BEST PRACTICES FOR CONSTRUCTING CONFIDENCE INTERVALS FOR THE GENERAL LINEAR MODEL UNDER NON-NORMALITY

Mark C. Adkins

A THESIS SUBMITTED TO

THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF ARTS

GRADUATE PROGRAM IN PSYCHOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO

November 2017

© Mark C. Adkins, 2017

Abstract

Given the current climate surrounding the replication crisis facing scientific research, a subsequent call for methodological reform has been issued which explicates the need for a shift from null hypothesis significance testing to reporting of effect sizes and their confidence intervals (CI). However, little is known about the relative performance of CIs constructed following the application of techniques which accommodate for non-normality under the general linear model (GLM). We review these techniques of normalizing data transformations, percentile bootstrapping, bias-corrected and accelerated bootstrapping, and present results from a Monte Carlo simulation designed to evaluate CI performance based on these techniques. The effects of sample size, degree of association among predictors, number of predictors, and different non-normal error distributions were examined. Based on the performance of CIs in terms of coverage, accuracy, and efficiency, general recommendations are made regarding best practice about constructing CIs for the GLM under non-normality.

Dedication

I dedicate this to my amazing wife, Sara, who is unquestionably the funniest person I know. Without her unwavering love and support this present work would not have been possible.

TABLE OF CONTENTS

Abstract.....ii

Dedication.....iii

Table of Contents iv

List of Tables v

List of Figures vi

References.....31

List of Tables

Table 1: Confidence intervals for empirical example.....11

Table 2: Confidence Interval Properties by Sample Size 20

Table 3: Confidence Interval Properties by Error Distribution.....21

Table 4: Confidence Interval Properties by the Degree of
Association Among Independent Variables.....22

Table 5: Confidence Interval Properties by Number of Independent Variables 23

Table 6: Best Efficiency per Condition with Proper Coverage.....25

List of Figures

Figure 1: Average Confidence Interval Coverages 17

Figure 2: Average Confidence Interval Efficiency 18

Figure 3: Average Confidence Interval Lower Proportions.....18

Figure 4: Average Confidence Interval Upper Proportions 19

Best Practices for Constructing Confidence Intervals For the General Linear Model under Non-normality

Renewed concerns regarding the dependability of psychological findings have highlighted the importance of research practice such as data analysis (Stangor & Lemay, 2016; Vazire, 2015) and reporting effect sizes (Kelley & Preacher, 2012) with their confidence intervals (CIs; Wilkinson, 1999; Nickerson, 2000; Cumming, 2014). The general linear model (GLM), which includes ANOVA and multiple linear regression as special cases, is a popular modeling framework within psychological research, and it typically assumes a normal distribution of the random errors with homogeneity of variance so that inferential information (e.g., CI coverage) is accurate. However, psychological variables are often non-normally distributed (Cain, Zhang, & Yuan, 2016; Micceri, 1989), raising the practical question of how best to address potential violation of the GLM assumption regarding the distribution of random errors. Several approaches of varying technicality have been developed to address violation of the structure of the GLM random errors, but these approaches are not often familiar to substantive researchers. The overall purpose of this current research is to present a comprehensive evaluation of alternative methods designed to accommodate violations of the assumption of normality in relation to effect sizes and their confidence intervals. Additionally, recommendations for best practice about which method to apply under different data conditions are presented.

General Linear Model

The general linear model can be expressed as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \qquad (1)$$

where $i = 1, \ldots, N$ denotes each individual in the sample of size N, and y_i is the observed value on the continuous dependent variable (DV) for individual i which is predicted by the $k = 1, \ldots, K$ independent variables (IV; e.g., x_k). The intercept, β_0, is interpreted as the expected value of y when all IVs equal zero. Each of the K IVs has an associated regression coefficient (e.g., β_k) which is interpreted as the expected

change in y for a one-unit increase in x_k, while holding all other IVs in the model constant. Finally, ε_i are the population random errors associated with each case i. Estimates of these random errors (hereafter called residuals and denoted as $\hat{\varepsilon}_i$) are the deviation of an individual's value on the DV y_i from the model-implied predicted value in Equation 1. In order to make inferences under ordinary least squares (OLS) estimation, it is assumed that these random errors are multivariate normally distributed with mean vector $\mathbf{0}$ and variance structure $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_N$. Here, σ^2 is the variance of the random errors around the model-implied expected value on the DV and $\mathbf{I}_N$ is an identity matrix of size N. This variance structure has the property of homogeneity of variance. If expanded into matrix form, the variance structure of the errors can be seen more clearly. As an example, consider the case of $N = 4$ cases: The variance structure matrix, $\boldsymbol{\Sigma}$, is expected to be as follows: $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{4x4} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$, where each case's error is normally distributed with variance equal to σ^2. Grouping all these assumptions together, we have $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\varepsilon}$ is the $1 \times N$ vector of random errors.

The key results of the GLM are the coefficient estimates and their confidence intervals (CIs). In the GLM, the estimated regression coefficients, $\hat{\beta}_k$, are effect size statistics that convey information about both the direction and magnitude of the expected change in y due to a change in x_k. Generally, an effect size in GLM is defined as "a measure of the magnitude of a phenomenon being studied" (Cohen, Cohen, West, & Aiken, 2003, p. 47). An important distinction between highlighting effect sizes over test statistics (along with their corresponding p-values) is that the former shifts focus away from making a dichotomous decision about the statistical significance of an effect toward the magnitude and direction of the effect. Test statistics and p-values do not convey information about the magnitude of an effect. In contrast, effect sizes also help answer the pertinent question of whether an effect has practical importance, where the interpretation of effects sizes "requires informed judgment in context" (Cumming, 2014). Effect size interpretations do not stand on their own apart from theory. The practical importance of an effect is determined by comparison to previous research or substantive theory. For example, if a mean difference

of 2 points was observed on some measure of pain ranging from 0 to 20 points, a researcher may understand enough about the pain measure to determine that a difference of this size is large enough to be noticeable for patients. By contrast, interpreting a *p*-value of 0.0035 does not answer the question of whether the reduction in pain would be noticeable to patients undertaking treatment. Such information cannot be assessed using *p*-values.

Confidence Intervals

The CIs of estimated regression coefficients convey their stochastic nature and precision. A $(1 - \alpha)100\%$ CI provides a range of values for a particular population parameter, such as a regression coefficient. Over repeated sampling, it is expected that $(1 - \alpha)100\%$ of similarly constructed CIs cover the unknown parameter (e.g., β_k), where α is the nominal Type I error rate. Assuming that the underlying population distribution of $\hat{\beta}_k$ is normal, confidence intervals for regression coefficients are constructed using the following formula:

$$\hat{\beta}_k \pm t_{1-\frac{\alpha}{2},df} \times SE_{\hat{\beta}_k} \qquad (2),$$

where $\hat{\beta}_k$ is the estimated value of the population parameter, $t_{1-\frac{\alpha}{2},df}$ is the critical value associated with the $(1 - \frac{\alpha}{2})$th quantile of the *t*-distribution with $df = (N - K - 1)$ degrees of freedom, and $SE_{\hat{\beta}_k}$ is the standard error of the estimate $\hat{\beta}_k$. The *t*-distribution in Equation 2 is based on the known sampling distribution of $\hat{\beta}_k$ given the normal distribution assumption about the distribution of errors. A sampling distribution is a theoretical distribution which represents the probability distribution of a sample statistic across repeated samples of the same size drawn from the same population. The assumption of multivariate normality and homoscedasticity of $\boldsymbol{\varepsilon}$ ensures that the sampling distribution of the regression coefficients follows a standard normal distribution. Because $\hat{\sigma}^2$ is estimated, a *t*-distribution is used in place of a standard normal distribution. CIs constructed using Equation 2 will be referred to as standard CIs.

The properties of standard CIs computed from Equation 2 depend on whether the underlying assumptions of the model concerning the variance structure of **ε** (i.e., normality and homoscedasticity) are met. When the assumed form of this variance structure is violated, three important properties of CIs can be affected. These three properties of CIs are coverage, accuracy, and efficiency. *Coverage* is the percent of CIs, over repeated sampling, that contain the population parameter. More formally, coverage is defined as

$$Coverage = \left[1 - \frac{\sum_{r=1}^{R} 1(\beta_k < lb) + \sum_{r=1}^{R} 1(\beta_k > ub)}{R} \right] 100\% \qquad (3),$$

where $r = 1, \ldots, R$ is the total number of replicates, 1(.) is an indicator function, β_k is the population regression parameter, and *lb* and *ub* are the lower and upper bounds of the CI, respectively. Typically, the Type I error rate in psychological research is set at $\alpha = 0.05$ which makes the expected confidence limit $(1 - \alpha)100\% = 95\%$. *Accuracy* of a CI relates to the tail proportions associated with the lower and upper bounds within the kernel of Equation 3. These tail proportions are estimates of the true population tail probabilities of the parameter not being captured by the nominal proportion of confidence intervals. The upper tail proportion is $\frac{1}{R}\sum_{r=1}^{R} 1(\beta_k > ub)$ and the lower tail proportion is $\frac{1}{R}\sum_{r=1}^{R} 1(\beta_k < lb)$. A CI is accurate when each of the two tail proportions is approximately equal to α/2 because the CI should capture the central $(1 - \alpha)100\%$ of parameter estimates. For example, an optimally performing 95% CI should have 2.5% in each of the tail proportions. *Efficiency* is defined by the width of the confidence interval. A CI is more efficient when there is less variability of an estimate, indicating a higher level of precision. When the model is correct in the population (i.e., no misspecifications about the regression equation and the nature of the errors), estimates are unbiased and maximally efficient. Efficiency is a reflection of estimated precision and sampling variability; narrow CIs reflect higher precision and lower sampling variability, and vice versa. One important note about efficiency is that it loses meaning if a CI does not have proper coverage and accuracy. In such a case, a narrow CI becomes a precise estimation

that is incorrect because it rarely, or never, captures the parameter of interest. Ultimately, such a CI becomes useless in terms of statistical inference. Efficiency can act as a method to arbitrate between two competing CI methods as long as both methods have proper coverage and accuracy first.

For example, if students' GPAs were regressed on age and the number of hours per week a student worked on homework, and the resulting coefficient for homework was 0.75, then this effect can be interpreted such that every one-unit increase in hours spent doing homework, there is an expected increase of 0.75 units in GPA, while holding age constant. This effect also has a 95% standard CI = [0.54, 0.96], indicating that over repeated sampling, 95% of similarly constructed CIs will capture the true coefficient. To illustrate the precision and efficiency of CIs, suppose the same data that produced the homework effect estimate was repeatedly resampled to create a 95% bootstrap interval = [0.58, 0.92]. Assuming that both types of intervals maintain proper coverage, this bootstrap interval has a narrower width indicating both a higher level of precision around the effect estimate and greater efficiency. Given that the interpretation of a CI is contingent upon repeated sampling, conclusions regarding the coverage of any single CI can be problematic. In practice, a single CI is calculated around a given effect estimate, and this particular CI has either 0% or 100% coverage of its associated true unknown population parameter.

Violation of Normality

When the assumptions of the GLM are met, estimates and standard CIs constructed using Equation 2 are unbiased; CIs have proper coverage, are accurate, and are maximally efficient (Cohen, Cohen, West & Aiken, 2003). However, the assumption of normality is often violated in practice, and frequently occurs alongside violations of homogeneity of variance (Kutner, Nachtsheim, & Neter, 2004). Heterogeneity of variance among the residuals can often result from skewed distributions, which are common in psychological research (Cain, Zhang, & Yuan, 2016). Heteroscedasticity can adversely affect CI properties for parameter estimates by producing standard errors which are biased and inconsistent (Hayes & Cai, 2007). For instance, if residuals vary less at the extremes of an IV, then OLS estimates of standard errors tend to be overestimated or upward biased (i.e., *bias* > 0). This upward bias results in CIs

which are wider than they would be if the upward bias was not present, communicating less precision than is correct. Alternatively, residuals which vary more at relatively high or relatively low values of an IV result in underestimated standard errors or a downward bias (i.e., *bias* < 0) compared with OLS standard errors. A downward bias in standard error estimates results in CIs which are narrower than proper, indicating more precision than warranted. Stated differently, coverage is less than the nominal rate. Alternatively, if standard errors are upward biased, then coverage is larger than the nominal rate.

Approaches to address violation of normality

When normality is violated in practice, researchers have several options. The more common approaches which will be reviewed here are the reliance on the robustness of the GLM methodology itself (i.e., assume that the effects of assumption violations are negligible), performing data normalizing transformations, and the application of bootstrap techniques.

Robustness of the method

When faced with violations of model assumptions such as normality, researchers often rely on the Central Limit Theorem. This theorem states that as sample size increases, the sampling distribution of regression coefficient becomes approximately normal (Moore, McCabe, & Craig, 2014). This claim avoids the need to assume normality of errors when using the GLM framework (or any other framework requiring distributional assumptions), provided sufficiently large sample size.

There is some historical precedent for the claim that the methodology surrounding the GLM is generally robust to assumption violations. Early studies examined the effect of small deviations from different assumptions in isolation (e.g., *t*-tests comparing two samples with variances of 1 and 4; Boneau, 1960) and concluded that parametric tests similar to the GLM were robust in many circumstances if sample sizes were at least 25 to 30 and the underlying distributions were comparable in shape. However, many studies did not examine large deviations from assumptions such as those found in real psychological data, which can have variance ratios (defined as the ratio between the largest and smallest

sample variances) much greater than the 4:1 ratios examined in early studies (e.g., variance ratios of 7:1 or even higher; Keselman et al., 1998). Yet, Pek, Wong, and Wong (2017) showed that "large" deviations from assumptions depend on how much the model residuals deviate from normal in terms of skewness and kurtosis. Although the above cited examples specifically examined special cases of the GLM (e.g., *t*-tests and ANOVA), there is evidence that results from conducting a multiple linear regression exhibit robustness as well. Specifically, the assumption of normality of model errors has been termed arbitrary as long as sample size is sufficiently large to invoke the CLT (Fox, 1991). A caveat to this statement is that even though the significance tests and CIs are correct, OLS estimation can suffer a loss in efficiency when the distribution of residuals is heavy-tailed. At this point, the belief that the GLM methodology is robust to assumption violations is still present and evidenced by the claims of robustness within many research methods textbooks (Erceg-Hurn & Mirosevich, 2008).

Although many of these early studies examined robustness in the sense of maintaining the nominal Type I error rate, the presence of assumption violations can result in CIs which are not robust when "real data are analyzed" (Erceg-Hurn & Mirosevich, 2008, p. 600). For instance, if residuals come from a heavy-tailed error distribution, there is a loss in efficiency due to larger standard errors which directly affects CI coverage and efficiency. Lower efficiency is synonymous with wider CIs and decreased precision about effect estimates. For a review of the effect assumption violations have on parametric tests (such as those in the GLM), see Glass, Peckham, and Sanders (1972).

Normalizing data transformations

Another common technique to address the violation of normality is the use of data transformations. A data transformation replaces raw scores, such as y_i, with new scores which have been rescaled by a monotonic transformation function, $f(.)$, such that transformed scores, y_i^*, can be expressed with $y_i^* = f(y_i)$ (Kutner, Nachtsheim, & Neter, 2004). Using an appropriate choice of transformation function can stabilize the variance of residuals, restore a linear relationship between the DV and residuals,

and normalize the distribution of residuals (Box & Cox, 1964). Often, a suitable transformation will accomplish these objectives simultaneously. For instance, a common transformation for psychological data is the logarithmic transformation. The logarithmic transformation can normalize a positively skewed distribution while also reducing residual variance (Bartlett, 1947), provided that the data undergoing transformation are all positive prior to transformation. Corrections can be made to data to ensure all values are positive, such as adding a constant. After applying a transformation, subsequent analyses are conducted using the newly transformed variable. Using the GLM framework, the logarithm of the DV is regressed on the IVs and slope estimates are produced. The interpretation of these estimates changes due to the transformation; under a logarithmic transformation, the regression coefficient becomes the expected change in the logarithm of the DV for a one-unit increase in a specific IV. Regression coefficients can also be interpreted as the expected percentage change of the DV on its original scale for a one-unit increase in a specific IV. Likewise, CIs around these effects sizes are also on the logarithmic scale of the DV. Construction for CIs around these effect sizes are calculated using Equation 2.

One criticism of data transformations is that, once applied, any subsequent analyses are performed on the transformed scale and are no longer logically connected to original research questions (Feng et al., 2014). Back-transformations of parameter estimates into the original scale of measurement may not directly map onto appropriate estimates of the original data.

An additional criticism regarding the logarithmic transformation specifically is that it changes the original model from being additive to a multiplicative model. By taking the logarithm of the DV, the model on the original scale of the DV becomes

$$y_i = e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i} . \tag{4}$$

By substituting $\alpha_k = e^{\beta_k}$ where $k = 0$ to K and $u = e^{\varepsilon_i}$ into Equation 4, the multiplicative nature of the model becomes clear:

$$y_i = \alpha_o * {\alpha_1}^{x_1} * \ldots * {\alpha_K}^{x_K} * u_i. \tag{5}$$

Thus, instead of the components of the model being additive, they have become multiplicative. This changes the distribution of the errors from $\varepsilon_i \sim N(0, \sigma^2)$ to $u_i \sim Lognormal(0, \sigma^2)$.This also implies that the IVs are no longer linearly related to the DV. Given the widespread usage of data transformations, using a logarithmic transformation will be examined to illustrate its efficacy for dealing with non-normality and its effects on CI properties.

Bootstrap approaches

Another common approach which can address violations of normality is to construct CIs using bootstrapping, which is a technique which avoids invoking assumptions about the distribution of the regression errors, $\boldsymbol{\varepsilon}$. When residual variance is not constant, using *case-resampling* is appropriate (Kutner, Nachtsheim, & Neter, 2004; Fox, 2002). This form of bootstrapping within the GLM resamples entire cases (sets of IVs and DV together). If there are *N* cases in the original sample, then each bootstrap sample will consist of *N* randomly sampled with replacement cases from the original sample. This resampling procedure is repeated until a predetermined number of bootstrap samples is produced. For each of these bootstrap samples, new estimates of regression coefficients are calculated, $\hat{\beta}_k$. If confidence intervals are the desired outcome from the bootstrap, then 1000 or more bootstrap samples are recommended (Fox, 2002). Like most statistical methods which use sample data to generalize about a population of interest, the bootstrap technique has an underlying assumption that the observed sample is a good representation of the actual population. To accommodate for this assumption, larger sample sizes are recommended (in a similar fashion to the CLT).

The percentile bootstrap CI is one of the most common types of bootstraps used, and it is formed by first rank ordering these bootstrapped estimates (e.g., $\hat{\beta}_k$) from smallest to largest. Next, a bootstrap percentile interval can be constructed by taking specific quantiles from these rank-ordered estimates. For instance, if *M* = 1000 bootstrap samples are created and a 95% bootstrap percentile interval is being

constructed, then the lower bound of the interval is simply the 25th ordered statistic and the upper bound is the 975th ordered statistic.

Another bootstrap variant of note is the bias-corrected and accelerated percentile bootstrap (BC_a CI; Efron, 1987). This variant accounts for skewness in the percentile CI. BC_a CIs are constructed similarly to percentile bootstrap CIs but with two adjustments. The first is a correction constant to adjust the CI for skewness. An acceleration parameter is also estimated which further adjusts the endpoints of the CI to account the fact that the distribution might change in shape or skewness at different levels of the statistic being estimated (Efron & Tibshirani, 1986; Efron, 1987; DiCiccio & Efron, 1996; Carpenter & Bithell, 2000). Progressing from percentile CIs to BC_a CIs requires less restrictive assumptions at the cost of greater computation, although the process itself can be carried out algorithmically without requiring researchers to formally calculate the parameters (Efron & Tibshirani, 1986).

Example

To illustrate each of the three approaches described above for addressing non-normality and highlight the motivation for this study, an empirical example is presented using a subset of the 1971 Canadian census focusing on occupational prestige, average income, years of education, and proportion of women within each occupation. The dataset contains 102 occupations and can be found within the *car* package for R (Fox & Weisberg, 2011). A GLM was fit to a sample of size $N = 30$ occupations regressing the average income measured in dollars on the average number of years of education for each occupation, the percentage of each occupation who are women, and the Pineo-Porter prestige score for each occupation. Each of these IVs was mean centered prior to fitting the model. Table 1 summarizes the confidence intervals around each regression coefficient using the four approaches described above, and it is readily apparent that each approach yields different 95% CIs, the properties (coverage and accuracy) of which remain unknown. Although efficiency can be calculated directly (i.e., the width of each CI), it is a poor metric to arbitrate among competing CIs techniques when the long-term coverage remains unknown. The rationale against using the most efficient CI among competing methods is that the efficiency of each

method may be simply a function of the data properties of the sample collected. Using a CI method which maintains the nominal coverage over repeated sampling should be the priority of a researcher. The question remains, which approach to addressing non-normality when constructing CIs around effect sizes should an applied researcher choose?

The purpose of this example is two-fold. The first goal is to draw attention to the difference in results based on the different approaches used. The focus here is to determine which approach to addressing non-normality is most appropriate in terms of CI properties. Given that the true coverage remains unknown, it is impossible to determine which (if any) of these CIs contain the true population regression coefficient, or which CI maintains the nominal rate of coverage if repeated sampling were possible. However, if certain data analytic conditions are known to be present, perhaps one method for constructing a CI outperforms the other methods and consequently be a researcher's best way to optimize CI properties. To help answer the question of which approach to utilize, this study presents a Monte Carlo simulation to evaluate the properties of the particular CI methods discussed earlier.

Table 1. Confidence intervals for empirical example

Confidence Interval Method	IV	CI Lower Limit	CI Upper Limit
Standard	Education	-259.49	699.95
Perc	Education	-213.10	561.17
BCa	Education	-186.02	574.92
Standard	Prestige	18.13	176.74
Perc	Prestige	16.59	181.33
BCa	Prestige	16.62	181.47
Standard	Women	-61.57	-21.22

Perc	Women	-54.01	-26.93
BCa	Women	-56.71	-29.38

Note. BCa = Bias corrected and accelerated bootstrap, Perc = Percentile bootstrap.

Methods

A Monte Carlo simulation was conducted to examine the relative efficacy of a variety of approaches to dealing with violations of normality of errors on the performance of CIs within the GLM framework. A fully crossed four-factor design was implemented examining the effects of sample size, number of IVs, degree of association among IVs, and non-normally distributed errors. The levels of each factor were varied to reflect data analytic scenarios commonly encountered in applied settings. The simulation was conducted using *R* (R Core Team, 2016) with the *SimDesign* package to organize the results (Chalmers, 2017; Sigal & Chalmers, 2016).

Sample size

The first factor is sample size. The levels of this factor were N = 10, 30, 50, 100, and 1000. The level of N = 30 was included to test the general rule of thumb that the CLT starts to correct non-normality at this sample size. As sample size increases, it is expected that coverage converges to the nominal rate of 95% using $\alpha = .05$, and the accuracy of the upper and lower boundaries of the CI converges to $\alpha/2$. The efficiency of confidence intervals improves due to the direct effect an increase in sample size has on estimates of the standard error for the regression coefficients.

Continuous independent variables

The second factor was the number of continuous IVs in the model. The number of multivariate normally distributed continuous IVs in each model varied from K = 2, 3, 4, 5, and 6. The IVs were generated using the *mvrnorm* function within the *R* package *MASS* (Venables & Ripley, 2002) using pre-defined covariance matrices specific to each condition in the simulation and each IV was mean-centered.

With everything else held constant, it was expected that as the number of IVs increased, the CI coverage would decrease. The rationale was that including more IVs would reduce the residual variance in the model and shrink the standard errors of the regression coefficients. This results in CIs with smaller ranges which are less likely to capture the population parameter of interest. Accuracy was likewise expected to deviate from the set rate of α/2 per tail.

Association among independent variables

The third factor was the degree of association among the continuous IVs. The degree of association among the IVs in each model was controlled by varying the value of the condition number, κ, which is the square root of the ratio between the largest and smallest eigenvalues produced from the population correlation matrix of the IVs. Following Dudgeon's (2017) simulation, values of κ were set at 3, 6, and 9 to create population correlation matrices for each of the different conditions within the simulation. These condition numbers were used to specify eigenvalues and the *genCorr* function from the *R* package *fungible* was used to create the population correlation matrices (Waller & Jones, 2016) for each model. A higher degree of association among the IVs was expected to increase the standard errors of the regression estimates and result in decreased efficiency, lower coverage, and adversely affect accuracy.

Error distributions

The fourth and final factor was the population error distributions. The distributions examined followed Dudgeon's (2017) simulation which used three levels for this factor: normal, contaminated-normal, and highly kurtotic. Using normally distributed errors with a mean of 0 and variance = 1, $\varepsilon \sim N(0,1)$, shows how a model's estimates and associated inferential statistics behave with no assumption violations and serves as a basis for comparison with the remaining levels of this factor. The contaminated-normal distribution was created by sampling from a mixture distribution defined as $\varepsilon = 0.9W_1 + 0.1W_2$, where the distributions of W_1 and W_2 differ only in terms of their variances, $W_1 \sim N(0, \sigma_1^2)$, $W_2 \sim N(0, \sigma_2^2)$, with $\sigma_1^2 = 0.09$, and $\sigma_2^2 = 100 * \sigma_1^2$. Taken together, the errors at this level are distributed

with a mean of 0 and variance of 1, with skewness around 0 and kurtosis of about 25.73. This contaminated-normal distribution samples more heavily from the tails of the distribution relative to the unit-normal distribution. The next distribution was even more highly kurtotic, and was simulated using a moment matching method for generating non-normal data as described in Fleishman (1978) using the *rValeMaurelli* function within the *SimDesign* package (Chalmers, 2016). This distribution has a mean of 0 and variance of 1, but with an approximate skewness of 0 and a kurtosis of 100. The high kurtosis of this distribution results in more errors clustering around the mean of 0 than the normal distribution. In general, as the errors deviate further from normality, the OLS estimates for the regression coefficients may no longer be the best linear unbiased estimates compared to other estimators.

Data generation

The fully crossed four-factor design of this simulation yielded 225 unique conditions to be examined. Using each condition's degree of association and number of IVs, a unique correlation matrix was constructed using Marsaglia and Olkin's (1984) method. Each of these condition-specific correlation matrices was rescaled using preset values for the population standard deviations of each IV. The resulting covariance matrix was used to simulate a $N*K$ matrix of sample data, where N was the condition's sample size and K was the number of IVs for that condition. This data matrix remained identical across all R = 1000 replications of each condition. The only difference across each condition's replications was the random sample of errors.

Log transformation

To assess the effect of performing a logarithmic data transformation on the DV in terms of CI properties, a separate unique condition was also simulated to mimic the way in which applied researchers typically approach the model building process. In practice, researchers will often regress a DV onto a set of IVs and use regression diagnostics (e.g., Fox, 1991) to determine whether the assumption of normality (among others) was violated. If a violation was detected, the researcher can apply an appropriate data

transformation or decide that the violation was not of sufficient consequence and the robustness of the GLM should minimize the impact of the violation. If the GLM was assumed to be robust to the violation of normality, then a researcher could fit the model without transforming the DV. However, if diagnostics indicated that the error distribution was asymmetric in the presence of a positively skewed DV, then the logarithm of the DV could be regressed onto the IVs instead. To illustrate how CI properties are affected by a decision to log-transform the DV, both models (using log-transformed and untransformed DVs) were fit in a unique condition which is described below.

This separate unique condition used a sample size of N = 1000, a condition value of 3 (degree of association), three IVs, and normally distributed errors to represent the ideal situation for fitting a GLM. The multiplicative model expressed in Equation 4 was used to generate the untransformed DV, y_i. The multiplicative model was used to ensure that the logarithm of y_i was linearly related to the three IVs and that the errors would be distributed normally. The linear relationship between the logarithm of y_i, the three IVs, and the errors can be seen in Equation 6, which can be found by taking the logarithm of both sides of Equation 4:

$$log(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \quad . \tag{6}$$

This equation represents the "true" model in which the IVs are linearly related to the logarithm of y_i with normally distributed errors. To fit data to the "incorrect" (or misspecified) model in which the IVs are not linearly related to the DV, y_i was regressed onto the exact same three IVs used to originally generate y_i from the multiplicative model. The form of the "incorrect" model is in Equation 7 which uses y_i, x_1, x_2, and x_3 to find estimates for the regression coefficients and residuals:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{\varepsilon}_i \tag{7}$$

This "incorrect "model will have non-normally distributed errors because the generation process for y_i ensures that the IVs are not linearly related to the DV. The properties of the CIs around the regression coefficients from both models are calculated similar to the main simulation as described above.

Results

To compare the performance of CI properties (coverage, efficiency, and accuracy) across all 225 unique design conditions, average CI properties for each regression parameter estimate of effect size were computed. For instance, in design conditions with three IVs, all three coverages, efficiencies, and tail proportions were averaged together for each of the three methods for constructing CIs (standard method, percentile bootstrap, and the BCa bootstrap). Both bootstrap methods used M = 1000 bootstrap samples to construct CIs. The first CI property examined was coverage. In general, coverage for the standard method performed well across all conditions (see Figure 1). Coverage for the percentile bootstrap method was close to the nominal 95%, but was slightly lower than the standard method. The BCa bootstrap method had the most problems, especially for conditions in which residuals were sampled from non-normal error distributions. A plot depicting the average coverage per condition is presented in Figure 1. The plot is divided horizontally into thirds based on the method used to construct the confidence intervals. The first section used the standard method for calculating confidence intervals, the second section used the percentile bootstrap method, while the third section used the BCa bootstrap method to construct the confidence intervals. A blue confidence region was superimposed on the plot to highlight which average coverages are within two standard errors of the nominal 0.95 value, with the standard error of the percentage of coverage defined as

$$SE(\pi) = \sqrt{\frac{0.95*(1-0.95)}{R}} \tag{5}$$

where R = 1000 was the number of replications per condition. Well-behaved average coverages should fall within this 95% confidence region.

Figure 1 shows spikes in many conditions across both bootstrap methods in which average coverage approached 100%. To determine which factor in the design corresponded to these spikes, averages of all four CI properties were calculated for each level of the four factors and summarized in Tables 2 to 5. The N = 10 level of the sample size factor was responsible for the spikes in average coverage. The average

efficiency for this level of sample size was quite large relative to every other level of this factor. These larger CI widths reflect the expected imprecision around the estimates of effect size for small sample sizes, and it was these larger widths that increased the average coverage of these confidence intervals.

Figure 2 shows a plot of average efficiency constructed like the plot for average coverage. Overall, all three methods for constructing CIs performed quite similarly. The marginal means of average efficiency are in Table 2. Two additional plots, Figures 3 and 4, showing the average lower and upper tail proportions of the CIs, did not have any easily visible trends. This result is an indication that, on average, each method for constructing CIs was centered around the true parameter value. Both Figure 3 and 4 use a 95% confidence region as with previous plots, except that these two plots have a 95% confidence region centered around the expected tail proportions.

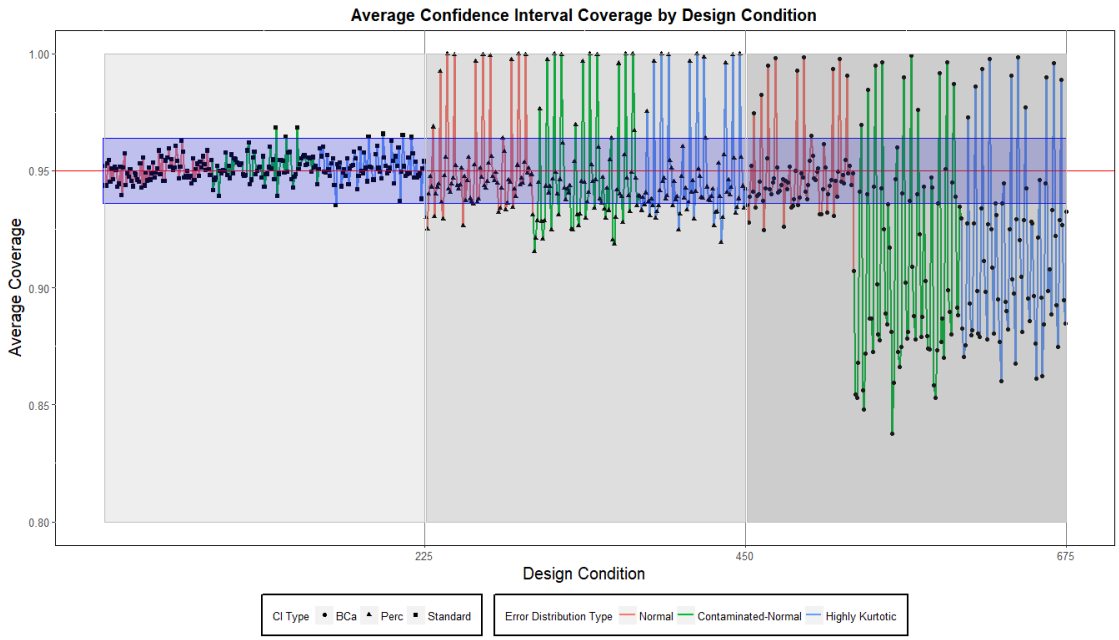

Figure 1. Average Confidence Interval Coverages

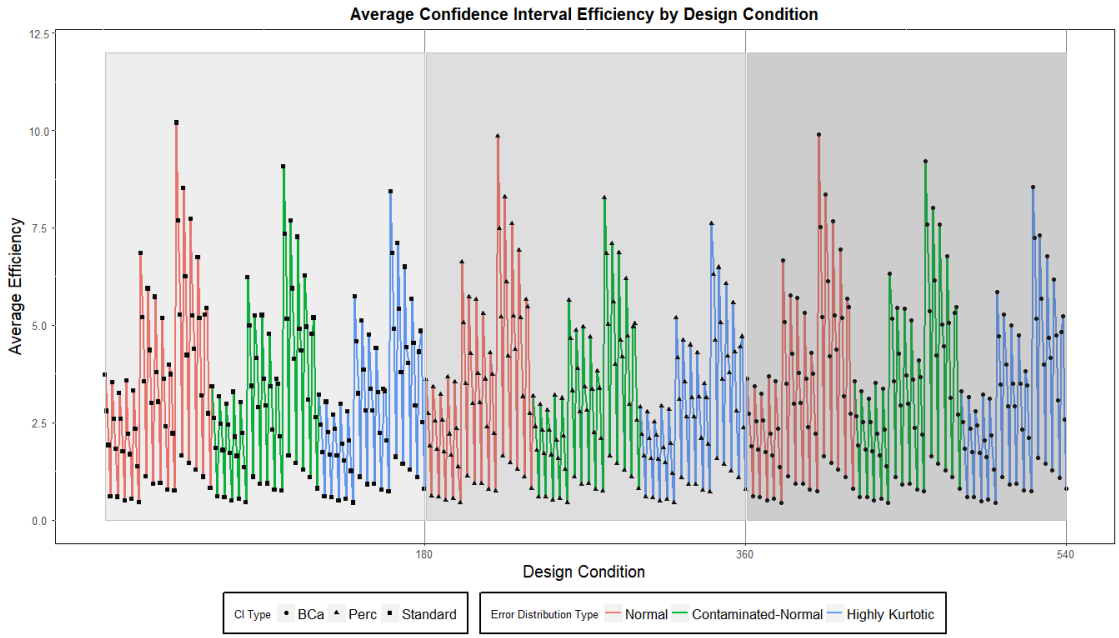

Figure 2. Average Confidence Interval Efficiencies

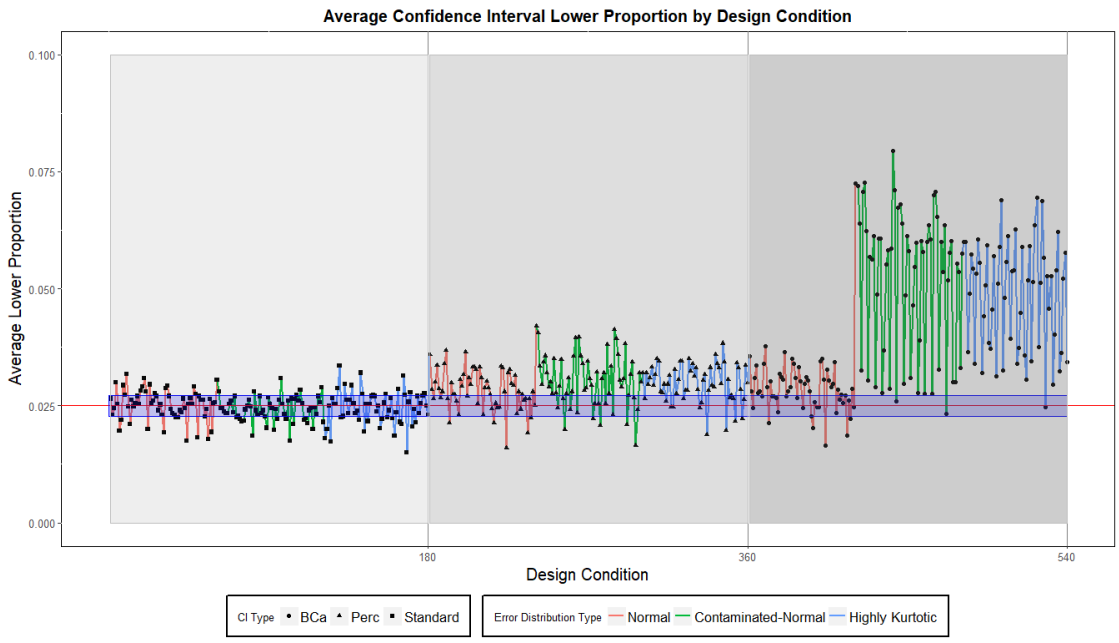

Figure 3. Average Confidence Interval Lower Proportions

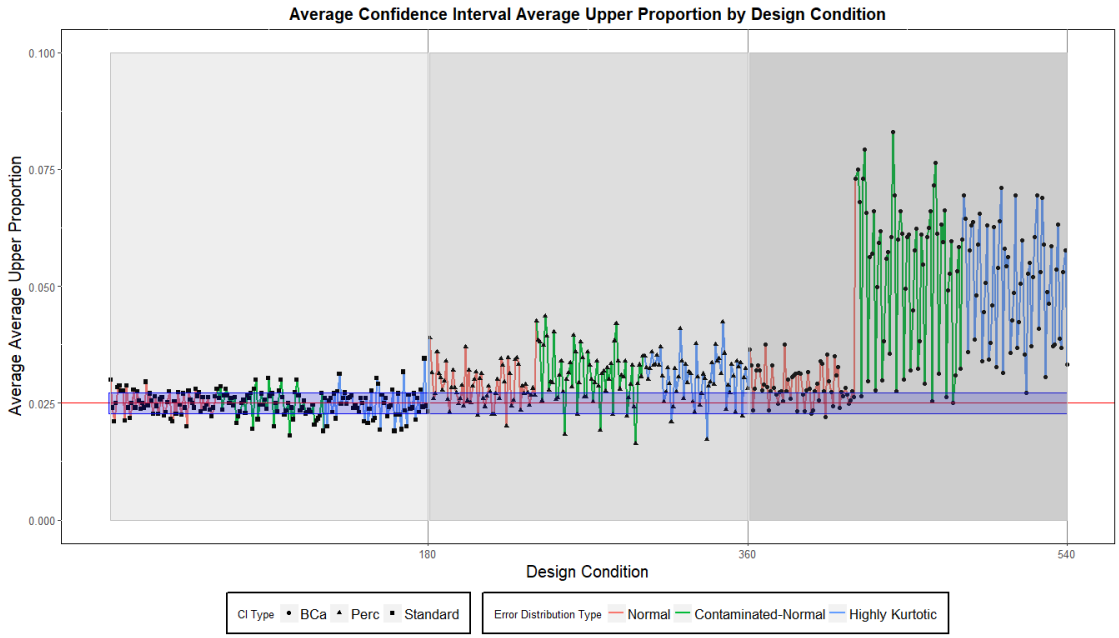

Figure 4. Average Confidence Interval Upper Proportions

CI Properties by Factor

Sample Size

The marginal means of CI properties by sample size are in Table 2. The sample size factor did not exhibit the expected results for all levels. At the smallest sample size, N = 10, both bootstrap CIs had coverage levels exceeding the nominal 95% and fell outside of the 95% confidence region. The percentile bootstrap had the highest average coverage (M = 0.98, SD = 0.03), followed by the BCa bootstrap (M = 0.97, SD = 0.03), and then the standard confidence interval (M = 0.95, SD = 0.01). Furthermore, the efficiencies of all three CIs at the N = 10 level were substantially worse than every other level. This effect was anticipated because of the instability of estimates at low sample sizes. In essence, the width of these intervals indicates very imprecise interval estimates for the regression coefficients and inflated the coverage to exceed the nominal level. For this reason, the N = 10 condition was removed prior to investigating the effects of the other factors of the simulation design.

Next, at the highest level of sample size, N = 1000, the coverage and efficiency were similar for all three types of CIs, with each around the 95% benchmark. The coverage of the three remaining levels of sample size (N = 30, 50, and 100) remained quite stable, and these sample sizes did exhibit the improvement in efficiency as expected. As sample size increased, CI width decreased. In each of these middle three levels of sample size, the standard method achieved the 95% coverage, while the coverage for the percentile bootstrap CI consistently had 94% coverage. The BCa method performed the worst of the three methods and had a coverage of approximately 90% across all three levels.

Table 2. Confidence Interval Properties by Sample Size

		Average Coverage		Average Efficiency		Average Lower Proportion		Average Upper Proportion	
Sample Size	Confidence Interval Type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
10	BCa	0.97	0.03	91.61	107.07	0.014	0.014	0.014	0.014
10	Perc	0.98	0.03	40.88	36.51	0.010	0.013	0.010	0.012
10	Standard	0.95	0.01	13.02	5.86	0.023	0.004	0.023	0.004
30	BCa	0.91	0.03	5.29	1.87	0.044	0.013	0.044	0.014
30	Perc	0.94	0.01	4.94	1.79	0.029	0.006	0.029	0.006
30	Standard	0.95	0.01	5.10	1.92	0.025	0.003	0.024	0.003
50	BCa	0.90	0.03	4.06	1.53	0.049	0.016	0.050	0.016
50	Perc	0.94	0.01	3.80	1.44	0.030	0.005	0.032	0.005
50	Standard	0.95	0.01	3.96	1.50	0.025	0.004	0.026	0.003
100	BCa	0.90	0.03	2.79	1.10	0.050	0.016	0.051	0.017
100	Perc	0.94	0.01	2.65	1.05	0.031	0.004	0.032	0.004
100	Standard	0.95	<0.01	2.73	1.08	0.025	0.002	0.025	0.002
1000	BCa	0.94	0.01	0.89	0.35	0.030	0.005	0.030	0.005
1000	Perc	0.95	0.01	0.89	0.35	0.027	0.004	0.027	0.004
1000	Standard	0.95	<0.01	0.89	0.35	0.024	0.003	0.024	0.002

Note. BCa = Bias corrected and accelerated bootstrap, Perc = Percentile bootstrap.

Error Distribution

A full summary of the marginal statistics for CI properties by error distribution is in Table 3. When model errors were drawn from a normal distribution, the coverage of all three CIs behaved as

expected. The standard method (*M* = 0.95, *SD* = 0.01) performed slightly better than the bootstrap methods (BCa *M* = 0.94, *SD* = 0.01; percentile *M* = 0.94, *SD* = 0.01). For models with errors drawn from the contaminated-normal distribution, coverage for the standard method performed best (*M* = 0.95, *SD* = 0.01), followed by the percentile bootstrap (*M* = 0.94, *SD* = 0.01), and the BCa performed the worst (*M* = 0.90, *SD* = 0.03). For models with errors drawn from the highly kurtotic distribution, coverage for the standard method was largest (*M* = 0.95, *SD* = 0.01), followed by the percentile bootstrap (*M* = 0.94, *SD* = 0.01), and the BCa, once again, performed the worst (*M* = 0.90, *SD* = 0.02). All three error distributions exhibited the same pattern regarding the rank order of each method's efficiencies. The percentile bootstrap systematically had the worst level of efficiency, followed by the standard method, and then the BCa bootstrap. The normal, contaminated-normal, and the highly kurtotic distribution levels all had similar efficiencies.

Table 3. Confidence Interval Properties by Error Distribution

		Average Coverage		Average Efficiency		Average Lower Proportion		Average Upper Proportion	
Error Distribution Type	Confidence Interval Type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Normal	BCa	0.94	0.01	3.37	2.22	0.028	0.004	0.029	0.004
Normal	Perc	0.94	0.01	3.36	2.21	0.028	0.004	0.029	0.004
Normal	Standard	0.95	<0.01	3.38	2.25	0.025	0.003	0.025	0.002
Contaminated-Normal	BCa	0.90	0.03	3.30	2.14	0.052	0.016	0.053	0.016
Contaminated-Normal	Perc	0.94	0.01	3.04	1.91	0.030	0.006	0.031	0.006
Contaminated-Normal	Standard	0.95	<0.01	3.18	2.05	0.024	0.003	0.025	0.003
Highly Kurtotic	BCa	0.90	0.02	3.11	1.97	0.049	0.012	0.050	0.012
Highly Kurtotic	Perc	0.94	0.01	2.80	1.73	0.030	0.004	0.031	0.005
Highly Kurtotic	Standard	0.95	0.01	2.96	1.87	0.025	0.004	0.025	0.003

Note. BCa = Bias corrected and accelerated bootstrap, Perc = Percentile bootstrap.

Degree of association

The average coverage for each type of CI was unchanged across all three levels of the degree of association among the IVs (κ = 3, 6, and 9). Coverage for the standard method was highest (*M* = 0.95, *SD* = 0.01), followed by the percentile bootstrap (*M* = 0.94, *SD* = 0.01), and then by the BCa bootstrap (*M* = 0.91, *SD* = 0.03). As expected, as the degree of association among IVs increased, the efficiency of each method increased resulting in greater imprecision for their respective CIs. A full summary is in Table 4.

Table 4. Confidence Interval Properties by the Degree of Association Among Independent Variables

		Average Coverage		Average Efficiency		Average Lower Proportion		Average Upper Proportion	
κ	Confidence Interval Type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
3	BCa	0.91	0.03	1.99	1.05	0.044	0.015	0.045	0.017
3	Perc	0.94	0.01	1.87	0.98	0.030	0.004	0.031	0.005
3	Standard	0.95	<0.01	1.93	1.01	0.025	0.003	0.025	0.002
6	BCa	0.91	0.03	3.24	1.71	0.043	0.016	0.043	0.016
6	Perc	0.94	0.01	3.05	1.59	0.030	0.005	0.029	0.005
6	Standard	0.95	<0.01	3.15	1.67	0.025	0.003	0.025	0.003
9	BCa	0.91	0.03	4.55	2.46	0.042	0.016	0.043	0.016
9	Perc	0.94	0.01	4.28	2.29	0.029	0.006	0.030	0.005
9	Standard	0.95	0.01	4.43	2.42	0.024	0.003	0.024	0.003

Note. κ = Degree of association among independent variables, BCa = Bias corrected and accelerated bootstrap, Perc = Percentile bootstrap.

Number of Independent Variables

The results for the factor pertaining to the number of IVs are summarized in Table 5. The average coverage for the standard method CIs (*M* = 0.95, *SD* < 0.01) was stable across all levels of this factor (*K* = 2, 3, 4, 5, and 6). The average coverage for percentile bootstrap CIs remained relatively stable (*M* = 0.94, *SD* = 0.01) for all conditions except when the number of IVs was six. With six IVs, average coverage increased to *M* = 0.95. The average coverage for BCa bootstrap CIs was lower for each level of this factor than both of the other CI methods. The average BCa coverage remained close to 0.92 with a standard

deviation ranging from 0.02 to 0.03. There was a decreasing trend in average efficiencies across all levels of this factor, such that as the number of IVs in the model increases, the average efficiency decreases as does the stability of these means (which can be seen via the standard deviations).

Table 5. Confidence Interval Properties by the Number of Independent Variables

		Average Coverage		Average Efficiency		Average Lower Proportion		Average Upper Proportion	
Number of Independent Variables	Confidence Interval Type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
2	BCa	0.90	0.04	4.00	2.62	0.047	0.018	0.048	0.019
2	Perc	0.94	0.01	3.74	2.43	0.031	0.005	0.032	0.006
2	Standard	0.95	0.01	3.96	2.61	0.024	0.004	0.024	0.003
3	BCa	0.91	0.03	3.43	2.16	0.046	0.017	0.047	0.018
3	Perc	0.94	0.01	3.21	2.00	0.032	0.004	0.032	0.005
3	Standard	0.95	<0.01	3.38	2.13	0.025	0.003	0.025	0.003
4	BCa	0.91	0.03	3.22	1.98	0.043	0.014	0.044	0.015
4	Perc	0.94	0.01	3.04	1.85	0.030	0.004	0.030	0.005
4	Standard	0.95	0.01	3.16	1.93	0.025	0.003	0.025	0.003
5	BCa	0.92	0.03	2.95	1.85	0.041	0.014	0.041	0.014
5	Perc	0.94	0.01	2.80	1.74	0.028	0.004	0.029	0.003
5	Standard	0.95	<0.01	2.84	1.74	0.024	0.003	0.025	0.002
6	BCa	0.92	0.03	2.68	1.66	0.038	0.013	0.039	0.013
6	Perc	0.95	0.01	2.55	1.57	0.027	0.004	0.027	0.005
6	Standard	0.95	<0.01	2.53	1.52	0.025	0.003	0.025	0.003

Note. BCa = Bias corrected and accelerated bootstrap, Perc = Percentile bootstrap.

Relative performance of confidence interval properties

To compare the performances of each CI method within the remaining 180 design conditions after removing the 45 conditions with a sample size of $N = 10$, a subset of only the CI methods which had an average coverage inside of the 95% confidence region constructed about the 95% confidence limits was analyzed. Coverage is arguably the most important of the three CI properties under consideration because

inference using CIs can be misleading without proper coverage. Using this subset of CI methods with proper coverage within each design condition, the CI methods with the best and worst efficiency (i.e. smallest and largest CI width) were cross-tabulated to broadly consider CI performance. The full results are in Table 6. For instance, in 93 of the 180 design conditions, the percentile bootstrap method had the best efficiency compared to all other methods. In 89 conditions, the percentile bootstrap method had the best efficiency while the standard method had the worst, and in three conditions the percentile bootstrap method was best while the BCa bootstrap method was the worst for these five conditions. There was one design condition in which only the percentile bootstrap method had proper coverage and by default had the best efficiency. The standard method had 51 conditions in which it was the only method that maintained proper coverage, and there were an additional 33 cases in which the standard method outperformed the all other methods in terms of efficiency. There were two conditions in which the BCa had proper coverage and was the most efficient of the three methods. Lastly, there was one design condition in which no method had proper coverage.

Despite using R = 1000 replications per condition, in many instances the efficiency of one method was not much different than another method. Using only methods that maintained proper coverage, a full breakdown of the 180 design conditions showcasing the methods with the best relative efficiency is presented in Table 6. There were three general outcomes. The first was that in many conditions only a single method maintained proper coverage, and consequently would be the optimal method. The second outcome was that multiple methods had proper coverage and comparable efficiencies, such that was no singularly optimal method. The third outcome was when multiple methods had proper coverage, but the width of one method was at least 5% smaller than the width of the next smallest CI width.

Table 6. Best Efficiency per Condition with Proper Coverage.

			Number of Independent Variables				
Error Distribution	Sample Size	κ	2	3	4	5	6
Normal	30	3	S	S	-	-	s
Normal	30	6	S	-	-	-	s
Normal	30	9	S	S	S	-	s
Normal	50	3	-	-	S	-	-
Normal	50	6	-	-	-	-	-
Normal	50	9	S	-	-	-	-
Normal	100	3	-	-	-	-	-
Normal	100	6	-	-	-	-	-
Normal	100	9	-	-	-	-	-
Normal	1000	3	-	-	-	-	-
Normal	1000	6	-	-	-	-	-
Normal	1000	9	-	-	-	-	-
Contaminated-Normal	30	3	S	S	-	-	-
Contaminated-Normal	30	6	p	S	p	-	-
Contaminated-Normal	30	9	S	S	p	-	s
Contaminated-Normal	50	3	S	S	p	-	-
Contaminated-Normal	50	6	S	S	S	-	-
Contaminated-Normal	50	9	P	S	p	-	-
Contaminated-Normal	100	3	S	S	S	S	-
Contaminated-Normal	100	6	S	S	-	S	S
Contaminated-Normal	100	9	S	S	S	S	S
Contaminated-Normal	1000	3	-	-	-	-	-
Contaminated-Normal	1000	6	-	-	-	-	-
Contaminated-Normal	1000	9	-	-	-	-	-

Highly Kurtotic	30	3	S	p	S	-	-
Highly Kurtotic	30	6	p	p	p	-	s
Highly Kurtotic	30	9	p	p	p	-	-
Highly Kurtotic	50	3	p	S	-	p	-
Highly Kurtotic	50	6	S	p	p	-	p
Highly Kurtotic	50	9	S	S	p	S	-
Highly Kurtotic	100	3	S	p	S	-	S
Highly Kurtotic	100	6	S	S	S	p	-
Highly Kurtotic	100	9	S	S	p	S	S
Highly Kurtotic	1000	3	-	S	-	-	-
Highly Kurtotic	1000	6	-	-	-	-	-
Highly Kurtotic	1000	9	S	-	-	-	-

Note. An S indicates that only the standard CI maintained proper coverage. A P indicates that only the percentile CI maintained proper coverage. Lowercase letters indicate that their corresponding CI method had an efficiency at least 5% better than the next best method.

Logarithmic transformation

The unique condition comparing the two models which used transformed and untransformed DVs yielded the expected results. When the untransformed DV was regressed on the three IVs, the average coverage suffered greatly (M = 0.16) compared to the log-transformed DV (M = 0.94), after R = 1000 replications. The average proportions for the upper and lower tails for the model fitted to the untransformed DV were M = 0.00 and M = 0.84, respectively, whereas the average proportions for the upper and lower tails for the correct, linear model were M = 0.028 and M = 0.029, respectively. Given the vastly different scales for each of these two models, a comparison of their average efficiencies is not meaningful and will not be reported.

Discussion

Given the renewed concerns regarding the dependability of psychological research findings and the call for researchers to report effect sizes and their CIs, the present study sought to further inform researchers about how best to approach violations of non-normality for CI construction within the GLM.

Using a Monte Carlo simulation, the impact of common data analytic conditions was assessed regarding three important CI properties: coverage, accuracy, and efficiency.

In terms of overall performance, a few facets of the simulation results need highlighting. First, if errors are distributed similarly to the contaminated-normal distribution used here, then at lower sample sizes the percentile bootstrap surpasses the standard method in terms of efficiency, given that they have similar coverage. It is recommended that the percentile bootstrap be routinely applied. However, when sample size is $N = 50$ or greater, the standard method starts to systematically outperform the percentile bootstrap method across all levels of the other simulation factors. As a reminder, the high kurtosis in the simulation's contaminated-normal error distribution is not only indicative of the "peakedness" of the distribution, it also represents the heavy tails of the distribution. The relevance of this reminder is that if a distribution is heavy tailed, a bootstrap method (i.e., the percentile bootstrap) might be a superior choice in terms of efficiency. It was expected that the standard CI would suffer a loss in efficiency when using OLS estimation, as was stated earlier and supported by Fox (1991).

When errors are distributed with extremely high kurtosis, the percentile bootstrap method should be employed as it outperforms the standard and BCa bootstrap methods more often, especially for sample sizes close to $N = 30$ or 50. When sample sizes approach 100, the standard method takes over as the optimal method for constructing CIs. This recommendation is similar to advice presented above regarding the contaminated-normal distribution, but it is important to note that that even with the more extreme kurtosis the recommendation remains about the same.

Another general recommendation is that when the residual distribution is consistent with a normal distribution for the underlying errors, the standard method performs similarly or better than the two bootstrap methods. Thus, regardless of the degree of association among the IVs, number of IVs, or sample size, the standard method was either the only method with nominal coverage or had nominal coverage and efficiency comparable with any other method of CI construction. In both cases, the standard method was most likely to have optimal CI properties.

The effect of sample size was as expected. First, a small sample size (N = 10) led to large standard errors and thus each method's CIs had higher coverage than the nominal 95% limit, and efficiency up to thirty times larger than CIs constructed using larger sample sizes. Both bootstrap methods were more adversely affected by low sample size than the standard method. It would be hard to justify that any sample of size N = 10 is a sufficient approximation for a normal distribution; there simply is not enough information to ascertain the veracity of this assumption. Second, the large-sample properties of each method showed the expected convergence of coverage to the nominal 95% with CI widths nearly identical to each other. This result simply means that with sufficiently large sample sizes (approximately N = 1000) the choice between CI method is arbitrary and researchers are recommended to default to the computationally lighter standard method. Lastly, coverage tends to reach the 95% confidence limit around N = 30 for all methods except the BCa bootstrap. The BCa bootstrap still does not reach proper coverage at N = 100, though as stated above, proper coverage for this method is achieved with samples of N = 1000.

The unique condition comparing the two models which used log-transformed and untransformed DVs indicated that a transformation could improve CI properties. However, data transformations should be applied thoughtfully to ensure that they achieve the desired effect of normalizing the distribution of the model residuals. There are many possible transformation methods, and, as pointed out by Feng et al. (2014), a logarithmic transformation in no way guarantees that model residuals are normalized after a transformation. Again, researchers are reminded that using a log-transformed DV within the GLM is tantamount to selecting a multiplicative model once back-transformed to the original scale of the DV. The question of whether to transform a DV is a matter of balancing how well the transformation normalizes residuals, whether the transformation restores linearity, and the ease of interpretation of the regression coefficients.

An empirical example was provided earlier to illustrate how competing approaches to dealing with non-normality result in different CIs whose properties remain largely unknown. Using the

recommendations summarized in Table 6, the decision about which CI method to choose can be made after first determining the data properties of the empirical example. The example used three IVs, had a condition value of 4.20 (degree of association among the IVs), and a sample size of $N = 30$. By superimposing the different error distributions examined in the simulation, the density of the empirical example's residuals most closely resembles either contaminated-normal error distribution. The simulation condition in Table 6 which provides the closest approximation to the data properties of the empirical example suggests that the percentile bootstrap method should outperform the other CI methods by achieving nominal coverage and exhibiting at least a 5% improvement in efficiency. While this suggestion to choose the percentile bootstrap method for constructing CIs is completely data-driven, the simulation results suggest that, on average, this method should be preferred for the current situation. Given that all inference requires varying degrees of uncertainty, the current findings can help make informed decisions regarding how to address non-normality, and even slight adjustments such as those suggested in this present study could improve the dependability of research findings.

Overall, small patterns were discernable from the simulation results which can give researchers support regarding decisions among competing approaches to non-normality. From the present results, it was clear that no single approach for dealing with non-normality was optimal across the board. By identifying conditions in which a non-standard method outperformed the standard method, researchers are encouraged to explore alternate approaches to deal with non-normal errors to improve the properties of their desired CIs. This finding alone is a useful contribution given the widespread belief that the GLM is robust to assumptions of violations of normality. Simply ignoring the assumption violation can result in suboptimal CI performance when the sample size is approximately $N = 50$ or less, especially when the distribution of errors deviates strongly from a normal distribution. When applied researchers are faced with the inevitable situation of violations of normality, they might not rely solely on methods with which they are familiar. Granted considerations based on known statistical theory outweigh recommendations based on simulated results, such theoretical considerations are usually based on large sample properties

and do not generalize to every applied situation, particularly with the small sample sizes typically found within psychological research. Often, applied researchers have no theoretical basis for deciding how to deal with assumption violations beyond the choice of modelling framework. Researchers should be aware of the impact that deviations from normality have on CI properties and expand their statistical toolbox to include additional methods to help achieve optimal CI performance.

References

Bartlett, M. S. (1947). The use of transformations. *Biometrics*, 3 (1), 39-52.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, *57*(1), 49.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, *26*, 211–252.

Cain, M. K., Zhang, Z., & Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, 49(5), 1716-1735.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19(9), 1141-1164.

Chalmers, P. (2017). SimDesign: Structure for organizing Monte Carlo simulation designs. *R* package version 1.6. https://CRAN.R-project.org/package=SimDesign

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Routledge.

Cumming, G. (2014). The new statistics why and how. *Psychological science*, 25(1). doi: 10.1177/0956797613504966.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3), 189-212.

Dudgeon, P. (2017). Some improvements in confidence intervals for standardized regression coefficients. *Psychometrika*, published online 13 March 2017. doi: *10.1007/s11336-017-9563-z*

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171-185.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, 1(1), 54-75.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591-601. doi: 10.1037/0003-066X.63.7.591

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, 26(2), 105-109.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika* 43(4), 521-532.

Fox, J. (1991). *Regression Diagnostics: An Introduction* (Vol. 79). Newbury Park, CA: Sage.

Fox, J. (2002). Bootstrapping Regression Models: An Appendix to an R and an S-Plus Companion to Applied Regression. Retrieved from http://cran.r-project.org/doc/contrib/Fox-companion/appendix-bootstrapping.pdf on 20 December 2016.

Fox, J., & Weisberg, S. (2011). An *R* companion to applied regression (2nd ed.). Thousand Oaks, CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237-288.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, 39(4), 709-722.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*(3), 350-386.

Kutner, M. H., Nachtsheim, C., & Neter, J. (2004). *Applied Linear Regression Models* (4th ed. / Michael H. Kutner, Christopher J. Nachtsheim, John Neter.). Boston: McGraw-Hill/Irwin.

Marsaglia, G. & Olkin, I. (1984). Generating correlation matrices. *Society for Industrial and Applied Mathematics Journal on Scientific and Statistical Computing*, 5(2), 470-475.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166. doi: 10.1037/0033-2909.105.1.156

Moore, D. S., McCabe, G. P., & Craig, B. A. (2014). *Introduction to the Practice of Statistics* (8th ed.). New York, NY: W. H. Freeman.

Nickerson, R. S. (2000). Null hypothesis significance testing: A Review of an old and continuing controversy, *Psychological Methods*, 5(2), 241-301.

Pek, J., Wong, A. C. M., & Wong, O. C. Y. (2017). Confidence intervals for the mean of non-normal distribution: transform or not to transform. *Open Journal of Statistics*, 7(3). doi: 10.4236/ojs.2017.73029

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

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with Monte Carlo simulation. *Journal of Statistics Education*, *24*(3), 136-156.

Stangor, C., & Lemay, E. P. (2016). Introduction to the special issue on methodological rigor and replicability. *Journal of Experimental Social Psychology*, 66, 1-3.

Waller, N. G. & Jones, J. A. (2016). fungible: Fungible coefficients and Monte Carlo functions. R package version 1.5.

Wilkinson, L., & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54 (8), 594-604.

Vazire, S. (2015). Editorial. *Social Psychological and Personality Science*, 7(1), 3-7.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0