asymmetrical, 4% of misses occurred on the low end of the interval and 1% of misses occurred on the high end of the interval). In general, there was an imbalance in the miss pattern. Confidence intervals tended to miss such that the upper end of the confidence interval was lower than the actual population value slightly more than the converse. Table V shows this.

The next consideration was power. The number of times 0 was outside the lower bounds of the confidence interval was calculated for each condition for each confidence interval method for which the null hypothesis was incorrect and should be rejected (i.e., excluding conditions where the population value was 0). Table VI shows the power of all confidence interval methods marginalized across the simulation conditions as a proportion of times that zero was outside the confidence intervals. An additional table in Appendix A shows the same calculations for power in each simulation and for each CI method. Lower numbers indicate lower power. Overall, the highest power was found in the bootstrapping methods that used bootstrap samples equal in size to the sample being bootstrapped and the typical regression confidence interval. The bootstrap intervals that performed well in terms of their Type I error rate (i.e., those that double entered and took bootstrap sample sizes half the size of the double entered sample) and the square root of two corrected typical CI perform poorer in terms of power. This is exactly in line with the Type I error rate results given the typical tradeoffs between power and Type I errors.

The fact that the bootstraps that double entered and then took bootstrap samples half the size of double entered data set did much better than all the other methods was surprising. In order to make sure that this was not due to a coding error, we implemented a small simulation to confirm the behavior outside of DF models. The results are both confirmatory and illuminating. For this simulation 100 bivariate normal observations with standard deviations of 1, means of 5, and correlations of .3 were generated (Table VII). The cor.test function in the *R stats* package (R Core Team, 2015) was used to obtain the standard confidence intervals. A univariate bootstrap and typical bootstrap confidence interval were constructed, followed by a univariate and typical bootstrap that used bootstrap resamples twice the size of the original sample. Then the data were double entered and univariate bootstraps, typical bootstraps using both the full double entered data sample size and a sample size equal to half that were used to obtain four more confidence intervals. This simulation was repeated 1,000 times. The results match the results above and provide some insight into the process underlying the results. When the bootstrap sample size is greater than the actual effective sample size it reduces the variability of the bootstrap resamples' parameter estimates around the sample parameter estimate, producing confidence intervals that are too narrow and that have an alpha level far higher than the nominal rate. Table VII shows the actual alpha rate

for the various confidence interval methods. The highest alpha rate expected with 1,000 simulations is 6.2%.

Empirical Analysis

Method

We now turn to a short empirical analysis. Our analysis uses data from the National Longitudinal Survey of Youth 1979 (NLSY79) sample. The NLSY79 is a household probability sample that followed adolescents from 1979 to the present on an approximately biennial basis (the survey was annual in early years of the study). The NLSY79 provides a rich, biometrically informed, dataset for analysis. For the present analysis we chose to look at individual BMI and its change over time. Individual BMI has a significant impact on individual health over time (e.g., Kopelman 2007). Previous research using the Framingham Heart Study has demonstrated heritability of BMI but a lack of heritability for change in BMI (Coady et al. 2002), we replicate the latter finding here. Anecdotally, one might expect that if one's parents or older siblings had a notable change in their BMI over time that one would experience a similar change. This belief reflects either a shared genetic or shared environmental influence on BMI trajectory.

In the NLSY79 information was available for nearly every survey administration for individual's weight. Height was not measured as consistently, but was measured in the first few years of the survey and then several times during the most recent survey administrations. Height and weight are the only measures need to calculate BMI. For years with an observed weight but not observed height we imputed the missing height on a person by person basis. Our imputation method was fairly simple as we reason that adult height is not generally highly variable, and height can be measured with a high degree of accuracy. For imputation we used a weighted mean of the height measures available in 1985 and 2006. If an individual was missing either of those observations we simply replaced the missing observations with the other observed height (if both were missing no imputation was done). Although this imputation process is somewhat ad hoc, height should not vary substantially for an individual in this sample between those times. There were over 11,000 individuals with at least one measure at either time point and over 7,000 with measures at both time points. Years earlier than 1985 had more observations, however many of the participants were not adults during those years and may not have attained adult height. In 1985 the youngest observed participant was 20 and so all participants were likely at (or very nearly at) their adult height.

The imputation process allowed us to calculate 185,843 BMI's for 12,575 individuals. This gave us, on average, approximately 14 observations per individual. For each individual we then calculated the slope of the regression of BMI on Year. We excluded individuals whose absolute value of their slope was greater than 1. For a man of average height a slope of one or more would suggest an average annual weight gain or loss of 7 pounds for the duration of their available data, for a woman of average height this would be a weight gain or loss of approximately 6 pounds. The vast majority of individuals were included under this criterion. Our interest was in the heritability of this slope.

Results

Using a publically available package, NlsyLinks (Beasley 2018; see Rodgers et al, 2016, for background), we then used known biometrical relationships to create pairs of observations for full and half siblings. There were 284 half sibling pairs and 3,881 full sibling pairs. A DF analysis indicated BMI a^2 = 0.20, and BMI c^2 = .03. Neither the a^2 nor c^2 component was statistically significant. However the width of the confidence intervals does vary considerably across the various methods (Table VIII).

These results suggest that the variability in the trend of weight gain or loss is due to non-shared environmental factors (or measurement error, which is confounded with E), and is not due to genetics or shared environmental factors, consistent with previous findings. We also notice that the confidence interval widths follow nearly the same pattern as in the simulation study. The narrowest confidence intervals are the standard confidence interval and bootstrap intervals that do not correct for double entry. The widest are bootstrap intervals that correct for double entry (either by double entering data after bootstrapping or by taking bootstrap samples only half the size of the double entered data), the square-root-of-two correction and the Kohler-Rodgers robust confidence interval. The Kohler-Rodgers robust interval behaving similarly to the corrected bootstrap and the square-root-of-two correction is somewhat different than in the simulation study, but otherwise the pattern is similar.

Discussion

Overall it would appear that, if more weight is given to avoiding Type I errors than Type II errors, bootstrapping or a correction using the square root of two should be the preferred method to construct CIs in DF analysis. In particular, when bootstrapping, data should be double entered and then a bootstrap sampling scheme that samples half the size of the double entered data (i.e., the original sample size) should be used. This method had somewhat lower power and slightly wider CIs, however it captured the true population value at a far higher rate than other methods across all conditions. The square-root-of-two correction might be an option in some settings. That method had slightly higher power in general, but was generally more liberal than bootstraps that corrected for double entry. This greater liberalism arguably outweighs any power benefits. Results for the univariate bootstrap and the standard bootstrap were similar. However, previous work by Beasley et al. (2007) suggested that the univariate bootstrap may be preferred because of the potential for superior performance in other settings, particularly with skewed distributions.

The final simulation illuminates why the most effective bootstrap method was double entry followed by a bootstrap half the size of the double entered sample size (an m < n bootstrap). When the sample size is inflated, either by using a bootstrap resample that is larger than the original sample size, or using a sample that is double entered, the bootstrap appears to lack the necessary variability; as a result, overly narrow confidence intervals are obtained, which then have Type I error rates substantially lower than expected. In DF models, although heteroscedasticity and non-independence of errors exist, the primary driver of CI inaccuracy is the doubling of the sample size with double entry. Although the Type I error rate was substantially better in the m < n bootstraps, the power was lower. Researchers may be tempted to use the other methods for the sake of improved power, but that cannot be

recommended here. Although power could be increased to virtually one with increasing sample size, no method is appropriate if Type I error rates are not controlled to avoid a liberal direction. This study shows that several methods are flawed with regards to Type I error rates and only correcting for the sample size can resolve this.

Overall the univariate bootstrap lived up to expectations, performing similarly to, or better than, the standard bootstrap. Further, its simplicity compared to a standard bootstrap with bias correction and acceleration is notable. Given that this represents the first full-scale application of the univariate bootstrap beyond bivariate correlations, this finding is encouraging for future research regarding the application of the univariate bootstrap to more advanced applications. With regards to DF models specifically, researchers can reasonably use a sample size corrected univariate bootstrap. The square root of two correction presented here, while better than the other non-bootstrap approaches, was rather liberal and we cannot advise its use despite its advantage in simplicity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors express their appreciation to Stacy Cherny, Mike Hunter, and Niels Waller, whose comments improved this paper.

References

Beasley W (2018). NlsyLinks: Utilities and kinship information for research with the NLSY. http://liveoak.github.io/NlsyLinks, https://github.com/LiveOak/NlsyLinks, https://r-forge.r-project.org/projects/nlsylinks.

Beasley WH, DeShea L, Toothaker LE, Mendoza JL, Bard DE, & Rodgers JL (2007). Bootstrapping to test for nonzero population correlation coefficients using univariate sampling. Psychological Methods, 12(4), 414–433. https://doi.Org/10.1037/1082-989X.12.4.414 [PubMed: 18179352]

Beasley WH & Rodgers JL (2012). Bootstrapping and Monte Carlo methods. APA Handbook of Research Methods in Psychology, 2, 407–425.

Cherny SS, Cardon LR, Fulker DW, & DeFries JC (1992). Differential heritability across levels of cognitive ability. Behavior Genetics, 22(2), 153–162. [PubMed: 1596255]

Coady SA, Jaquish CE, Fabsitz RR, Larson MG, Cupples LA, & Myers RH (2002). Genetic variability of adult body mass index: a longitudinal assessment in Framingham families. Obesity research, 10(7), 675–681. [PubMed: 12105290]

DeFries JC, & Fulker DW (1985). Multiple regression analysis of twin data. Behavior Genetics, 15(5), 467–473. [PubMed: 4074272]

DiCiccio TJ, & Efron B (1996). Bootstrap confidence intervals. Statistical science, 189–212.

Domingue BW, Conley D, Fletcher J, & Boardman JD (2016). Cohort effects in the genetic influence on smoking. Behavior genetics, 46(1), 31–42. [PubMed: 26223473]

Edgington ES (1987). Randomization tests. New York: Marcel Dekker.

Efron B (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia: Society for Industrial and Applied Mathematics.

Efron B (1987). Better bootstrap confidence intervals. Journal of the American Statistical Association, 82(397), 171 10.2307/2289144

Efron B, & Tibshirani R (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 54–75.

Jackson DB (2016). The link between poor quality nutrition and childhood antisocial behavior: A genetically informative analysis. Journal of Criminal Justice, 44, 13–20.

Kaiser HF, & Dickman K (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. Psychometrika, 27(2), 179–182.

Kohler H-P, & Rodgers JL (2001). DF-analyses of heritability with double-entry twin data: Asymptotic standard errors and efficient estimation. Behavior Genetics, 31(2), 179–191. [PubMed: 11545535]

Kopelman P (2007). Health risks associated with overweight and obesity. Obesity reviews, 8, 13–17. [PubMed: 17316295]

Lee W-C, & Rodgers JL (1998). Bootstrapping correlation coefficients using univariate and bivariate sampling. Psychological Methods, 3(1), 91.

Maczulskij T & Bockerman P (2019). harsh times: Do stressors lead to labor market losses? The European Journal of Health Economics, 20, 357–373. [PubMed: 30178149]

Meldrum RC, & Barnes JC (2017). Unstructured socializing with peers and delinquent behavior: A genetically informed analysis. Journal of youth and adolescence, 46(9), 1968–1981. [PubMed: 28451940]

O'Keefe P (2019). Omisc: Univariate Bootstrapping and Other Things. R package version 0.1.2 https://CRAN.R-project.org/package=Omisc

Purcell S & Sham PC (2003). A model-fitting implementation of the DeFries-Fulker model for selected twin data. Behavior Genetics. 33 (3), 271–278. doi: 10.1023/a:1023494408079 [PubMed: 12837017]

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria URL https://www.R-project.org/.

Rodgers JL (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. Multivariate Behavioral Research, 34(4), 441–456. 10.1207/S15327906MBR3404_2 [PubMed: 26801635]

Rodgers JL, & Beasley WH (2012). Fisher, Gosset, and alternative hypothesis significance testing (AHST): Using the bootstrap to test scientific hypotheses about the multiple correlation In Edwards MC & MacCallum RC (Eds.), Current topics in the theory and application of latent variable models (pp. 217–239). Routledge.

Rodgers JL, Beasley WH, Bard DE, Meredith KM, Hunter MD, Johnson AB, Buster M, Li C, May KO, Garrison SM, Miller WB, van den Oord E, Rowe DC (2016). The NLSY kinship links: Using the NLSY79 and NLSY-Children data to conduct genetically-informed and family-oriented research. Behavior Genetics, 46, 538–551. DOI 10.1007/s10519-016-9785-3. [PubMed: 26914462]

Rodgers JL, & Kohler H-P (2005). Reformulating and simplifying the DF analysis model. Behavior Genetics, 35(2), 211–217. [PubMed: 15685433]

Rodgers JL, Kohler H-P, Kyvik K & Christensen K (2001). Genes affect human fertility via fertility motivations: Findings from a contemporary Danish twin study. Demography, 38, 29–42. [PubMed: 11227843]

Rodgers JL and McGue M 1994 A simple algebraic demonstration of the validity of the DeFries-Fulker analysis in unselected samples with multiple kinship levels. Behavior Genetics, 24, 259–62. [PubMed: 7945155]

Rodgers JL, Rowe DC, & Li C (1994). Beyond nature versus nurture: DF analysis of nonshared influences on problem behaviors. Developmental Psychology, 30, 374–384.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria URL http://www.R-project.org/.Stine, R. (1989).

Roos JM & Nielsen F (2019). Outrageous fortune or destiny? Family influences on status achievement in the early life course. Social Science Research, 80, 30–50. 10.1016/j.ssresearch.2018.12.007 [PubMed: 30955560]

Schwartz JA, Connolly EJ, Nedelec JL, & Beaver KM (2017). An investigation of genetic and environmental influences across the distribution of self-control. Criminal justice and behavior, 44(9), 1163–1182.

Stine R (1989). An introduction to bootstrap methods: Examples and ideas. Sociological Methods & Research, 18(2-3), 243–291.

Tibshirani R (1985). How many bootstraps? Stanford University CA Department of Statistics.

Waller NG (1994). A DeFries and Fulker regression model for genetic nonadditivity. Behavior Genetics, 24, 149–53. [PubMed: 8024531]

Wu CFJ (1986). Jackknife, bootstrap and other resampling methods in regression analysis. the Annals of Statistics, 1261–1295.

York C (2017). A regression approach to testing genetic influence on communication behavior: Social media use as an example. Computers in Human Behavior, 73, 100–109.

Table I:

Example of kinship data.

K_1	K_2	R
9	20	1
8	18	1
21	16	1
7	19	1
19	17	.5
7	21	.5

Table II:

Example of double entered kinship data.

K_1	K_2	R
9	20	1
8	18	1
21	16	1
7	19	1
19	17	.5
7	21	.5
20	9	1
18	8	1
16	21	1
19	7	1
17	19	.5
21	7	.5

Table III:

Proportion of confidence intervals out of 192 simulated intervals (each replicated 10,000 times) with expected coverage or non-liberal coverage.[a,b]

	Proportion with expected coverage	Proportion with non liberal coverage
Robust CI	0.13	0.08
Typical CI	0.01	0.10
Typical Adjusted by $\sqrt{2}$	0.19	0.72
Standard DEA Bootstrap standard CI	0.16	0.11
Univariate DEA WGS Bootstrap standard CI	0.30	0.26
Standard DEB .5 Bootstrap standard CI	0.15	0.87
Univariate DEB .5 WGS Bootstrap standard CI	0.05	1.00

[a]DEA: Double entry after bootstrapping; DEB: Double entry before bootstrapping; .5: bootstrap resample size was half the (double entered) sample size; 1: bootstrap resample size was equal to the size of the (double entered) sample size; WGS: within group sampling was used for the univariate bootstrap; UGS: ungrouped sampling, or sampling without regard to class membership was used for the univariate bootstrap.

[b]Test for proper cases was two-tailed, test for non-liberal was one tailed. This resulted in some intervals faring poorer in the non-liberal Type I error case than in the expected Type I error test.

Table IV:

Table coverage (in proportion) marginalized across different simulation settings.[c]

	Population Beta Weights			MZ:DZ Twin Pair Sample Size				Parameter		Population		
	0	**0.3**	**0.69**	**16:32**	**24:24**	**166:332**	**249:249**	**a2**	**c2**	χ^2_1	χ^2_{10}	**normal**
Robust CI*	0.91	0.89	0.89	0.88	0.87	0.92	0.92	0.90	0.89	0.84	0.92	0.93
Typical CI*	0.84	0.83	0.86	0.88	0.86	0.83	0.81	0.86	0.82	0.75	0.88	0.90
Typical CI Corrected by $\sqrt{2}$	0.95	0.94	0.95	0.96	0.95	0.93	0.92	0.95	0.93	0.89	0.97	0.97
Standard DEA Bootstrap standard CI*	0.93	0.91	0.90	0.91	0.91	0.91	0.92	0.91	0.91	0.87	0.93	0.93
Univariate DEA WGS Bootstrap standard CI*	0.94	0.92	0.92	0.93	0.93	0.92	0.93	0.93	0.93	0.9	0.94	0.94
Standard DEB .5 Bootstrap standard CI	0.96	0.97	0.99	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
Univariate DEB .5 WGS Bootstrap standard CI	0.97	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.97

[c] Significantly lower coverage than expected at $p < .05$ across all conditions. Note: The lowest admissible coverage varies slightly across conditions, however the lowest admissible rate is .949. All numbers are proportions.

Table V:

Number of tested intervals (simulated 10,000 times each) out of 192 (with proportion in parentheses) that had significantly many misses [d]

	Miss Interval too Low	Miss Interval too High
Robust CI	146(76)	133(.69)
Typical CI	163(.85)	159(.83)
Typical CI Correct by $\sqrt{2}$	53(.28)	44(.21)
Standard DEA Bootstrap standard CI	159(.83)	100(.52)
Univariate DEA WGS Bootstrap standard CI	130(.68)	99(.52)
Standard DEB .5 Bootstrap standard CI	40(.21)	11(.06)

[d] using this method we would expect each CI method to have approximately nine or ten (0.05 × 192) cases in which the method was found to have too many misses due to chance [i.e., the expected value of each cell under perfect conditions is 9.6 (0.05)].

Table VI:

Power to detect population deviation from zero for each method marginalized across simulation conditions.

	Population Beta Weight		MZ:DZ Twin Pair Sample Size				Parameter		Population		
	0.3	**0.69**	**16:32**	**24:24**	**166:332**	**249:249**	**a2**	**c2**	χ^2_1	χ^2_{10}	**normal**
Robust CI	0.47	0.87	0.50	0.48	0.77	0.76	0.64	0.61	0.54	0.66	0.68
Typical CI	0.52	0.86	0.42	0.47	0.86	0.86	0.63	0.68	0.60	0.67	0.68
Standard Adjusted by $\sqrt{2}$	0.36	0.75	0.24	0.29	0.77	0.77	0.47	0.56	0.47	0.53	0.54
Standard DEA Bootstrap standard CI	0.44	0.84	0.45	0.43	0.76	0.76	0.64	0.56	0.50	0.64	0.66
Univariate DEA WGS Bootstrap standard CI	0.44	0.83	0.43	0.41	0.77	0.76	0.63	0.56	0.49	0.63	0.66
Standard DEB .5 Bootstrap standard CI	0.30	0.71	0.23	0.21	0.71	0.7	0.43	0.49	0.32	0.51	0.55
Univariate DEB .5 WGS Bootstrap standard CI	0.28	0.67	0.21	0.18	0.69	0.67	0.42	0.46	0.27	0.49	0.55

Table VII:

Type I error rates for various confidence interval methods using bivariate normal data with a correlation of .3 [e]

Confidence Interval Method	Type I error rate
Standard Confidence interval	4.5%
Univariate Bootstrap NDE k=1	4.4%
Standard Bootstrap NDE k=1	5.5%
Univariate Bootstrap NDE k=2 [*]	18.5%
Standard Bootstrap NDE k=2 [*]	19.2%
Univariate Bootstrap DE k=1 [*]	17.8%
Standard Bootstrap DE k=1 [*]	19.2%
Univariate Bootstrap DE k=.5	5.1%
Standard Bootstrap DE k=.5	5.7%

[e](N)DE: (Not) Double Entered; k: Bootstrap resample size is that multiple of input data size (e.g., if not double entered, and k =1, bootstrap resample size is n, if double entered 2n).

[*]Significantly greater Type I error rate than expected $p < .05$.

Table VIII:

Confidence intervals for A and C components from a DF analysis of adult BMI slope

	a^2 lower	a^2 upper	c^2 lower	c^2 upper
Robust CI	−0.30	0.70	−0.21	0.28
Typical CI	−0.14	0.54	−0.13	0.20
Typical CI adjusted by $\sqrt{2}$	−0.20	0.76	−0.18	0.28
Standard DEA Bootstrap CI	−0.30	0.68	−0.20	0.27
Univariate DEA WGS Bootstrap CI	−0.28	0.65	−0.19	0.27
Standard DEB .5 Bootstrap CI	−0.34	0.72	−0.22	0.29
Univariate DEB .5 WGS Bootstrap CI	−0.34	0.69	−0.21	0.29