UNIVERSITY of PENNSYLVANIA

Department of Criminology

Working Paper No. 2015-13.01

Calibrated Percentile Double Bootstrap for Robust Linear Regression Inference

Daniel McCarthy
University of Pennsylvania

Kai Zhang
University of North Carolina at Chapel Hill

Richard Berk
Lawrence Brown
Andreas Buja
Edward George
Linda Zhao

Department of Statistics,
University of Pennsylvania

Calibrated Percentile Double Bootstrap For Robust Linear Regression Inference

Daniel McCarthy
Department of Statistics, University of Pennsylvania
and
Kai Zhang*
Department of Statistics, University of North Carolina at Chapel Hill

Richard Berk, Lawrence Brown, Andreas Buja, Edward George and Linda Zhao
Department of Statistics, University of Pennsylvania

November 3, 2015

Abstract

When the relationship between a response variable Y and covariates $\vec{\mathbf{X}}$ is nonlinear with possibly heteroskedastic noise, and where the covariates $\vec{\mathbf{X}}$ are themselves random, the empirical coverage probability of traditional confidence interval ('CI') methods decreases considerably. We propose a double bootstrap-based calibrated percentile method, `perc-cal`, as a general-purpose CI method which performs very well relative to alternative methods in challenging situations such as these. For the first time, we prove that under relatively mild regularity conditions, the rate of coverage error of `perc-cal` for a two-sided confidence interval of the best linear approximation between Y and a p-dimensional $\vec{\mathbf{X}}$ is $\mathcal{O}(n^{-2})$. We then show that `perc-cal` performs

*Kai Zhang was partially supported by NSF grant DMS-1309619. His work is also partially supported by the NSF under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

very well in practice, demonstrating its performance in a thorough, full-factorial design synthetic data study as well as a real data example involving the length of criminal sentences. We have provided an `R` package, available through CRAN and coded primarily in `C++`, to make it easier for practitioners to use `perc-cal`.

Keywords: Confidence intervals; Edgeworth expansion; Second-order correctness; Resampling

1 Introduction

Literature on bootstrap-based inference has primarily considered scenarios in which the true relationship between a response variable Y and covariates $\vec{\mathbf{X}} = (1, X_1, \ldots, X_p)^T$ is linear. We show that when the relationship between Y and $\vec{\mathbf{X}}$ may be non-linear with noise that is possibly heteroskedastic, and where $\vec{\mathbf{X}}$ is itself random, empirical coverage of population regression slopes deteriorates considerably for all traditional confidence interval methods. This is troubling because practitioners are frequently confronted with inference problems that involve data which have these features. We propose a double bootstrap-based calibrated percentile method, `perc-cal`. This method has been previously discussed when studying univariate data without model misspecification, see Hall (1992). For the first time, we prove that even when the relationship between Y and $\vec{\mathbf{X}}$ is non-linear in their joint distribution, under relatively mild regularity conditions, the rate of coverage error of `perc-cal` for two-sided confidence intervals of the best linear population slopes between a response variable Y and p-dimensional covariates $\vec{\mathbf{X}}$ is $\mathcal{O}(n^{-2})$, in contrast to that in conventional methods of $\mathcal{O}(n^{-1})$. We then show in a Monte Carlo study that `perc-cal` performs better than traditional confidence interval methods, including the BCa method (Efron (1987)), and other Sandwich-based estimators discussed in Cribari-Neto et al. (2007) and MacKinnon (2013). Our study is similar in structure to the simulation study that was performed in Gonçalves and White (2005), but modified to study a very wide variety of misspecified mean functions. We follow up this synthetic simulation study with a real data example involving a criminal sentencing dataset, and show that `perc-cal` once again performs satisfactorily. We have released an `R` package, available through CRAN and coded primarily in `C++`, so that practitioners may benefit from a fast implementation of `perc-cal` for their own analyses.

We argue the combination of theoretical and empirical justification presented in this paper supports the claim that `perc-cal` is a reliable confidence interval methodology that performs well in general, even in the presence of relatively severe model misspecification. The remainder of the paper is organized as follows. Section 2 provides a review of confidence interval methods. Section 3 presents the theoretical results. Section 4 compares the performance of `perc-cal` with that of other often more commonly used confidence interval estimators in synthetic and real data settings. Section 5 provides concluding remarks, and an Appendix gives all of the proofs.

2 Literature Review

2.1 Review of Bootstrap Confidence Intervals

There is a very wide variety of bootstrap methods that have been proposed in the literature to compute $(1-\alpha)$ confidence intervals. These methods include Efron's percentile method (Efron (1981), page 146), Hall's percentile approach (Hall (1992), page 7), and Hall's percentile-t method (Hall (1988), page 937). Other forms of bootstrap CIs include symmetric CIs (Hall (1992), page 108) and short bootstrap confidence intervals (Hall (1992), page 114). In general, performance of these methods depends upon the properties of the data generating process and/or the sample size. We are primarily interested in confidence interval methods that assume much less about the true underlying data generating process, which is usually unknown and often not well behaved in real data applications, making these methods less relevant to the work which follows.

Hall advocates the use of pivotal bootstrap statistics because pivotal bootstrap statistics have higher asymptotic accuracy when the limiting distributions are indeed pivotal (Hall

(1992), page 83). We emphasize that Hall's preference for pivotal bootstrap statistics, and much of the discussion regarding the relative merits of various confidence interval methods, are based on the *asymptotic* properties of these methods. When the sample size is small, these asymptotic considerations do not necessarily reflect the empirical performance of these methods. For example, Hall cautioned that "our criticism of the percentile method and our preference for percentile-t lose much of their force when a stable estimate of σ^2 is not available" (Hall (1988) page 930). Simulation studies that reinforce this include Scholz (2007).

Another class of confidence intervals may be formed by replacing the standard error estimator in the standard `z` or `t` interval with a so-called 'Sandwich' estimator (White (1980)), or one of the many extensions of the Sandwich estimator (Cribari-Neto et al. (2007)). A comprehensive review of Sandwich estimators can be found in MacKinnon (2013), and we will compare these methods with our proposed method in Section 4.

2.2 Review of Iterative Bootstrap Confidence Intervals

The idea of the iterative bootstrap (or double-bootstrap) was first introduced in Efron (1983). The improvement on coverage probability of CIs was first analyzed in Hall (1986) and later discussed in more details in Hall and Martin (1988). A comprehensive review can be found in Section 3.11 in Hall (1992) (see also Efron and Tibshirani (1994), page 268). In general, the iterative bootstrap provides more accurate coverage probability at the cost of more computing.

To fix ideas, in this section we shall introduce the proposed double-bootstrap confidence interval method in a univariate case with generic notations. We will extend this procedure to the regression setting in Section 2.4. We assume that we observe $Z_1, \ldots, Z_m \overset{iid}{\sim} F$ for some distribution F. Let $\theta = \theta(F)$ be a parameter of our interest. We will estimate θ

through the empirical distribution $\hat{F}(z) = \frac{1}{m}\sum_{i=1}^{m} I(Z_i \leq z)$. The estimator is denoted by $\hat{\theta} = \theta(\hat{F}) = \theta(Z_1, \ldots, Z_m)$. The construction of the confidence interval is illustrated in Figure 1 and is described as follows.

1. For chosen bootstrap sample size B_1, obtain bootstrap samples $(\mathbf{Z}_1^*, \ldots, \mathbf{Z}_{B_1}^*)$. Each $\mathbf{Z}_j^*$ consists of m i.i.d. samples with replacement from $\hat{F}$. For chosen bootstrap sample size B_2, obtain double bootstrap samples corresponding to all bootstrap samples, $(\mathbf{Z}_{1,1}^{**}, \ldots, \mathbf{Z}_{1,B_2}^{**}, \mathbf{Z}_{2,1}^{**}, \ldots, \mathbf{Z}_{2,B_2}^{**}, \ldots, \mathbf{Z}_{B_1,1}^{**}, \ldots, \mathbf{Z}_{B_1,B_2}^{**})$ in the same manner as in the first-level bootstrap. Denote the empirical distributions by $\hat{F}_j^*$'s, $j = 1, \ldots, B_1$, and $\hat{F}_{j,k}^{**}$'s, $j = 1, \ldots, B_1, k = 1, \ldots, B_2$, respectively.

2. Obtain parameter estimates corresponding to the observed sample, $\hat{\theta} = \theta(\hat{F})$, all bootstrap samples, $(\hat{\theta}_1^*, \ldots, \hat{\theta}_{B_1}^*)$ with $\hat{\theta}_j^* = \theta(\hat{F}_j^*)$ and all double bootstrap samples corresponding to all bootstrap samples, $(\hat{\theta}_{1,1}^{**}, \ldots, \hat{\theta}_{1,B_2}^{**}, \hat{\theta}_{2,1}^{**}, \ldots, \hat{\theta}_{2,B_2}^{**}, \ldots, \hat{\theta}_{B_1,1}^{**}, \ldots, \hat{\theta}_{B_1,B_2}^{**})$ with $\hat{\theta}_{j,k}^{**} = \theta(\hat{F}_{j,k}^{**})$.

3. Form B_1 double-bootstrap histograms $\hat{\boldsymbol{\theta}}_1^{**}, \ldots, \hat{\boldsymbol{\theta}}_{B_1}^{**}$, where each histogram $\hat{\boldsymbol{\theta}}_j^{**}$ is comprised of all B_2 double bootstrap estimates $(\hat{\theta}_{j,1}^{**}, \ldots, \hat{\theta}_{j,B_2}^{**})$ corresponding to the jth bootstrap sample and estimate, $\mathbf{Z}_j$ and $\hat{\theta}_j$, respectively, $j \in \{1, 2, \ldots, B_1\}$.

4. Find the smallest $\hat{\lambda}$ such that $1/2 < \hat{\lambda} < 1$ and that $\hat{\theta}$ lies in the $1 - \hat{\lambda}$ percentile and the $\hat{\lambda}$ percentile of the histograms $1 - \alpha$ proportion of the time.

5. We know that $\hat{\theta}$ lies between the $(1 - \hat{\lambda}, \hat{\lambda})$ percentiles of the second-level bootstrap distributions $1 - \alpha$ proportion of the time. Therefore our `perc-cal` $1 - \alpha$ interval for θ is equal to the $(1 - \hat{\lambda}, \hat{\lambda})$ percentiles of the first-level bootstrap distribution, $[\hat{\boldsymbol{\theta}}_{(1-\hat{\lambda})}^*, \hat{\boldsymbol{\theta}}_{(\hat{\lambda})}^*]$.

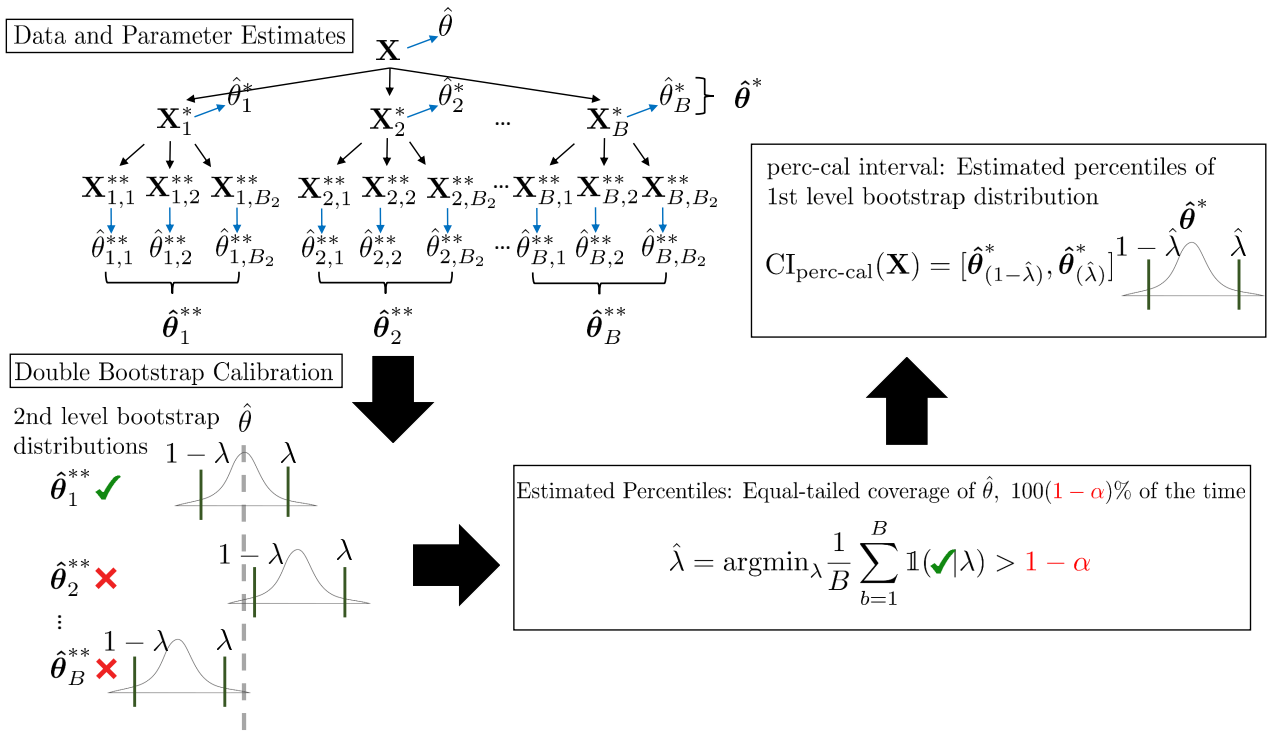

Figure 1: `perc-cal` diagram

For a $(1-\alpha)$ left-sided `perc-cal` confidence interval for θ, the only change in the procedure is in Step 4, where one uses the histograms to find the smallest $\hat{\lambda}$ such that $\hat{\theta}$ lies below the $\hat{\lambda}$ percentile of the histograms $1-\alpha$ proportion of the time. In what follows, we shall refer the two-sided `perc-cal` interval as $\mathcal{I}_2 = [\hat{\boldsymbol{\theta}}^*_{(1-\hat{\lambda})}, \hat{\boldsymbol{\theta}}^*_{(\hat{\lambda})}]$ and the one-sided `perc-cal` interval as $\mathcal{I}_1 = (\infty, \hat{\boldsymbol{\theta}}^*_{(\hat{\lambda})}]$.

A similar double-bootstrap confidence interval is the double-bootstrap-t method which uses the second-level bootstrap to calibrate the coefficient of the bootstrap standard deviation estimate. In practice, both methods can be applied. Hall commented in his book (Hall (1992), page 142) that "either of the two percentile methods could be used, although

the 'other percentile method' seems to give better results in simulations, for reasons that are not clear to us." Here "the 'other percentile method"' refers to confidence intervals $\mathcal{I}_1$ and $\mathcal{I}_2$. Our simulation studies in Section 4 demonstrate the same phenomenon.

Research on optimizing the trade-off between the number of simulations in double-bootstrap and the CI accuracy can be found in Beran (1987), Beran (1988), Booth and Hall (1994), Booth and Presnell (1998), Lee and Young (1999), among many others. Since the computation of `perc-cal` is reasonably efficient as discussed in Section 5, we do not pursue this type of optimization here but note that further performance gains are a promising area for future research.

2.3 Review of Bootstrap Applications in Conventional Linear Models

Bootstrap in linear models is studied in Section 4.3 in Hall (1992). Hall refers the fixed design case the "regression model" and the random design case the "correlation model." Bootstrap estimation and confidence intervals for the slopes, as well as simultaneous confidence bands, are described.

Bootstrap is widely used in regression models because of its robustness to the sample distributions since the seminal paper Freedman (1981). A review of bootstrap methods in economics can be found in MacKinnon (2006). Gonçalves and White (2005) consider bootstrapping the sandwich estimator for the standard error when the observations are dependent and heterogeneous. Bootstrap applications under other types of model misspecifications are recently considered in Kline and Santos (2012) and Spokoiny and Zhilova (2014). In this paper, we focus on a different case when observations of $(Y, \vec{\mathbf{X}})$ are i.i.d. but the joint distribution is assumption-lean – we elaborate upon this further in the next

section.

2.4 The Assumption-Lean Framework and Double-Bootstrap Applications

Conventional linear models assume $\mathbf{E}[Y|\vec{\mathbf{X}}] = \vec{\mathbf{X}}\boldsymbol{\beta}$ for some $(p+1)$-vector $\boldsymbol{\beta}$ as the slope coefficients, so that Y depends on $\vec{\mathbf{X}}$ only through a linear function. While ths is commonly assumed in the bootstrap literature, we may not want to require it when performing inference in real data settings because the true relationship may not be linear. Moreover, as first noted in White (1980), a non-linear relationship between Y and $\vec{\mathbf{X}}$ and randomness in $\vec{\mathbf{X}}$ can lead to serious bias in the estimation of standard errors. Buja et al. (2015) reviewed this problem, and proposed an "assumption-lean" framework for inference in regressions. In this framework, we do not posit any assumptions on the relationship between Y and $\vec{\mathbf{X}}$. We only assume the existence of certain moments of the joint distribution of $\vec{\mathbf{V}} = (X_1, \ldots, X_p, Y)^T$, which we denote by G. This consideration makes the model very general and thus widely applicable.

Even though a linear relationship between $\mathbf{E}[Y|\vec{\mathbf{X}}]$ and $\vec{\mathbf{X}}$ is not assumed in an assumption-lean framework, the slope coefficients that are estimated are always well-defined through a population least-squares consideration: the population least-squares coefficients $\boldsymbol{\beta}$ minimize squared error risk over all possible linear combinations of $\vec{\mathbf{X}}$:

$$\boldsymbol{\beta} = argmin_{\mathbf{b}}\mathbf{E}\|Y - \mathbf{b}^T\vec{\mathbf{X}}\|_2^2 = \mathbf{E}[\vec{\mathbf{X}}\vec{\mathbf{X}}^T]^{-1}\mathbf{E}[Y\vec{\mathbf{X}}]. \quad (1)$$

This definition of linear coefficients $\boldsymbol{\beta}$ is meaningful in addition to being well defined: $\boldsymbol{\beta}$ provides us with the best linear approximation from $\vec{\mathbf{X}}$ to Y, whether or not $\vec{\mathbf{X}}$ and Y are linearly related to one another. This setup allows for situations including but not limited to random $\vec{\mathbf{X}}$, non-linearity and heteroskedasticity and we show later that the proposed

`perc-cal` method provides better empirical coverage of the true population least-squares coefficients $\boldsymbol{\beta}$ on average over a wide variety of data generating processes, even if those data generating processes involve random $\vec{\mathbf{X}}$, non-linearity and/or heteroskedasticity.

To estimate $\boldsymbol{\beta}$, denote the i.i.d. observations of $\vec{\mathbf{V}}$ by $\vec{\mathbf{V}}_1, \ldots, \vec{\mathbf{V}}_n$ and denote the $n \times (p+1)$ matrix with rows $\vec{\mathbf{V}}_1, \ldots, \vec{\mathbf{V}}_n$ by $\mathbf{V}$. The multivariate empirical distribution of $\vec{\mathbf{V}}$ is then $\hat{G}(\vec{\mathbf{V}}) = \hat{G}(x_1, \ldots, x_p, y) = \frac{1}{n}\sum_{i=1}^{n} I(X_{1,i} \leq x_1, \ldots, X_{p,i} \leq x_p, Y_i \leq y)$. The least squares estimate for $\boldsymbol{\beta}$ defined in (1) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2}$$

where $\mathbf{X}$ is the $n \times (p+1)$ matrix and $\mathbf{Y}$ is the $n \times 1$ vector containing the i.i.d. observations of $\vec{\mathbf{X}}$ and Y respectively. Note that each estimate $\hat{\beta}_j$ of β_j can be written as a function of $\hat{G}$, $\hat{\beta}_j = \beta_j(\hat{G})$.

The `perc-cal` confidence intervals for each of the slopes β_j in $\boldsymbol{\beta}$ are constructed similarly as described in Section 2.2. We use the pairs bootstrap first proposed by Freedman (1981) and create B_1 i.i.d. bootstrap samples $\mathbf{V}_k^*$'s, where each matrix $\mathbf{V}_k^*$ consists n i.i.d. samples with replacement from $\hat{G}$ and has empirical distribution $\hat{G}_k^*$. To find the proper calibration in $\mathcal{I}_1$ or $\mathcal{I}_2$, we sample B_2 i.i.d. pairs bootstraps $\mathbf{V}_{k,h}^{**}$ with empirical distributions $\hat{G}_{k,h}^{**}$. The other steps in the construction are identical to those in Section 2.2 with $\theta(\cdot)$ replaced by $\beta_j(\cdot)$ with respective empirical distributions as arguments.

3 Asymptotic Theory

In this section, we discuss the theoretical properties of `perc-cal`. The following theorem describes the accuracy on the coverage probability of `perc-cal` confidence intervals.

Theorem 3.1 *Consider n i.i.d. observations of the $(p+1)$-dimensional random vector $\vec{\mathbf{V}} = (X_1, \ldots, X_p, Y)^T$. Suppose that*

1. *(Non-degeneracy condition) $\exists C_0 > 0$, such that the minimal eigenvalue of the covariance matrix $Var(\vec{\mathbf{V}})$ is larger than C_0.*
2. *(Moment condition) $\mathbf{E}[\|\vec{\mathbf{V}}\|_2^{16}] < \infty$.*
3. *(Density condition) The distribution of $\vec{\mathbf{V}}$ is absolutely continuous with respect to the Lebesgue measure.*

Consider the population least squares parameter defined in (1) *whose estimate is the sample least squares defined in* (2)*. Then for each j, the $(1-\alpha)$* `perc-cal` *CI for β_j described in Section 2.2 and Section 2.4 have the coverage probabilities*

$$\mathbf{P}(\boldsymbol{\beta}_j \in \mathcal{I}_1) = 1 - \alpha + O(n^{-1}). \tag{3}$$

$$\mathbf{P}(\boldsymbol{\beta}_j \in \mathcal{I}_2) = 1 - \alpha + O(n^{-2}). \tag{4}$$

The proof of Theorem 3.1 is in Section A in the Appendix. It uses proof techniques similar to those used in Section 3.11.3 of Hall (1992), with a focus on the regression setting.

Remark 1. Note that general one-sided confidence intervals have a coverage probability of $1 - \alpha + O(n^{-1/2})$, and two-sided confidence intervals have a coverage probability of $1 - \alpha + O(n^{-1})$ (see Section 3.5.4 and Section 3.5.5 in Hall (1992)). Thus, the double bootstrap method provides better coverage but at a higher computational cost.

Remark 2. The requirement of the finiteness of the 16th moment is a technical requirement to guarantee the asymptotic theory. Simulation results show that `perc-cal` works for many general distributions. Furthermore, in practice, stronger conditions such as sub-Gaussian or sub-exponential tails are often assumed, which require all moments of the distribution to be finite.

Remark 3. A weaker version of the density condition is the Cramér's condition that the characteristic function $\chi(\mathbf{t})$ of $\vec{V}$ satisfies $\limsup_{\|\mathbf{t}\|_2 \to \infty} |\chi(\mathbf{t})| < 1$. This condition requires that the atoms in the distribution of the sample mean of $\vec{V}$ to have exponentially small mass (see Hall (1992), page 57 for a related discussion).

Remark 4. There are two possible extensions of the results in Theorem 3.1: (1) In practice one often wants to construct simultaneous confidence intervals for $\boldsymbol{\beta}_j$'s, and (2) in modern datasets the number of predictors is usually large and it is possible $p \to \infty$. One possible approach to these two problems utilizes the results in Portnoy (1986) who considered the central limit Gaussian approximation of the joint distribution of a random vector whose dimension tends towards infinity. While these extensions are interesting, they are not central to the focus of our paper, and so we leave these problems to future work.

4 Numerical Studies

In this section, we study the performance of `perc-cal` compared to alternative (often more common) methods for forming confidence intervals, including other double bootstrap methods. We first compare `perc-cal` to these other methods using simulated data under a very wide variety of true data generating processes. We then illustrate our approach in a real data example. We will see that `perc-cal` performs very satisfactorily in general.

4.1 Synthetic Simulation Study

4.1.1 Design: Synthetic Simulation

We compare the performance of `perc-cal` with 10 other methods that are commonly used for constructing confidence intervals:

1. Standard normal interval: `z` (Efron and Tibshirani (1994), pp. 168).

2-6. Five sandwich variants: `sand1`, `sand2`, `sand3`, `sand4` and `sand5` (MacKinnon (2013) provides a review of these methods, denoted there by H1, H2, H3, HC4 and HC5).

7. Hall's Studentized interval: `stud` (Hall (1988)).

8. Hall's "bootstrap-t" method: `boot-t` (Efron and Tibshirani (1994), pp. 160-162).

9. Efron's BCa interval: `BCa` (Efron (1987)).

10. Single percentile method: `perc` (Efron and Tibshirani (1994), pp. 170).

We consider a very wide range of underlying true data generating models, to obtain a more general understanding for how these confidence interval methods compare against one another in a wide variety of data settings, for large sample sizes as well as small. The data generating models represent a full factorial design of the following factors, excluding non-denerate combinations (i.e., combinations for which the conditional mean is not finite):

- Simple regression - one predictor, $Y = \beta_0 + \beta_1 X + \epsilon$
- Sample size $n = 32, 64, 128, 256$
- Relationships between Y and X: (1) $Y = X + e$; (2) $Y = \exp(X) + e$; (3) $Y = X^3 + e$.
- Distribution of X: (1) $X \sim \mathcal{N}(0, 1)$, (2) $X \sim \exp(\mathcal{N}(0, 1))$.
- Noise: $\epsilon \sim (1)N(0, 1); (2)\ |X| * \mathcal{N}(0, 1); (3)\ \exp(\mathcal{N}(0, 1))$.

In each of the above cases, we use 2000 first and second-level bootstrap samples for all bootstrap methods ($B = B_2 = 2000$). We obtain empirical coverage figures for the slope

coefficient in the regression, β_1. Results are averaged over 500 replications to reduce the empirical standard error of the resulting intervals to below 1.5% on average across scenarios and methods. We present the results for a target coverage of 90% ($\alpha = .05$), without loss of generality (results for a target coverage of 95% are qualitatively the same).

4.1.2 Results

To more easily visualize the performance of many methods under many different scenarios, we begin with a coverage scatterplot in Figure 2. On the x-axis, we provide the average coverage proportion of β_1 using `sand1`. On the y-axis, we provide the coverage proportion of alternative methods. We exclude `sand2`, `sand3`, and `sand4` but include `sand5`, because `sand5` performs better than `sand2`, `sand3`, and `sand4`. For ease of comparison, we add a 45 degree diagonal line to each graph – all points below this line represent methods which had empirical coverage less than `sand1`. We also add a horizontal line to the graph at the desired target coverage level of 90%.

In general, none of the methods were "perfect" in the sense of always providing coverage at or above the target level of coverage. All noticeably undercover in particular cases and in these cases, `perc-cal`'s relative performance is generally noticeably strong, achieving higher empirical coverage than alternative methods.

We see that across scenarios, `perc-cal` provided the most consistent empirical coverage. `sand5` does not provide empirical coverage which is as consistent (coverage is frequently below 85%), but was itself generally better than the other alternative methods, including `BCa`, which has favorable asymptotic properties. `perc-cal` occassionally over-covers by a moderate amount, between 5% and 10%. The other methods rarely over-cover, although most methods provide good coverage (above 85%) as we move to the right hand side of the plot, which represents the region in which `sand1` covers satisfactorily (above 85%). At the

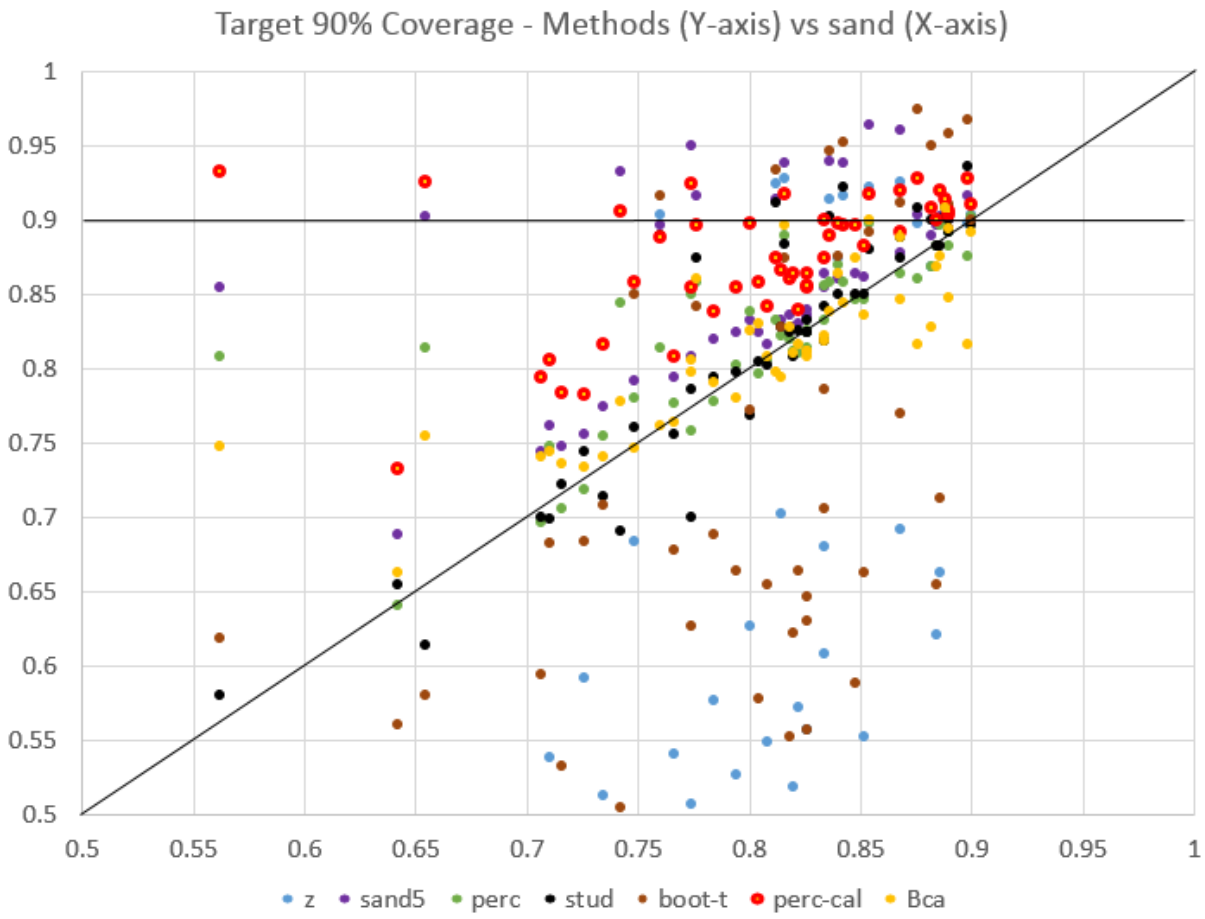

Figure 2: Scatterplot of coverage proportion of methods versus `sand` – 90% Target Coverage

same time, we see some notable cases where `BCa` does not cover well in this region (i.e., coverage for `BCa` falls below 85% when `sand1` has coverage above 85%).

In summary, there is a clear preference here for `perc-cal`; especially because moderate over-coverage (i.e., overcoverage by less than 3%) is a less serious defect than very noticeable under-coverage.

To better appreciate the ability of `perc-cal` to provide empirical coverage equal to the

target level in the presence of misspecification, consider one illustrative scenario: $Y = X + e$; $X \sim \mathcal{N}(0,1)$; $\epsilon \sim |X| * \mathcal{N}(0,1)$; $n = 64$. This scenario is challenging because the errors are heteroskedastic.

The various methods performed unevenly, despite a relatively large sample size. While target coverage was 90%, the traditional z-interval had empirical coverage of only 60.8%. `boot-t` also had very poor performance at 70.6%. `sand5`, the best performing sandwich estimator averaged across scenarios, achieved empirical coverage of 86.4%. The traditional percentile method covered 85.6% of the time, better than BCa which covered 81.8% of the time, but all were below the 90.0% coverage that `perc-cal` was able to achieve.

`perc-cal`'s improved coverage came at the cost of longer interval lengths – `perc-cal` had an average interval length of 0.74, at the upper end of other methods – excluding the poor performance of `z` and `boot-t`, these other methods had interval lengths between 0.64 and 0.73. While these other methods had shorter interval lengths, it would not be acceptable to a practitioner for this shortness to come at the expense of falling below desired target coverage. Only when target coverage is achieved do considerations like average interval length become a primary concern. In this particular scenario, one must reach further into the tails of the bootstrap distribution for β_1 to get close to target coverage.

4.2 Real Data Example: Criminal Sentencing Dataset

4.2.1 Design: Real Data Example

We turn now to an example of how well `perc-cal` performs in practice, on real data. In this section, we compare `perc-cal` to other methods on a criminal sentencing dataset. This dataset contains information regarding criminal sentencings in a large state from January 1st 2002 to December 31st 2007. There are a total of 119,983 offenses in the dataset,

stemming from a variety of crimes – murder (899 cases), violent crime (80,402 cases), sex-related crime (7,496 cases), property-related crime (92,743 cases), firearm-related crime (15,326 cases) and drug crime (93,506 cases). An individual offense can involve multiple types of crime, and an offender's case can involve multiple charges of each type of crime. This is truly a random X setting because the predictors themselves are stochastic, coming to us from an unknown distribution. $\vec{\mathbf{X}}$ is stochastic, the relationship between $\vec{\mathbf{X}}$ and Y is unknown and possible non-linear, and error may have heteroskedastic variance.

Our modeling objective is to form marginal confidence intervals for the slope coefficients of a linear regression. The response variable of our regression is the number of days of jail time an offender must serve (log-transformed), which we predict with the following 8 covariates:

1. `race`: Binary variable for the race of the offender (1 if white, 0 if non-white).
2. `seriousness`: A numerical variable scaled to lie between 0 and 10 indicating the severity of the crime. A larger number denotes a more serious crime.
3. `age`: Age of offender at the time of the offense.
4. `race`: The percent of the neighborhood that is not of Caucasian ethnicity in the offender's home zip code.
5. `in-state`: Binary variable for whether the offender committed the crime in his/her home state (1 if in offenders home state, 0 otherwise).
6. `juvenile`: Binary variable for whether the offender had at any point committed a crime as a juvenile (1 if yes, 0 otherwise). 18% of all offenses involved offenders who had committed a crime as a child.

7. `prior-jaildays`: Number of days of jail time the offender had previously served.

8. `age-firstcrime`: The age of the offender when the offender was charged with his/her first crime as an adult.

We run a regression of log-transformed sentence length against race, severity of crime, age at time of offense, the percent of the neighborhood which is black in the offenders home zip code, whether the offender has committed a crime as a child or not, the total amount of prior jail time the offender served prior to the current offense, and the age of the offender at the time of the offenders first crime.

We first run a linear regression upon the full dataset containing all 119,983 offenses. We treat the coefficients as if this is a population-level regression. We then proceed as if we do not have the full dataset and instead only have the ability to observe random subsets of the dataset of size 500 – large, but not so large that all coefficient estimates are over-powered. We study the empirical coverage performance of confidence intervals formed using the methods in the simulation exercise over repeated realizations which are obtained through random subsamples of size 500.

Although this dataset is highly relevant, the response variable is contaminated by mis-recorded jail length figures. The contamination does not affect our ability to compare the empirical coverage fidelity of various confidence interval methods, but we lose the ability to read into the actual regression coefficient estimates thus obtained.

4.2.2 Results: Real Data Example

Linear regression across the full dataset has an R^2 of 16.9% with 6 of the 8 predictors coming up as significant. We then take repeated random subsamples of size 500 from this population of offenses and treat these subsamples as if they were the observed dataset.

Presupposing that each crime represents an *iid* draw, this framework allows us to compare and contrast the empirical coverage performance of confidence interval methods.

In Figure 3, we present the empirical coverage for each of our predictors when we form 90% confidence intervals. The y-axis of the plot below represents the empirical coverage over 10,000 realizations for each of the methods in question (i.e., 90% empirical coverage for a particular method implies that 9,000 of the 10,000 realizations had confidence intervals for that method which contained the true but unknown population-level parameters). Along the x-axis, we have the predictors listed above. We include a bold horizontal line at the target level of empirical coverage of 90%. The standard error associated with the coverages presented below average to .002 across scenarios, predictors and methods.

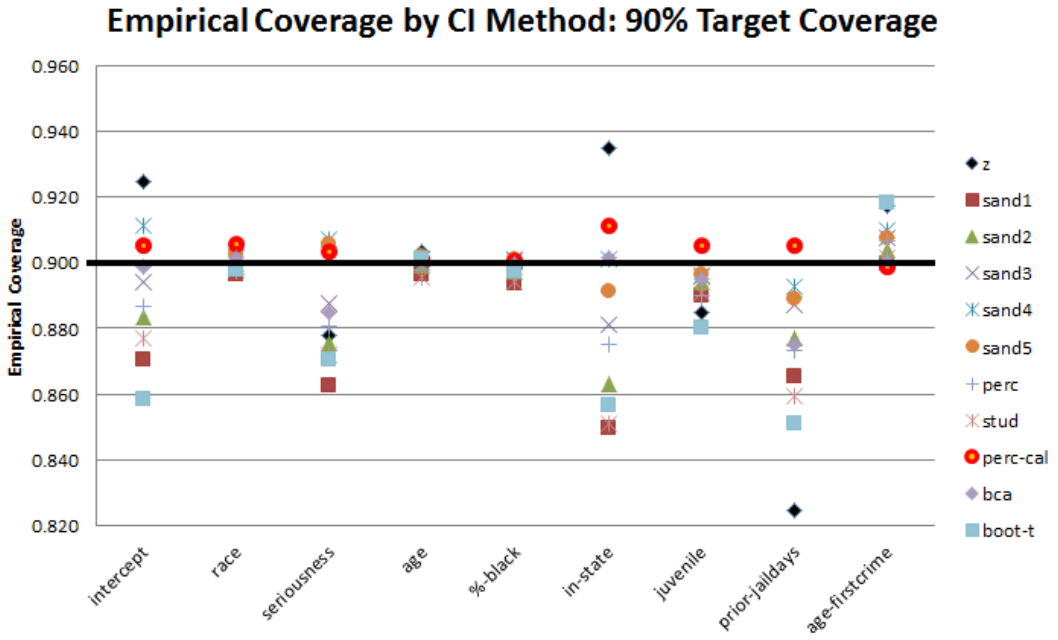

Figure 3: Scatterplot of coverage proportion of methods – 90% Target Coverage

There are a number of inferences that we can draw from the above chart:

- All methods generally perform as expected, with empirical coverage proportions generally falling between 85% and 92%.

- `perc-cal` is the only method that consistently achieves empirical coverage over 90%.

- `prior-jaildays` appears to be the predictor with the most disappointing empirical coverage. All methods except for perc-cal do not achieve 90% empirical coverage. The average empirical coverage of `prior-jaildays` for all non-`perc-cal` methods was 87.0%.

- There is also considerable disparity in the ability of various methods to cover the coefficients associated with the intercept term and the `in-state` covariate. Although the `BCa` method has near-90% empirical coverage of the `in-state` super-population coefficient, its coverage is less satisfactory for the `seriousness` and `prior-jaildays` covariates.

When we plot the relationship of jail length (log transformed) against prior total jail length in Figure 4, adjusted for all of the other covariates in the super-population, we see an almost bi-model relationship.

It is clear from the plot in Figure 4 that the highly misspecified relationship between Y and $\vec{\mathbf{X}}$ is likely to be driving the large disparity (and general deterioration) in coverage performance across the various non-`perc-cal` confidence interval methods. Overall, these results support the notion that `perc-cal` is a good all-purpose confidence interval method, and that all other methods, while performing well for some of the covariates, do not perform well for all of the covariates as was the case for `perc-cal`. The results assuming target coverage of 95% are qualitatively the same as the results presented above.

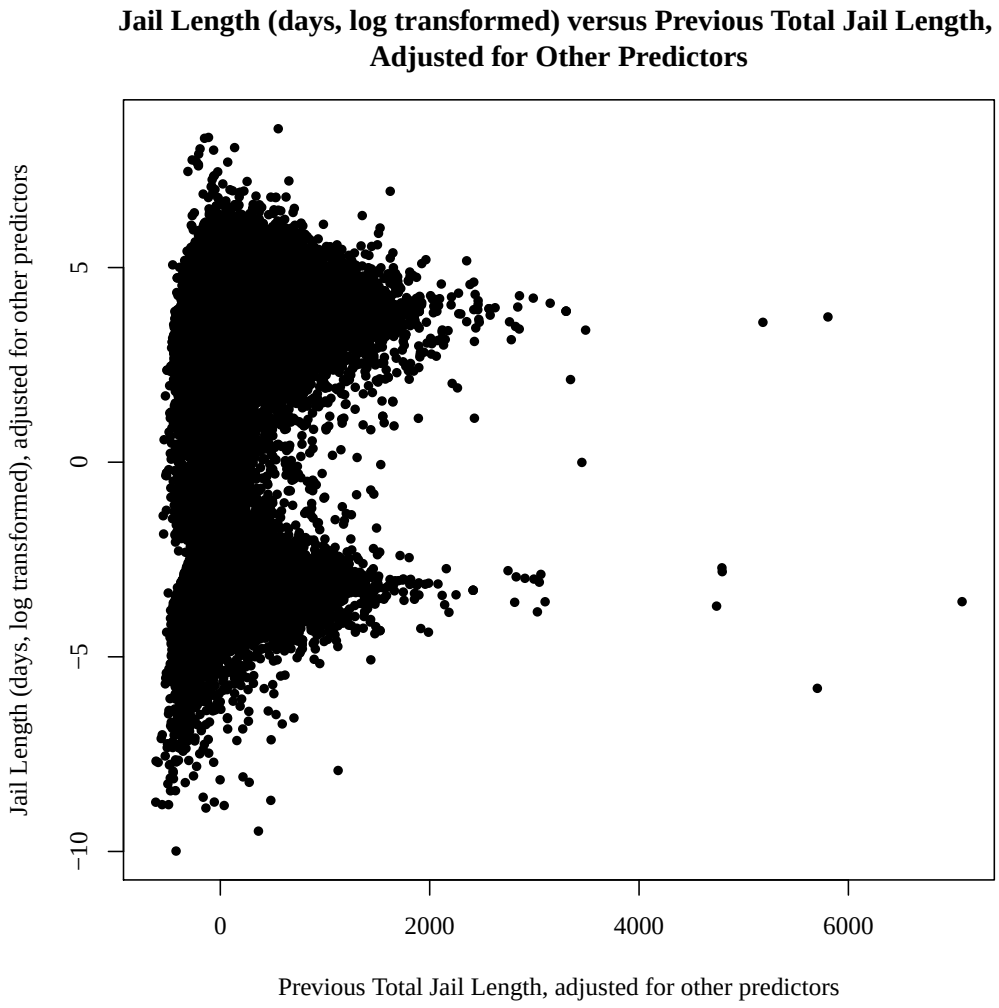

Figure 4: Jail Length (log transformed) versus Previous Total Jail Length, Adjusted for Other Predictors

5 Discussion and Concluding Remarks

If `perc-cal` performs so well relative to alternative more popular CI methods, why is it not used more in practice? We believe the use of double bootstrap methods in general have not

been widely adopted primarily because of their computational cost. Although it is true that double bootstrap methods in general and `perc-cal` in particular require more computation, the computational burden of these procedures is far less problematic than in the past because of current computational advances. For example, the rise of grid computing has greatly facilitated parallel computation. Because `perc-cal` is trivially parallelizable, it is relatively straightforward to compute all second-level bootstrap calculations in parallel, allowing researchers to compute `perc-cal` at a computational "cost" that is on the order of a single bootstrap. Furthermore, the perceived computational cost of double bootstrap methods may be inflated due to the inefficiency with which the calculations are carried out in popular statistical programming languages, most notably `R` – the very same calculations are orders of magnitude faster in lower level languages, such as `C++`. The rising popularity and adoption of packages integrating `R` with `C++` (Eddelbuettel et al. (2011)) can greatly reduce the cost of double bootstrap methods for practitioners performing data analysis in `R` who do not know `C++`. In the spirit of this, the `R` package we have created allows users to compute `perc-cal` intervals in `R` efficiently using `C++` code via `Rcpp`. We are optimistic that the use of double bootstrap methods will only increase further as the cost of computing declines further over the next 10 years.

We have restricted our attention to equal-tailed intervals for all methods considered here. It is natural and certainly possible to extend our approach to compute the shortest *unequal*-tailed interval, even if other methods cannot or would not, because of the symmetry of the asymptotic distribution underlying those alternative methods. At the same time, this advantage should not be over-stated – for example, one may be forced so far into the tails of the bootstrap distribution that a considerably larger number of first and second-level bootstrap samples are required. Because this is not the focus of our paper, we do not pursue it further here.

The asymptotic theory we developed, examined, and compared the more traditional percentile and "bootstrap-t" methods to their double bootstrap analogs in our "assumption lean" setting. We did not study the asymptotic properties of alternative confidence interval methods in our setting. Although it would be interesting to do so, there are a very wide range of methods in the literature, making systematic theoretical study impractical. We leave this for future work.

In summary, randomness in $\vec{\mathbf{X}}$, non-linearity in the relationship between Y and $\vec{\mathbf{X}}$, and heteroskedasticity "conspire" against classical inference in a regression setting (Buja et al. (2015)), particularly when the sample size is small. We have shown that in theory, the percentile-calibrated method `perc-cal` provides very satisfactory empirical coverage – the asymptotic rate of coverage error under mild regularity conditions for a two-sided confidence interval of the best linear approximation between Y and $\vec{\mathbf{X}}$ is $\mathcal{O}(1/n^2)$. Furthermore, `perc-cal` performs very well in practice, both in synthetic and real data settings. We believe that `perc-cal` is a good general-purpose CI method and merits consideration when confidence intervals are needed in applied settings by practitioners.

A Proof of Theorem 3.1

The proof consists of two parts. The first part shows the existence of the Edgeworth expansion of the pivoting quantity. The second part derives the asymptotic order of the error term by using the first terms in this expansion and the Cornish-Fisher expansion, which can be regarded as the inverse of the Edgeworth expansion.

Part I: Existence of Edgeworth Expansion.

This part of proof relies on the following general theorem on the existence of the Edgeworth expansion and its bootstrap version. We state it as in Section 5.2 in Hall (1992) with

modifications for later confidence interval constructions.

Theorem A.1 *Let $\mathbf{W}_1, \ldots, \mathbf{W}_n$ be a random sample from a d-variate population $\mathbf{W}$, and write $\bar{\mathbf{W}} = \frac{1}{n}\sum_{i=1}^n \mathbf{W}_i$ as the sample mean vector. Denote the population mean vector as $\boldsymbol{\mu} = \mathbf{E}[\mathbf{W}]$. For a parameter $\theta_0 = g(\boldsymbol{\mu})$ with $g : \mathbb{R}^d \to \mathbb{R}$, consider the estimate of the form $\hat{\theta} = g(\bar{\mathbf{W}})$, whose asymptotic variance is $n^{-1}\sigma^2 = n^{-1}h(\boldsymbol{\mu})^2$ which is estimated by $\hat{\sigma}^2 = h(\bar{\mathbf{W}})^2$. Suppose the following regularity conditions hold:*

1. *(Smoothness conditions) g and h are smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$.*
2. *(Moment condition) $\mathbf{E}[\|\mathbf{W}\|_2^4] < \infty$.*
3. *(Density condition) The distribution of $\mathbf{W}$ is absolutely continuous with respect to the Lebesgue measure.*

Then the following results hold:

1. *For the statistic: $A(\bar{\mathbf{W}}) = \frac{\hat{\theta}-\theta_0}{\sigma}$, we have*

$$\mathbf{P}(\sqrt{n}A(\bar{\mathbf{W}}) < x) = \Phi(x) + \sum_{j=1}^{2} \pi_j(x)\phi(x) + o(n^{-1}) \tag{5}$$

   uniformly for any x, for some polynomials $\pi_j(x)$ whose coefficients depend only on the moments of $\mathbf{W}$ up to the 4-th moment. Moreover, $\pi_1(x)$ is an even function and $\pi_2(x)$ is an odd function.

2. *Consider n-out-of-n bootstrap with $\bar{\mathbf{W}}^* = \frac{1}{n}\sum_{i=1}^n \mathbf{W}_i^*$ as the resample mean, and $\hat{\theta}^* = g(\bar{\mathbf{W}}^*)$ and $\hat{\sigma}^* = h(\bar{\mathbf{W}}^*)$ as the bootstrap estimates. Then the resample statistic $\hat{A}(\bar{\mathbf{W}}^*) = \frac{\hat{\theta}^*-\hat{\theta}}{\hat{\sigma}}$ satisfies*

$$\sup_{-\infty<x<\infty} \left|\mathbf{P}(\sqrt{n}\hat{A}(\bar{\mathbf{W}}^*) \le x | \mathbf{W}_1, \ldots, \mathbf{W}_n) - \Phi(x) - \sum_{j=1}^{2} \hat{\pi}_j(x)\phi(x)\right| = O(n^{-3/2}). \tag{6}$$

where the coefficients in $\hat{\pi}_j$ are sample versions of the moments of $\mathbf{W}$*, so that the error in each estimate is* $O_p(n^{-1/2})$.

In what follows, we check the regularity conditions in Theorem A.1 for the population least squares parameter $\boldsymbol{\beta}$ defined in (1). We consider

$$\mathbf{W} = (X_1, \ldots, X_p, Y, X_1^2, X_1X_2, \ldots, Y^2, X_1^3, \ldots, Y^3, X_1^4, \ldots, Y^4)^T$$

being the $(\binom{p+2}{4}-1)\times 1$ vector consisting all first through fourth moments of $(X_1, \ldots, X_p, Y)$. Thus, with $\boldsymbol{\mu} = \mathbf{E}[\mathbf{W}]$, we have

$$\boldsymbol{\beta}_{(p+1)\times 1} = \boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{\mu}) = \mathbf{E}[\vec{\mathbf{X}}\vec{\mathbf{X}}^T]^{-1}\mathbf{E}[Y\vec{\mathbf{X}}] \tag{7}$$

where each component of the mapping $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^{p+1}$ is a smooth function from $\mathbb{R}^d$ to $\mathbb{R}$ due to the non-degeneracy condition in Theorem 3.1. The sample least squares estimate $\hat{\boldsymbol{\beta}}$ defined in (2) can now be written as $g_{\boldsymbol{\beta}}(\hat{\mathbf{W}})$.

Similarly, the $(p+1)\times 1$ vector consisting of the asymptotic variance of each component in $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\sigma}^2_{(p+1)\times 1}$, can be written as a function of $\boldsymbol{\mu}$, that is,

$$\boldsymbol{\sigma}^2 = \boldsymbol{h}(\boldsymbol{\mu})^2 = diag\{\mathbf{E}[\vec{\mathbf{X}}\vec{\mathbf{X}}^T]^{-1}\mathbf{E}[(Y - \vec{\mathbf{X}}^T\boldsymbol{\beta})^2\vec{\mathbf{X}}\vec{\mathbf{X}}^T]\mathbf{E}[\vec{\mathbf{X}}\vec{\mathbf{X}}^T]^{-1}\}. \tag{8}$$

By the non-degeneracy condition again, each component of $\boldsymbol{h}$ is a smooth function from $\mathbb{R}^d$ to $\mathbb{R}$.

The moment condition and the density condition follow directly from the assumptions in Theorem 3.1. Thus, Theorem A.1 applies and the Edgeworth expansion for $A(\bar{\mathbf{W}}^*)$ exists.

Part II: The Asymptotic Accuracy of Double Bootstrap CIs

With the existence of the Edgeworth expansion, we develop the asymptotic accuracy of the two-sided double-bootstrap CI for regression. In this section, we use θ_0 to denote a

generic β_j and use $\hat{\theta}$ to denote the corresponding β_j. We show here only the proof for the two-sided `perc-cal` confidence intervals $\mathcal{I}_1$. The one-sided case for $\mathcal{I}_1$ is proved in a similar (and easier) manner. The techniques used in this proof are patterned after those in Section 3.11 in Hall (1992) but are reorganized for readability and included so that our analysis is self-contained.

Consider the distribution of $\hat{A}(\bar{\mathbf{W}}^*) = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}}$. Note that for any $0 < \gamma < 1$, the quantile estimate $\hat{v}_\gamma$ satisfies

$$\mathbf{P}(\sqrt{n}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma} \leq \hat{v}_\gamma | \mathbf{W}_1, \ldots, \mathbf{W}_n) = \gamma. \tag{9}$$

Due to the Edgeworth expansion, we can write $\hat{v}_\gamma$ in the standard normal quantile z_γ through the Cornish-Fisher expansion:

$$\hat{v}_\gamma = z_\gamma + n^{-1/2}\hat{p}_1(z_\gamma) + n^{-1}\hat{p}_2(z_\gamma) + O_p(n^{-3/2}) \tag{10}$$

where $\hat{p}_1$ and $\hat{p}_2$ are polynomials whose coefficients are sample estimates of that of p_1 and p_2, and the coefficients of p_1 and p_2 depend only on up to 4-th moments of $\mathbf{W}$. Given the moment condition that the 16th moments of $\mathbf{W}$ are finite, all of these estimates are root-n consistent.

Now consider the quantile $\hat{w}_\lambda$ in the bootstrap distribution of $\hat{\theta}^*$ such that

$$\mathbf{P}(\hat{\theta}^* \leq \hat{w}_\lambda | \mathbf{W}_1, \ldots, \mathbf{W}_n) = \lambda. \tag{11}$$

By comparing (11) and (9) with $\gamma = \lambda$, we see

$$\hat{w}_\lambda = \hat{\theta} + n^{-1/2}\hat{\sigma}\hat{v}_\lambda = \hat{\theta} + n^{-1/2}\hat{\sigma}(z_\lambda + n^{-1/2}\hat{p}_1(z_\lambda) + n^{-1}\hat{p}_2(z_\lambda) + O_p(n^{-3/2})) \tag{12}$$

Thus, by Proposition 3.1 in Hall (1992) (page 102), we have

$$\begin{aligned} &\mathbf{P}(\theta_0 \in (-\infty, \hat{w}_\lambda)) \\ =&\mathbf{P}(\theta_0 \leq \hat{\theta} + n^{-1/2}\hat{\sigma}(z_\lambda + n^{-1/2}\hat{p}_1(z_\lambda) + n^{-1}\hat{p}_2(z_\lambda) + O_p(n^{-3/2}))) \\ =&\lambda + n^{-1/2}r_1(z_\lambda)\phi(z_\lambda) + n^{-1}r_2(z_\lambda)\phi(z_\lambda) + O(n^{-3/2}) \end{aligned} \tag{13}$$

where ϕ is the density of the standard normal distribution, and r_1 and r_2 are even and odd polynomials whose coefficients can be root-n consistently estimated.

Let $\xi = 2(1-\alpha/2-\lambda)$ and $\lambda = 1-\alpha/2+\xi/2$. To find a proper λ for the `perc-cal` interval $\mathcal{I}_2$ is now to find ξ such that

$$\mathbf{P}(\theta_0 \in (\hat{w}_{1-\lambda}, \hat{w}_\lambda)) = \mathbf{P}(\theta_0 \in (\hat{w}_{\alpha/2-\xi/2}, \hat{w}_{1-\alpha/2+\xi/2})) = 1-\alpha. \tag{14}$$

Note that the coverage probability of a two-sided CI can be written as

$$\begin{aligned} &\mathbf{P}(\theta_0 \in (\hat{w}_{1-\lambda}, \hat{w}_\lambda)) \\ =&\mathbf{P}(\theta_0 \in \hat{w}_\lambda) - \mathbf{P}(\theta_0 \in (\hat{w}_{1-\lambda})) \\ =&2\lambda - 1 + 2n^{-1} r_2(z_\lambda)\phi(z_\lambda) + O(n^{-2}). \\ =&1-\alpha+\xi + 2n^{-1} r_2(z_{1-\alpha/2+\xi/2})\phi(z_{1-\alpha/2+\xi/2}) + O(n^{-2}) \end{aligned} \tag{15}$$

The cancellation of the $O(n^{-1/2})$ term due to that $-z_{1-\lambda} = z_\lambda$ and that r_1 is an even polynomial is crucial for the improvement in double-bootstrap. To achieve the accuracy of the coverage in Theorem 3.1, we would like to choose ξ such that

$$\xi = -2n^{-1} r_2(z_{1-\alpha/2+\xi/2})\phi(z_{1-\alpha/2+\xi/2}) + O(n^{-2}) \tag{16}$$

Now consider the second-level bootstrap, in which we calibrate $\hat{\xi}$ for $\hat{\lambda} = 1-\alpha/2+\hat{\xi}/2$ in the `perc-cal` intervals. Through a similar argument for the first-level bootstrap, we see that the calibrated $\hat{\xi}$ satisfies that

$$\hat{\xi} = -2n^{-1} \hat{r}_2(z_{1-\alpha/2+\hat{\xi}/2})\phi(z_{1-\alpha/2+\hat{\xi}/2}) + O_p(n^{-2}) \tag{17}$$

so that

$$\hat{\xi} - \xi = O_p(n^{-3/2}). \tag{18}$$

Finally, consider the coverage probability of the double-bootstrap CI $(\hat{w}_{\alpha/2-\hat{\xi}/2}, \hat{w}_{1-\alpha/2+\hat{\xi}/2})$. Note that by (12), the Taylor expansion

$$z_{\gamma+\epsilon} = z_\gamma + \epsilon\phi(z_\gamma)^{-1} + O(\epsilon^2), \tag{19}$$

and the derivations for (3.36) in Hall (1992), we have

$$\begin{aligned}
&\mathbf{P}(\theta_0 \in (-\infty, \hat{w}_{1-\alpha/2+\hat{\xi}/2})) \\
=&\mathbf{P}(\sqrt{n}(\hat{\theta}-\theta_0)/\hat{\sigma} > -z_{1-\alpha/2+\hat{\xi}/2} - n^{-1/2}\hat{p}_1(z_{1-\alpha/2+\hat{\xi}/2}) - n^{-1}\hat{p}_2(z_{1-\alpha/2+\hat{\xi}/2}) + \ldots)) \\
=&\mathbf{P}(\sqrt{n}(\hat{\theta}-\theta_0)/\hat{\sigma} > -z_{1-\alpha/2+\xi/2} - n^{-1/2}\hat{p}_1(z_{1-\alpha/2+\xi/2}) - \frac{1}{2}(\hat{\xi}-\xi)\phi(z_{1-\alpha/2})^{-1} - \\
&n^{-1}\hat{p}_2(z_{1-\alpha/2+\xi/2}) + O_p(n^{-2})) \\
=&\mathbf{P}(\sqrt{n}(\hat{\theta}-\theta_0)/\hat{\sigma} > -z_{1-\alpha/2+\xi/2} - n^{-1/2}\hat{p}_1(z_{1-\alpha/2+\xi/2}) - n^{-1}\hat{p}_2(z_{1-\alpha/2+\xi/2}) + \ldots + \\
&\frac{1}{2}(\hat{\xi}-\xi)\phi(z_{1-\alpha/2})^{-1}) + O(n^{-2}) \\
=&\mathbf{P}(\theta_0 < \hat{w}_{1-\alpha/2+\xi/2}) + n^{-3/2}bz_{1-\alpha/2}\phi(z_{1-\alpha/2}) + O(n^{-2})
\end{aligned} \tag{20}$$

where the constant b is defined through

$$\mathbf{E}[\sqrt{n}(\hat{\theta}-\theta_0)/\hat{\sigma}n^{3/2}(\hat{\xi}-\xi)/2] = b + O(n^{-1}). \tag{21}$$

The $O(n^{-1})$ term is derived as in equation (3.35) in Hall (1992) (page 100). Similarly,

$$\mathbf{P}(\theta_0 \in (-\infty, \hat{w}_{\alpha/2-\hat{\xi}/2})) = \mathbf{P}(\theta_0 \in (-\infty, \hat{w}_{\alpha/2-\xi/2})) - n^{-3/2}bz_{\alpha/2}\phi(z_{\alpha/2}) + O(n^{-2}) \tag{22}$$

Now

$$\begin{aligned}
&\mathbf{P}(\theta_0 \in (\hat{w}_{\alpha/2-\hat{\xi}/2}, \hat{w}_{1-\alpha/2+\hat{\xi}/2})) \\
=&\mathbf{P}(\theta_0 \in (-\infty, \hat{w}_{1-\alpha/2+\hat{\xi}/2})) - \mathbf{P}(\theta_0 \in (-\infty, \hat{w}_{\alpha/2-\hat{\xi}/2})) \\
=&\mathbf{P}(\theta_0 \in (-\infty, \hat{w}_{1-\alpha/2+\xi/2})) + n^{-3/2}bz_{1-\alpha/2}\phi(z_{1-\alpha/2}) + O(n^{-2}) \\
&- (\mathbf{P}(\theta_0 \in (-\infty, \hat{w}_{\alpha/2-\xi/2})) - n^{-3/2}bz_{\alpha/2}\phi(z_{\alpha/2}) + O(n^{-2})) \\
=&1-\alpha+O(n^{-2}),
\end{aligned} \tag{23}$$

which concludes our proof.

References

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika 74*(3), 457–468.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association 83*(403), 687–697.

Booth, J. and B. Presnell (1998). Allocation of monte carlo resources for the iterated bootstrap. *Journal of Computational and Graphical Statistics 7*(1), 92–112.

Booth, J. G. and P. Hall (1994). Monte carlo approximation and the iterated bootstrap. *Biometrika 81*(2), 331–340.

Buja, A., R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zhao, and K. Zhang (2015). Models as approximationsa conspiracy of random regressors and model deviations against classical inference in regression.

Buja, A., R. Berk, L. Brown, E. George, M. Traskin, K. Zhang, and L. Zhao (2015). Models as approximations - a conspiracy of random regressors and model deviations against classical inference in regression.

Cribari-Neto, F., T. C. Souza, and K. L. Vasconcellos (2007). Inference under heteroskedasticity and leveraged data. *Communications in StatisticsTheory and Methods 36*(10), 1877–1888.

Eddelbuettel, D., R. François, J. Allaire, J. Chambers, D. Bates, and K. Ushey (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software 40*(8), 1–18.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics 9*(2), 139–158.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association 78*(382), 316–331.

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

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*, Volume 57. CRC press.

Freedman, D. A. (1981, 11). Bootstrapping regression models. *Ann. Statist. 9*(6), 1218–1228.

Gonçalves, S. and H. White (2005). Bootstrap standard error estimates for linear regression. *Journal of the American Statistical Association 100*(471), 970–979.

Hall, P. (1986). On the bootstrap and confidence intervals. *The Annals of Statistics*, 1431–1452.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, 927–953.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer.

Hall, P. and M. A. Martin (1988). On bootstrap resampling and iteration. *Biometrika 75*(4), 661–671.

Kline, P. and A. Santos (2012). Higher order properties of the wild bootstrap under misspecification. *Journal of Econometrics 171*(1), 54–70.

Lee, S. M. S. and G. A. Young (1999). The effect of monte carlo approximation on coverage error of double-bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(2), 353–366.

MacKinnon, J. G. (2006). Bootstrap methods in econometrics*. *Economic Record 82*(s1), S2–S18.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In *Recent advances and future directions in causality, prediction, and specification analysis*, pp. 437–461. Springer.

Portnoy, S. (1986). On the central limit theorem in r^p when $p \to \infty$. *Probability Theory and Related Fields 73*(4), 571–583.

Scholz, F. (2007). The bootstrap small sample properties. *Boeing Computer Services, Research and Technology, Tech. Rep*.

Spokoiny, V. and M. Zhilova (2014). Bootstrap confidence sets under a model misspecification. *arXiv preprint arXiv:1410.0347*.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.