Post-Selection Inference

Arun K. Kuchibhotla,[1] John E. Kolassa,[2] and Todd A. Kuffner[3]

[1]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, United States, 15232; email: arunku@cmu.edu
[2]Department of Statistics, Rutgers University, Piscataway, United States, 08854, email: kolassa@stat.rutgers.edu. Supported in part by NSF DMS-1712839.
[3]Department of Mathematics & Statistics, Washington University, St. Louis, United States, 63130, email: kuffner@wustl.edu. Supported in part by NSF DMS-1712940.

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–24

https://doi.org/10.1146/((please add article doi))

Keywords

Data transformations, Exploratory data analysis, Model selection, Post-selection inference, Sample splitting, Selective inference.

Abstract

We discuss inference after data exploration, with a particular focus on inference after model or variable selection. We review three popular approaches to this problem: sample splitting, simultaneous inference, and conditional selective inference. For each approach, we explain how it works, and highlight its advantages and disadvantages. We also provide an illustration of these post-selection inference approaches.

1. Introduction

The classical inferential theory of mathematical statistics is based on the philosophy that all the models to fit, all the hypotheses to test, and all the parameters to do inference for are fixed prior to seeing the data. This is not how statistics is practiced. The analyst often explores the data to find the "right" model to fit to the data, the "right" hypothesis to test, and so on. As Ronald Coase once said (Tullock 2001, page 205),

> if you torture the data long enough, it will confess.

Once the data has been explored to find the hypothesis or model, the assumptions of a fixed model and fixed hypothesis are no longer appropriate. Classical inference procedures may no longer have the properties established by classical theory. This can invalidate inferences, nullifying the claimed error rates or interpretations. Test statistics and estimators may exhibit completely different distributions than what classical theory prescribes. There can arise biases in estimation caused by data exploration. Procedures designed to control false discovery rates may no longer achieve the desired error control. Power calculations which do not account for data exploration should be viewed suspiciously. The selection of any aspect of a model or hypothesis using the data introduces sampling variability into the model or hypotheses, rendering random the specification process itself.

Many authors, e.g. Benjamini et al. (2009) and Gelman and Loken (2014), attribute this failure of expected behavior of inferential processes as contributing to the failure of scientific replicability, which is considered important by the American Statistical Association (Kafadar 2021).

The potential problems for classical inference procedures arising from model selection or data exploration procedures have long been acknowledged. For example, in the context of variable selection, Hotelling (1940) warned against the "fallacies of selection among numerous results of that one which appears most significant and treating it as if it were the only one examined." Breiman (1992) referred to this as the "quiet scandal in the statistical community."

Post-selection inference has a long and rich history, and the literature has grown beyond what can reasonably be synthesized in our review. Our selection of topics and references should not be misconstrued as a judgment about the relative merits of contributions. Rather than embarking on a futile attempt at being comprehensive, we have chosen a subset of topics that can be coherently presented and that we feel will be of greatest interest to practitioners.

For the purposes of this review, we consider only the setting where the analyst genuinely believes there is model uncertainty, and therefore uses the data to select a model to be used for subsequent inference. There is an equally vast literature on inference for fixed, high-dimensional parameters defined by a linear model containing the full set of observed covariates. In that high-dimensional inference paradigm, what we call model selection is alternatively viewed as dimension reduction or regularization, yielding a lower-dimensional approximation to the original model in the sense of having fewer covariates, and hence a lower-dimensional parameter. In this latter paradigm, post-regularization or post-dimension reduction inference is sought for the full, often high-dimensional, parameter, based on a lower-dimensional approximating model. Within this framework, one may consider either inference for the original parameter, or inference for its appropriately defined representation in the lower-dimensional approximating model. Since the dimension reduction is not considered as selecting a model and its corresponding parameters for subsequent inference, this framework represents an alternative view of what are the relevant inferential targets. For more discussion of the differences, see the Appendix of Berk et al. (2013).

We consider frequentist post-selection inference in this review. The literature concerning Bayesian post-selection inference is comparatively small and authors are not in agreement about many fundamental issues that are essential to studying potential selection effects and correcting for them. Some notable developments include Yekutieli (2012) and Rasines and Young (2020); see the references therein. Selecting a single model for inference could even be considered non-Bayesian according to some interpretations of Bayesian orthodoxy, in the sense that the posterior distribution on the model space, as well as the posterior distributions for all candidate model parameters, constitute a more complete representation of posterior uncertainty than reporting the posterior only for a selected model.

We present post-selection inference as an example of the more general problem of providing Valid Inference after Data Exploration (VIDE). This includes inference after variable selection using, e.g., correlation plots, lasso, or residual diagnostics (Moore and McCabe 1998, Pardoe 2008, Whittingham et al. 2006, Cole 2020). Other than variable selection, data exploration can also include methods to choose a transformation for variables (Harrison and Rubinfeld 1978, Stine and Foster 2013, Weisberg 2005, Liquet and Riou 2013) or cut-off points for discretizing variables (Liquet and Commenges 2001). These widely-used data exploration methods are rarely accounted for when drawing statistical conclusions in practice.

In Section 2, we formulate the post-selection inference problem. In Section 3, we discuss three prominent solutions to VIDE in the literature: sample splitting, simultaneous, and conditional selective inference. In the context of post-selection inference, we discuss their advantages and disadvantages. Examples are presented for each approach. An on-line supplement performs calculations utilizing `R` packages. In Section 4, we consider uniform validity of these approaches and discuss the impossibility results of Leeb and Pötscher (2006). Finally, in Section 5, we consider the implications for practical data analysis.

1.1. Notation

In this paper, we use the following notation. The set of real numbers is denoted by $\mathbb{R}$ and the set of p-dimensional vectors of real numbers is denoted by $\mathbb{R}^p$. Convergence in distribution of a sequence of random variables/vectors T_n to T is denoted by $T_n \overset{d}{\to} T$. Convergence in probability of a sequence of random variables/vectors T_n to T is denoted by $T_n \overset{P}{\to} T$. A sequence of random variables T_n converging in probability to zero is also written as $T_n = o_p(1)$. We write $a := b$ to define a to be a quantity taking the value of b. Expectation and variance of a random variable/vector X are denoted by $\mathbb{E}[X]$ and $\mathrm{Var}(X)$. For any function $f : \mathcal{X} \to \mathbb{R}$, we denote the global minimizer of f by $\arg\min_{x \in \mathcal{X}} f(x)$. The coordinate-wise inequality between two vectors $a, b \in \mathbb{R}^p$ is denoted by $a \preceq b$, i.e., $a_j \leqslant b_j$ for all $j = 1, \ldots, p$ with a_j, b_j representing the j-th coordinates of a, b.

2. Formulation of the problem

The common practice of data analysis may be described as follows: Start with a question of interest, obtain a dataset, explore the data to find a suitable model or find the subset of covariates or find the transformations for variables, then fit the model to draw inferences or statistical conclusions. For example, in the context of fitting a linear regression with a treatment variable, the question of interest could be "is there a non-zero treatment effect?" In the presence of confounders, one might select a subset of confounders to be used in the final model, or one might select a transformation for the response/confounders. Then one fits the model with the selected set of confounders and transformations.

A mathematical formulation in the case of linear regression could be as follows. Suppose we have observations $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$; these need not be independent or identically distributed.

1. For each $\mathrm{M} \subseteq \{1, \ldots, p\}$ corresponding to indices of covariates, define the "target" of estimation by
$$\beta_{\mathrm{M}} \;:=\; \underset{\theta \in \mathbb{R}^{|\mathrm{M}|}}{\arg\min}\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[(Y_i - X_{i,\mathrm{M}}^{\top}\theta)^2\right], \qquad 1.$$
where $X_{i,\mathrm{M}}$ is the subvector of X_i with indices M of covariates.
2. Based on the data, select a subset $\widehat{\mathrm{M}} \subseteq \{1, \ldots, p\}$ of covariates using a method of the analyst's choice. The selection procedure could be formal (lasso, AIC, BIC, marginal screening), informal (correlation plots, residual diagnostics), or even post-hoc (such as changing the model because the conclusion is unexpected).
3. Calculate the estimator
$$\widehat{\beta}_{\widehat{\mathrm{M}}} \;:=\; \underset{\theta \in \mathbb{R}^{|\widehat{\mathrm{M}}|}}{\arg\min}\; \frac{1}{n}\sum_{i=1}^{n} (Y_i - X_{i,\widehat{\mathrm{M}}}^{\top}\theta)^2. \qquad 2.$$
This estimator "targets" $\beta_{\widehat{\mathrm{M}}}$ (the evaluation of the map $\mathrm{M} \mapsto \beta_{\mathrm{M}}$ at $\mathrm{M} = \widehat{\mathrm{M}}$).
4. A VIDE approach to inference for $\beta_{\widehat{\mathrm{M}}}$ based on $\widehat{\beta}_{\widehat{\mathrm{M}}}$ is to construct a valid confidence region $\widehat{\mathrm{CI}}_{\widehat{\mathrm{M}}}$, i.e., one that satisfies
$$\liminf_{n \to \infty} \mathbb{P}\left(\beta_{\widehat{\mathrm{M}}} \;\in\; \widehat{\mathrm{CI}}_{\widehat{\mathrm{M}}}\right) \;\geqslant\; 1 - \alpha, \qquad 3.$$
for the selection method leading to $\widehat{\mathrm{M}}$. In this context, the adjective "valid" means both that the intended nominal error rate α of the procedure for constructing such a confidence region is correct, which would require that the distribution used for the probability calculation is correct asymptotically, and that this error rate is correct for confidence regions $\widehat{\mathrm{CI}}_{\widehat{\mathrm{M}}}$ constructed by this procedure for any $\beta_{\widehat{\mathrm{M}}}$.

The selected set of covariates $\widehat{\mathrm{M}}$ is random through the data and hence potentially changes with the sample size n. For notational simplicity, we do not index $\widehat{\mathrm{M}}$ (and other selections below) with the sample size n.

Selection of variables is only one of many outcomes of data exploration. As described above, variable transformation can also be seen as an outcome. For each transformation $g : \mathbb{R} \to \mathbb{R}$, define the "target"
$$\beta_g \;:=\; \underset{\theta \in \mathbb{R}^p}{\arg\min}\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\{[g(Y_i) - X_i^{\top}\theta]^2\right\}. \qquad 4.$$
Similarly, the estimator $\widehat{\beta}_g$ is obtained as the minimizer of $n^{-1}\sum_{i=1}^{n}[g(Y_i) - X_i^{\top}\theta]^2$. Based on the data, the analyst chooses a transformation $\widehat{g} \in \mathcal{G}$ from a class of transformations. The class of Box-Cox transformations is one such example: $\{y \mapsto (y^{\lambda} - 1)/\lambda : \lambda \neq 0\}$. The

VIDE problem in this case is to construct a valid confidence region $\widehat{\mathrm{CI}}_{\hat{g}}$ for $\beta_{\hat{g}}$ in that it satisfies

$$\liminf_{n\to\infty} \mathbb{P}\left(\beta_{\hat{g}} \;\in\; \widehat{\mathrm{CI}}_{\hat{g}}\right) \;\geqslant\; 1-\alpha, \tag{5.}$$

for the selection method leading to $\hat{g} \in \mathcal{G}$.

The VIDE problems 3. and 5. represent the prototypical problems we will consider. Extensions are possible to logistic, Poisson, and Cox regression models. An even more general VIDE problem can be described as follows. Suppose $Z_1, \ldots, Z_n$ are observations taking values in a set $\mathcal{Z}$. Consider a universe $\mathcal{Q}$ of all possible selections and for every $q \in \mathcal{Q}$ define the estimator

$$\hat{\theta}_q \;:=\; \operatorname*{arg\,min}_{\theta\in\Theta_q} \frac{1}{n}\sum_{i=1}^{n} \ell_q(\theta, Z_i),$$

for a loss function $\ell_q(\cdot,\cdot)$ and a "parameter" set Θ_q that might depend on q. The data analyst can now choose an element $\hat{q} \in \mathcal{Q}$ and the inference is to be based on the estimator $\hat{\theta}_{\hat{q}}$. The VIDE problem is to construct a confidence region $\widehat{\mathrm{CI}}_{\hat{q}}$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\theta_{\hat{q}} \;\in\; \widehat{\mathrm{CI}}_{\hat{q}}\right) \;\geqslant\; 1-\alpha, \tag{6.}$$

for the selection method leading to $\hat{q} \in \mathcal{Q}$. Here the "target" $\theta_{\hat{q}}$ is defined as the evaluation of the map $q \mapsto \theta_q$, at $q = \hat{q}$, given by

$$\theta_q \;:=\; \operatorname*{arg\,min}_{\theta\in\Theta_q} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell_q(\theta, Z_i)]. \tag{7.}$$

Covariate selection and transformation selection can be seen as special cases.

- For covariate selection, take $Z_i = (X_i, Y_i)$, $\mathcal{Q} = \{\mathrm{M} : \mathrm{M} \subseteq \{1, \ldots, p\}\}$, for $q = \mathrm{M} \in \mathcal{Q}$, $\Theta_q = \mathbb{R}^{|\mathrm{M}|}$, and $\hat{\theta}_q = \hat{\beta}_{\mathrm{M}}$.
- For covariate selection, one can also take $Z_i = (X_i, Y_i)$, $\mathcal{Q} = \{\mathrm{M} : \mathrm{M} \subseteq \{1, \ldots, p\}, |\mathrm{M}| \leqslant k\}$. This represents selecting at most k covariates out of p covariates. See Berk et al. (2013, Section 4.5) for more examples.
- For transformation selection, take $\mathcal{Q} = \{g : \mathbb{R} \to \mathbb{R} : g \in \mathcal{G}\}$, for $q = g \in \mathcal{G}$, $\Theta_q = \mathbb{R}^p$, and $\hat{\theta}_q = \hat{\beta}_g$.

In the formulation of the problem, we have not assumed any parametric model for the data. The targets defined in 1., 4. and 7. can be called misspecification-robust targets. They are well-defined even if no parametric model is correct for the data. Further, if the parametric model is correct, then these targets match the usual parametric targets.

The targets in 1., 4. and 7. have different meanings for different values of M, g, and q. More concretely, in the context of variable selection, β_{M_1} and β_{M_2} for $\mathrm{M}_1 = \{1, 2\}$ and $\mathrm{M}_2 = \{1, 3\}$ have different meanings. For example, the first coordinate of β_{M_1}, $\beta_{1\cdot\mathrm{M}_1}$, is the population partial correlation of the response and the first covariate X_1 when adjusted for X_2, while the first coordinate of β_{M_2}, $\beta_{1\cdot\mathrm{M}_2}$, is the partial correlation of β_{M_2} is the population partial correlation of the response and the first covariate X_1 when adjusted for X_3. In general, $\beta_{1\cdot\mathrm{M}_1} \neq \beta_{1\cdot\mathrm{M}_2}$ and they may not even have the same sign. The same logic goes through for β_g as different transformations g.

The major hurdle to solving the VIDE problem is that the estimator $\hat{\theta}_{\hat{q}}$ with a data-driven choice of $\hat{q}$ is random also through $\hat{q}$. In most cases, for every fixed q, $\hat{\theta}_q$ behaves "nicely", i.e. it is asymptotically normal at a $\sqrt{n}$-rate with mean zero and some finite variance depending on q. Because of data exploration, $\widehat{\theta}_{\widehat{q}}$ in general does not have a normal distribution and can be quite biased, even asymptotically.

Figure 1 shows the distribution of the ordinary least squares estimator under forward stepwise selection in a Monte Carlo experiment. The simulation setting is as follows: the covariate vector $X = (X_1, X_2, X_3)$ is multivariate Gaussian with mean zero and a non-diagonal covariance matrix. The response Y is generated from a normal distribution with mean 1 and variance 9, independently of X, so the population coefficients (except the intercept) for linear regression of Y on any subset of covariates are zero. We select from the three covariates by first running a forward stepwise regression. The final model $\widehat{\mathrm{M}}$ is the one with the smallest C_p criterion. Figure 1 shows the histogram of the estimated coefficients of X_1 when fitting the estimated linear model for Y on $X_{\widehat{\mathrm{M}}}$. A density estimate is also laid over the histogram. The histogram of the coefficient of X_1 is drawn only from replications where $\widehat{\mathrm{M}}$ contains 1. A naive analyst who ignores the selection might use the normal distribution as an approximation to the distribution of $\widehat{\beta}_1$ when the selected model contains X_1. Figure 1 shows that such an approximation can be very misleading. The bimodal distributions shown in Figure 1 are expected because X_1 is selected by the variable selection strategy only when it has a reasonably large coefficient in absolute value. This is depicted in Figure 1 with the histogram spread away from zero.

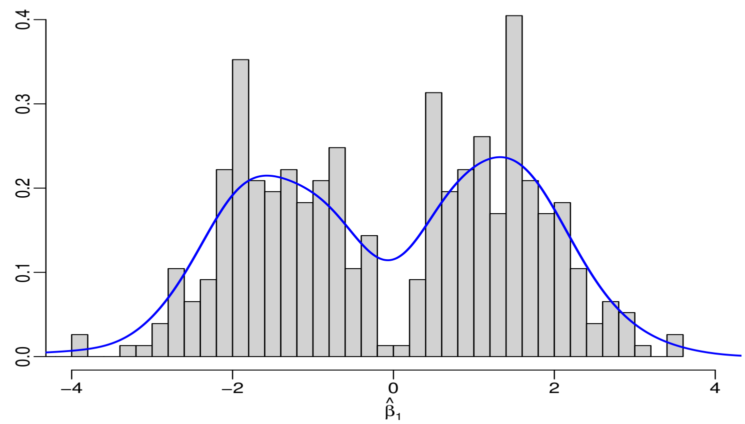

Figure 1: Distribution of $\widehat{\beta}_1$ under forward stepwise selection.

3. Approaches to post-selection inference

Approaches that attempt to provide solutions to VIDE can be characterized by the following terms, to be explained below,

1. sample splitting,
2. simultaneous inference, and
3. conditional selective inference.

These approaches increasingly restrict the selection method. To illustrate them, we use the

Boston housing data available in R package MASS. This data set was introduced in Harrison and Rubinfeld (1978) to understand the impact of air pollution (measured as concentration of nitrogen oxide (nox) on the median value (medv) of houses in different census tracts in Boston. This effect is estimated when adjusting with other covariates including crime rate (crim), proportion of land zoned for lots (zn), vicinity of Charles river (chas), number of rooms (rm), proportion of non-retail business acres per town (indus), proportion of owner-occupied units built prior to 1940 (age), weighted distances to five Boston employment centres (dis), index of accessibility to radial highways (rad), full-value property-tax rate per $10,000 (tax), pupil-teacher ratio by town (ptratio), proportion of African-Americans (black), and % lower status of the population (lstat).

3.1. Sample splitting

A classical and possibly the oldest solution for VIDE problems is sample splitting; see Rinaldo et al. (2019) for a brief history. The basic idea is to split the sample into two parts: training and test data. These could be of different sizes but are usually taken to be of almost equal sizes. First, the training data is used to explore the data and select $\widehat{q}$. Once the selection is made, one ignores the training data and computes the estimator $\widehat{\theta}_{\widehat{q}}$ based on the test data with $\widehat{q}$ from the training data. In this context, one division of the data is made, one model is selected, and standard inferential techniques are applied once. This procedure is different from other procedures, such as the jackknife and cross-validation, that repeatedly split the sample. Because $\widehat{q}$ is independent of the test data when the sample consists of independent observations, $\mathbb{P}(\widehat{\theta}_{\widehat{q}} - \theta_{\widehat{q}} \in A \mid \widehat{q} = q) = \mathbb{P}(\widehat{\theta}_q - \theta_q \in A)$ for all Borel sets A, i.e. the usual asymptotics work on the test data as if no selection was performed. A detailed presentation of sample splitting as a solution of VIDE was given in Zhang (2012, Chapter 2). Sample splitting in light of increasing dimension is discussed in Rinaldo et al. (2019).

3.1.1. Advantages.

One major advantage of sample splitting in comparison to the other two methods we discuss is the generality it allows on selection. There are no assumptions or restrictions on the selection procedure provided it uses only the training data. If the training and test data are approximately the same size, then the sample splitting confidence intervals are at most $\sqrt{2}$ times wider than those ignoring the selection, provided $\sqrt{n}(\widehat{\theta}_q - \theta_q)$ has a limiting distribution for every $q \in \mathcal{Q}$. Hence, if sample splitting applies, it would be recommended for reporting most statistically valid results.

3.1.2. Disadvantages.

The two main disadvantages of sample splitting in comparison to the other approaches we consider are:

- Sample splitting, in conjunction with some model selection procedure such as stepwise, might select a set of variables violating the analyst's "criterion" in the sense that a selected model may exhibit parameter estimates that are inconsistent with known mechanisms underlying the process generating the data. It is difficult to consistently apply sample splitting in a way that avoids unacceptable models.
- Sample splitting is invalid for dependent data. It inherently assumes independence of observations in the data. If the observations are dependent then sample splitting is invalid and no such simple alternative yet exists. Dependent data can easily be accommodated in the simultaneous inference method. Recently, Lunde (2019) proved that sample splitting

guarantees can be extended to weakly dependent data. The subject, however, is not mature enough to apply the results for a wide range of dependent data.

There are other more minor issues with sample splitting. The effect of split sizes is not understood in many problems and there is no clear guidance for choosing the splits. The randomness also causes trouble with interpretation, since with a change in the split sample there can be a change in the selection and hence the target of estimation. This effect of randomness is different from that of the randomness in bootstrap or subsampling, where the randomness disappears with the number of replications diverging. The quantity being estimated using test data changes with every split sample.

3.1.3. Application to the Boston housing example. We apply sample splitting, and other VIDE approaches described below, to the Boston housing data. The dataset was randomly split in half, with one subsample used for training and the other to be used for testing. This particular split only chooses 10 covariates instead of the 11 selected based on the full data. Table **??** contains incorrect p-values resulting from stepwise regression applied to the training set, and p-values correctly calculated from the test set after model selection using the training set.

All covariates selected are significant at level 0.05. The difference in inference implied in the two columns of Table **??** points to a drawback in model splitting, in that the model selected by the training sample may not match that based on the full data. One should not compare the p-values from sample splitting to those in the model selected from the full data. The p-values from training data are in general much smaller than those in the testing data, indicating spurious significance; the test data must be used for inference on the selected model.

3.2. Simultaneous inference approach to VIDE

The simultaneous inference approach, or the uniform inference approach, is proposed by Berk et al. (2013) and extended by Bachoc et al. (2020). The basic idea is to express valid post-selection inference as a simultaneous inference problem. Suppose $\{\theta_q : q \in \mathcal{Q}\}$ are real-valued parameters (or functionals) indexed by the elements of $\mathcal{Q}$. Based on the data, the analyst selects $\widehat{q} \in \mathcal{Q}$ and uses $\widehat{\theta}_{\widehat{q}}$ as an estimator of $\theta_{\widehat{q}}$. To form a confidence region for $\theta_{\widehat{q}}$, the simultaneous inference approach constructs the set of confidence regions $\{\widehat{\mathrm{CI}}_q : q \in \mathcal{Q}\}$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\bigcap_{q\in\mathcal{Q}} \left\{\theta_q \in \widehat{\mathrm{CI}}_q\right\}\right) \geqslant 1-\alpha, \qquad 8.$$

which implies for any $\widehat{q} \in \mathcal{Q}$ that

$$\liminf_{n\to\infty} \mathbb{P}\left(\theta_{\widehat{q}} \in \widehat{\mathrm{CI}}_{\widehat{q}}\right) \geqslant 1-\alpha, \qquad 9.$$

because for any $\widehat{q} \in \mathcal{Q}$,

$$\mathbb{P}\left(\theta_{\widehat{q}} \in \widehat{\mathrm{CI}}_{\widehat{q}}\right) \geqslant \mathbb{P}\left(\bigcap_{q\in\mathcal{Q}} \left\{\theta_q \in \widehat{\mathrm{CI}}_q\right\}\right). \qquad 10.$$

This bound can be conservative because the coverage guarantee is given for all models but is needed only for one *selected* model. Setting this aside for the moment, simultaneous inference has several interesting features.

- Simultaneity implies valid confidence guarantees for *arbitrary* selection procedures $\widehat{q}$, i.e., it does not restrict the practitioner except for the requirement $\widehat{q} \in \mathcal{Q}$.
- Simultaneity implies *infinite* revisions of the selection. For example, one can perform an initial selection, perform inference, and if this is not as expected, one can perform another selection procedure on the data and proceed without any further correction.
- Simultaneity also guarantees validity if multiple models are reported. This is a common occurrence in social sciences where the same question is investigated with several models and a significant outcome in all of them is seen as strengthening the conclusion.

Getting back to the conservativeness of the simultaneous approach, one can always construct a selection procedure $\widehat{q} \in \mathcal{Q}$ such that 10. is an equality; see Theorem 3.1 of Kuchibhotla et al. (2020). This implies that if valid inference is required for an arbitrary selection procedure, then one must perform simultaneous inference.

We now consider simultaneous inference. A generic method for obtaining simultaneous confidence sets is based on the assumption of uniform linear representation of the estimators around the target. This means that for the estimators $\{\widehat{\theta}_q : q \in \mathcal{Q}\}$ based on observations $Z_1, \ldots, Z_n$, there exist functions $\{\psi_q(\cdot) : q \in \mathcal{Q}\}$ such that

$$\max_{q\in\mathcal{Q}} \left| \Psi_{n,q}^{-1/2} \left(\widehat{\theta}_q - \theta_q - \frac{1}{n} \sum_{i=1}^{n} \psi_q\left(Z_i\right) \right) \right| = o_p\left(\frac{1}{\sqrt{n}}\right), \qquad 11.$$

where $\sum_{i=1}^{n} \mathbb{E}[\psi_q(Z_i)] = 0$ and $\Psi_{n,q} = n^{-1}\sum_{i=1}^{n} \mathrm{Var}[\psi_q(Z_i)]$ for all $q \in \mathcal{Q}$. We call assumption 11. the **Uniform Asymptotic Linear Representation**. Most widely-used estimators satisfy 11. when $\mathcal{Q}$ is a singleton (Kuchibhotla 2018) and the functions $\psi_q(\cdot)$ play the role of influence functions for $\widehat{\theta}_q$ for each $q \in \mathcal{Q}$. Assumption 11. implies that the estimators $\widehat{\theta}_q$ are approximately averages of n random variables, with the approximation errors disappearing uniformly over $q \in \mathcal{Q}$.

There is a rich literature on uniform asymptotic linear representations and they have been used in optimal M-estimation problems. See condition (2.3) of Theorem 2.1 in Arcones (2005) and Sections 10.2, 10.3, and equation (10.25) of Dodge and Jurevckova (2000) for examples where uniform asymptotic linear representations are obtained for a large class of M-estimators indexed by a subset of $\mathbb{R}$, an uncountably infinite index set. Their main goal is to choose a tuning parameter that asymptotically leads to an estimator with the "smallest" variance and to account for this randomness in proving that the resulting estimator has an asymptotic normal distribution with the "smallest" variance.

Assumption 11 can be verified for a selection universe $\mathcal{Q}$ for a large class of M-estimation problems, with mild conditions on the "complexity" $\mathcal{Q}$; see Kuchibhotla et al. (2021) and Kuchibhotla (2018, Sections 7.2, 7.3). Although these works deal specifically with covariate selection, their results can be used with variable transformations or a combination of covariate selection and variable transformations.

For each $q \in \mathcal{Q}$, assumption 11. under (weak) independence of $Z_1, \ldots, Z_n$ and integrability conditions such as the Lindeberg–Feller condition imply that $n^{1/2}\Psi_{n,q}^{-1/2}(\widehat{\theta}_q - \theta_q) \stackrel{d}{\to} N(0,1)$, and if $\mathcal{Q}$ is finite with cardinality bounded by a constant independent of the sample size n, then the vector

$$\left(n^{1/2}\Psi_{n,q}^{-1/2}(\widehat{\theta}_q - \theta_q) : q \in \mathcal{Q}\right) \stackrel{d}{\to} \left(G_q : q \in \mathcal{Q}\right), \qquad 12.$$

for a Gaussian random vector $(G_q : q \in \mathcal{Q})$ satisfying $\mathbb{E}[G_q] = 0$ and $\mathrm{Var}(G_q) = 1$ for all

$q \in \mathcal{Q}$. See, for example, Lemma 2.2 of Bachoc et al. (2020). Hence

$$\max_{q\in\mathcal{Q}} \left| n^{1/2}\Psi_{n,q}^{-1/2}(\widehat{\theta}_q - \theta_q) \right| \ \xrightarrow{d}\ \max_{q\in\mathcal{Q}} |G_q|. \tag{13.}$$

Therefore, for a constant $K_\alpha \geqslant 0$ such that $\mathbb{P}(\max_{q\in\mathcal{Q}} |G_q| \leqslant K_\alpha) = 1 - \alpha$,

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{q\in\mathcal{Q}} \left| n^{1/2}\Psi_{n,q}^{-1/2}(\widehat{\theta}_q - \theta_q) \right| \leqslant K_\alpha \right) \ = \ 1-\alpha. \tag{14.}$$

Equivalently, $\widehat{\mathrm{CI}}_q = [\widehat{\theta}_q - K_\alpha\sqrt{\Psi_{n,q}/n}, \widehat{\theta}_q + K_\alpha\sqrt{\Psi_{n,q}/n}], q \in \mathcal{Q}$, forms a simultaneous confidence region, i.e., it satisfies 11.. Usually, $\Psi_{n,q}$ is unknown and has to be replaced by an estimate $\widehat{\Psi}_{n,q}$ which may be conservative (i.e., asymptotically larger than $\Psi_{n,q}$). One only requires 13. and not the joint distributional convergence 12. for the simultaneous coverage guarantee 14.. This is important because the convergence result 13. can hold even if the cardinality of $\mathcal{Q}$ is growing with the sample size or infinite; see Paulauskas and Račkauskas (1989), Norvaiša and Paulauskas (1991), Chernozhukov et al. (2019, 2014), Kuchibhotla et al. (2021), and Kuchibhotla and Rinaldo (2020). A practical way to estimate the constant K_α and the variances $\Psi_{n,q}$ is via a bootstrap, pseudocode for which is given in Algorithm 1, whose validity for a selection universe $\mathcal{Q}$ of fixed cardinality follows from the results of Bachoc et al. (2020). The validity of the bootstrap when $\mathcal{Q}$ grows with sample size follows from Chernozhukov et al. (2014), Kuchibhotla et al. (2021, Section 4.1) and Belloni et al. (2018). The inference procedure in Algorithm 1 depends on the max-t statistic

$$\max_{q\in\mathcal{Q}} \left| n^{1/2}\widehat{\Psi}_{n,q}^{-1/2}(\widehat{\theta}_q - \theta_q) \right|. \tag{15.}$$

Algorithm 1: Bootstrap Procedure for Simultaneous Inference

Input: Data $Z_1, \ldots, Z_n$, coverage probability $1-\alpha$, and the universe of selection $\mathcal{Q}$.

Output: Simultaneous confidence intervals $\widehat{\mathrm{CI}}_q, q \in \mathcal{Q}$, satisfying 8..

Fix $B \geqslant 1$. For $b = 1, \ldots, B$, generate a bootstrap sample $Z_1^{*,b}, \ldots, Z_n^{*,b}$ from $Z_1, \ldots, Z_n$.

Compute the bootstrap estimators $\widehat{\theta}_q^{*b}$ based on $Z_1^{*b}, \ldots, Z_n^{*b}$ for $b = 1, \ldots, B$ and the bootstrap estimate of $\Psi_{n,q}$ as $\widehat{\Psi}_{n,q} := (B-1)^{-1}\sum_{b=1}^{B}[\sqrt{n}(\widehat{\theta}_q^{*b} - \widehat{\theta}_q)]^2$.

Compute the $(1-\alpha)$ quantile $\widehat{K}_\alpha$ of $T^{*b} := \max_{q\in\mathcal{Q}} |n^{1/2}\widehat{\Psi}_{n,q}^{-1/2}(\widehat{\theta}_q^{*b} - \widehat{\theta}_q)|$, for $b = 1, \ldots, B$.

Return the confidence intervals

$$\widehat{\mathrm{CI}}_q \ = \ \left[\widehat{\theta}_q - \widehat{K}_\alpha \frac{\widehat{\Psi}_{n,q}^{1/2}}{\sqrt{n}}, \widehat{\theta}_q + \widehat{K}_\alpha \frac{\widehat{\Psi}_{n,q}^{1/2}}{\sqrt{n}} \right], q \in \mathcal{Q}. \tag{16.}$$

One can compare the confidence intervals 16. to the unadjusted confidence intervals

$$\widehat{\mathrm{CI}}_q^{\text{unadj}} \ := \ \left[\widehat{\theta}_q - z_{\alpha/2}\frac{\widehat{\Psi}_{n,q}^{1/2}}{\sqrt{n}}, \widehat{\theta}_q + z_{\alpha/2}\frac{\widehat{\Psi}_{n,q}^{1/2}}{\sqrt{n}} \right], \tag{17.}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the $N(0,1)$ distribution. The simultaneous confidence intervals 16. inflate the unadjusted confidence intervals 17. by $\widehat{K}_\alpha/z_{\alpha/2} \geqslant 1$. In general, there is no simple expression for the ratio $\widehat{K}_\alpha/z_{\alpha/2}$, which depends on the correlations of $(G_q : q \in \mathcal{Q})$. In a simple setting, Figure 2 shows the coverage and width comparison of the unadjusted confidence interval 17. and the simultaneous confidence interval 16. in the simulation setting: for $d = 1, \ldots, 100$, we generate 500 observations from $(X_i, Y_i) \sim N_{d+1}(0, I_{d+1})$, the standard Gaussian distribution in $\mathbb{R}^{d+1}$. We select one covariate $\widehat{j} \in \{1, \ldots, d\}$ such that the absolute correlation between Y and $X_{\widehat{j}}$ is maximized; this is same as the first step of forward stepwise selection. For this selection, $\mathcal{Q} = \{1, \ldots, d\}$. We compute confidence intervals based on the slope estimator in the linear regression of Y on $X_{\widehat{j}}$. Figure 2 shows that an increase in the number of covariates d leads to a deterioration in the coverage of the unadjusted interval and hence requires more adjustment, as evidenced by the growth of the ratio of widths.

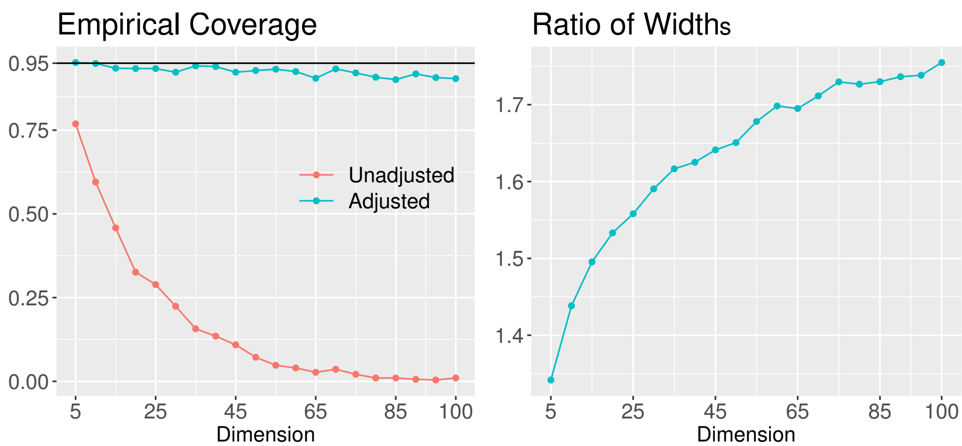

Figure 2: Comparison of unadjusted and simultaneous inference when selecting one covariate out of d. The comparison is based on 1000 replications for each dimension d. The right panel shows the ratio of the width of simultaneous confidence interval 16. to that of the unadjusted confidence interval 17., i.e., $\widehat{K}_\alpha/z_{\alpha/2}$.

The bootstrap procedure used in Algorithm 1 is the classical bootstrap of Efron (1979) which can be replaced by the m-out-of-n bootstrap or wild/multiplier bootstrap (Mammen 1992). The validity guarantee for a growing selection universe $\mathcal{Q}$ follows from the results of Chernozhukov et al. (2013, 2014, 2017) and Belloni et al. (2018); these works contain validity results for both the classical bootstrap and multiplier bootstrap. If the random variables $Z_1, \ldots, Z_n$ are dependent, then the classical bootstrap cannot capture the dependence and for asymptotic validity one must use a version of block bootstrap; see, for example, Zhang and Cheng (2014), and Zhang and Cheng (2018) for a description of the bootstrap and validity results under dependence. In general, subsampling procedures of Politis and Romano (1994) and Politis et al. (1999) provide asymptotic validity. When $\mathcal{Q}$ has infinite cardinality (e.g., Box-Cox variable transformation for the response), it suffices to take an increasingly dense grid of $\mathcal{Q}$ while computing T^{*b} in step 3 of Algorithm 1. In step 2 of Algorithm 1, we use bootstrap replication to estimate the asymptotic variance; this can be skipped if an estimate is otherwise available.

Max-t (in 15.) was one of the first aggregate statistics used for simultaneous inference. Tukey (1949, 1953) used such a statistic for all pairwise differences in ANOVA; in this case, $\mathcal{Q}$ is finite. Scheffé (1953) performed simultaneous inference for all contrasts in the ANOVA model; in this case, $\mathcal{Q}$ is (uncountably) infinite. Both assume a correct parametric model and approximate Gaussianity of the errors. One can use the bootstrap in Algorithm 1 to avoid such restrictions. Both approaches are specific to inference on contrasts of model parameters, and are not directly applicable to inference after model selection.

Any aggregate statistic such as the ℓ_2 or ℓ_p norms could be used instead of the maximum over $q \in \mathcal{Q}$; see Giessing and Fan (2020). Moreover, even with the maximum, there are different possibilities. For example, one can take $\max_{q\in\mathcal{Q}} f_q[|n^{1/2}\Psi_{n,q}^{-1/2}(\widehat{\theta}_q - \theta_q)|]$, for some monotone functions $f_q : \mathbb{R}_+ \to \mathbb{R}_+$; see Kuchibhotla (2020, Chapter 5). Such transformed max-t statistics can be motivated from the idea of balanced confidence intervals (Beran 1988) and using them can lead to significant shortening of the intervals.

3.2.1. Advantages. The simultaneous inference approach has several advantages compared to sample splitting, such as infinite revisions of selection and the ease of reporting inferences from multiple models. Furthermore, it applies to dependent data via subsampling or block bootstrap methods. Because the simultaneous approach allows for selection and inference based on the same data, it can lead to better selection than that from sample splitting. This leads to a trade-off between, respectively, selection and inference properties, when comparing sample splitting and simultaneous approaches; see Rinaldo et al. (2019, Section 3). Finally, simultaneity allows valid inference even when ad hoc selection is done via graphical diagnostics on the full data.

3.2.2. Disadvantages. The simultaneous inference approach requires the specification of $\mathcal{Q}$ *before* exploring the data, i.e., $\mathcal{Q}$ cannot depend on the data. This contrasts with sample splitting, which places no restrictions on $\mathcal{Q}$ provided the selection depends only on the first split. This restriction of simultaneous inference can prohibit its application when data analysis involves sequential modeling, wherein later steps depend on earlier ones and hence $\mathcal{Q}$ can expand without bound. If the selection method $\widehat{q}$ lies in a much smaller subset of $\mathcal{Q}$ with high probability, then the simultaneous approach can lead to conservative confidence intervals, thereby reducing the number of significant results. Finally, simultaneous inference using Algorithm 1 requires computing the estimators $\widehat{\theta}_q$ for all $q \in \mathcal{Q}$. In the context of linear regression with covariate selection, there exists a computationally efficient simultaneous inference procedure; see Kuchibhotla et al. (2020) for details.

3.2.3. Application to the Boston data set. As noted before, the methods of Tukey (1949, 1953) and Scheffé (1953) are appropriate for inference on contrasts. In the Boston housing data, the variable rad is a categorical variable taking 9 different values. *A priori* knowledge of the impact of rad is minimal; because convenience values of closeness to highways is balanced against nuisances associated with highway proximity, one would not expect the effect to be monotonic, let alone linear. In order to explore this effect, one might simultaneously bound all mean valuation differences for houses with differing accessibility to radial highways. For simplicity, all values of rad above 5 are set to 5. There are 20, 24, 38, 110, and 314 towns associated with values of this modified rad of 1 through 5, respectively.

Figure 3 shows the difference in sample means for each pair of values of rad. Simultaneous lower and upper confidence limits are also reported. Such intervals allow one to look at all differences and pick the largest or smallest and make a valid statistical claim. For example, `5-3` yields the most negative difference in sample means, and because the corresponding confidence interval does not contain zero, we can (at level 0.05) conclude that the median house price is different for census tracts with rad 5 and 3, even after taking selection into account. These contrasts can also be tested using the method of Scheffé (1953) (also displayed in Figure 3), but because this provides simultaneous inference over all contrasts (not just pairwise differences) it tends to be less powerful for pairwise differences than the method of Tukey (1949).

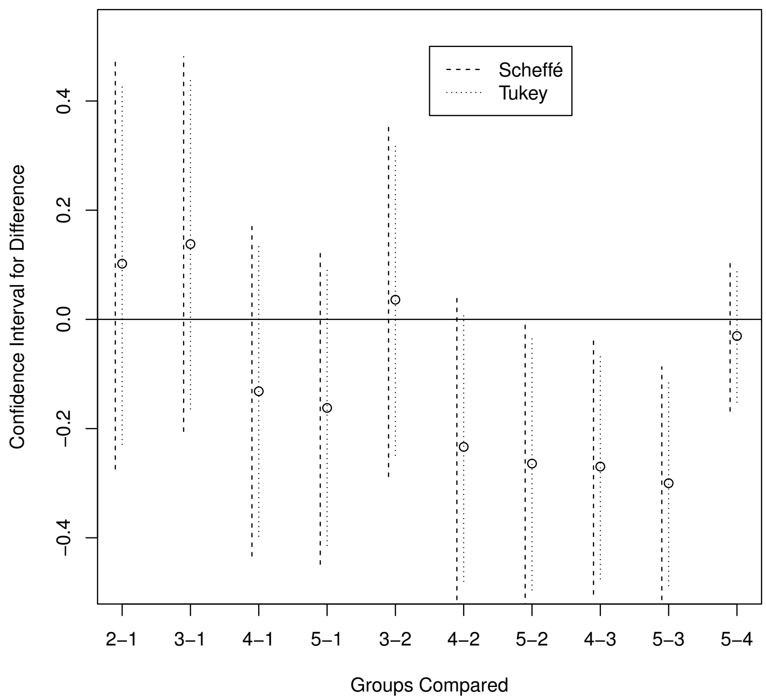

Figure 3: Tukey and Scheffé 95% confidence intervals for median housing value difference by redefined rad.

As mentioned before, Scheffé's test is based on Gaussianity and homoscedasticity assumptions, which might be invalid. Under these assumptions Scheffé's test is less powerful than the output of Algorithm 1 when covariate selection is performed, because Sheffé's method provides simultaneous inference on more contrasts than needed; see Berk et al. (2013, Section 4.8) for a detailed discussion. Algorithm 1 may be used for covariate selection under a well-specified linear model. Table 1 shows the $\widehat{K}_\alpha$ values to be used in 16.. This algorithm requires only the covariate matrix, because of the Gaussian linear model assumption. The PoSI constant shown above under the column `PoSI` is the smallest. Note that the Scheffé constant is also shown in the final column. Without specifying other arguments, the output of this algorithm provides adjustments for the universe of selection

Confidence Level	PoSI	Bonferroni	Scheffé
95%	3.591	4.904	4.729
99%	4.075	5.211	5.262

Table 1: Values of $\widehat{K}_\alpha$ for Various Adjustments for Simultaneous Inference

$\mathcal{Q} = \{(j, \mathrm{M}) : j \in \mathrm{M}, \mathrm{M} \subseteq \{1, \ldots, p\}\}$. Other arguments can be used to reduce the universe; reducing the universe reduces the computational complexity.

To go beyond the linear model assumptions and allow for potentially misspecification, we can use the bootstrap idea in Algorithm 1, which gives a value $\widehat{K}_{0.95} = 4.624$ corresponding to 95% confidence. This approach may also be applied to the regression model with covariates crim and chas to give results in Table 2.

Variable	Lower	Upper
Intercept	21.619	25.609
crim	-0.717	-0.095
chas	-3.864	15.020

Table 2: Confidence Intervals using the method of Berk et al. (2013).

Case studies involving covariate selection and also transformation selection can be found in Cai (2020). Finally, max-t style corrections in other VIDE problems including optimal cut-off detection and transformations are discussed in Liquet and Commenges (2001), Liquet and Riou (2013, 2019).

3.3. Conditional selective inference

The setting here is the same as in Section 3.2. Instead of considering the simultaneous statement as in 8., selective inference constructs $\widehat{\mathrm{CI}}_{\widehat{q}}$ such that for all $q \in \mathcal{Q}$,

$$\liminf_{n\to\infty} \mathbb{P}\left(\theta_{\widehat{q}} \in \widehat{\mathrm{CI}}_{\widehat{q}} \mid \widehat{q} = q\right) \geqslant 1 - \alpha. \qquad 18.$$

Kuffner and Young (2018) explain that conditioning on the selection event can be justified through the Fisherian proposition of relevance. Conventionally, this is achieved by following the conditionality principle, that relevance of the inference to the actual data under study requires the hypothetical repeated sampling to be conditioned on certain features of the observed data. In this case, relevance is achieved by conditioning on the subset of the sample space yielding the particular selection outcome. The construction of $\widehat{\mathrm{CI}}_{\widehat{q}}$ proceeds by approximating the conditional distribution of $\sqrt{n}(\widehat{\theta}_q - \theta_q)$ given $\widehat{q} = q$ for any $q \in \mathcal{Q}$ and computing/estimating the conditional quantile. For simplicity, we restrict our discussion to inference for a univariate target θ_q. The conditional selective inference framework can be understood using the following assumptions. Fix a $q \in \mathcal{Q}$.

(A1) There exists a random vector $D_{n,q} \in \mathbb{R}^{d_D}$ such that $\{\widehat{q} = q\} \equiv \{D_{n,q} \preceq 0\}$, with the symbol $\preceq$ between two vectors representing coordinate-wise inequality. The integer d_D represents the dimension of $D_{n,q}$ and the subscript is used to distinguish this from d, the dimension of covariates in our regression examples.

(A2) The selection event occurs with asymptotically non-zero probability, that is,

$$\liminf_{n\to\infty} \mathbb{P}(D_{n,q} \preceq 0) > 0. \qquad 19.$$

(A3) There exist a vector $\mu_{n,q} \in \mathbb{R}^{d_D}$ and a covariance matrix Ω_q such that

$$\begin{bmatrix} \sqrt{n}(\widehat{\theta}_q - \theta_q) \\ D_{n,q} - \mu_{n,q} \end{bmatrix} \xrightarrow{d} \begin{bmatrix} G_{\theta,q} \\ G_{D,q} \end{bmatrix} \sim N(0, \Omega_q).$$

(A4) There exists a consistent estimator $\widehat{\Omega}_q$ for Ω_q, i.e.,

$$\widehat{\Omega}_q = \begin{bmatrix} \widehat{\omega}_q^2 & \widehat{\Omega}_{\theta D} \\ \widehat{\Omega}_{D\theta} & \widehat{\Omega}_{DD} \end{bmatrix} \xrightarrow{P} \Omega_q = \begin{bmatrix} \omega_q^2 & \Omega_{\theta D} \\ \Omega_{D\theta} & \Omega_{DD} \end{bmatrix}. \qquad 20.$$

These assumptions are modeled after the selective inference framework of Markovic et al. (2017) and McCloskey (2020). All the assumptions relate only to $q \in \mathcal{Q}$ individually. Assumption **(A1)**. requires that the selection of a "model" q can be written in terms of a statistic $D_{n,q}$. The representation in terms of the negative orthant might seem very restrictive, but any inequality of the form $A_q D'_{n,q} \preceq \widehat{a}_{n,q}$ can be written as $A_q D'_{n,q} - \widehat{a}_{n,q} \preceq 0$, so **(A1)**. applies to any "polyhedral" selection event. Condition 19. is equivalent to insisting that the event $\{\widehat{q} = q\}$ occurs with a non-zero probability asymptotically. This has been relaxed in some works, but a condition on how fast the selection probability can converge to zero (Tian and Taylor 2017) is required to ensure that the denominator in the conditional probability 18. converges to its asymptotic counterpart; see 22. for an example. The distributional assumption **(A3)**. implicitly requires that the dimension of $(\widehat{\theta}_q, D_{n,q})$ is fixed as the sample size n diverges to infinity. Assumption **(A4)**. can be easily satisfied by bootstrapping or subsampling the vector $(\sqrt{n}(\widehat{\theta}_q - \theta_q), D_{n,q} - \mu_{n,q})$.

Because $\mu_{n,q}$ in **(A3)**. may depend on the sample size n, we need a "uniform" convergence result in addition to **(A3)**.. Assumption **(A3)**. implies such a uniform convergence result. If $\mathcal{C}$ is the set of all convex sets in $\mathbb{R}^{1+d_D}$, then Theorem 4.2 of Rao (1962) proves that **(A3)**. implies

$$\sup_{C\in\mathcal{C}} \left| \mathbb{P}\left(\begin{bmatrix} \sqrt{n}(\widehat{\theta}_q - \theta_q) \\ D_{n,q} - \mu_{n,q} \end{bmatrix} \in C \right) - \mathbb{P}\left(\begin{bmatrix} G_{\theta,q} \\ G_{D,q} \end{bmatrix} \in C \right) \right| \to 0, \qquad n \to \infty. \qquad 21.$$

Here the set C must be a continuity set for $[G_{\theta,q}^\top \; G_{D,q}^\top]^\top$, as would be true if the covariance matrix Ω_q were positive definite.

Before describing the selective confidence interval, let us provide two simple selection methods to which the framework applies.

3.3.1. Inference on Winners. The following example is discussed in Sampson and Sill (2005), Sill and Sampson (2009), and Andrews et al. (2019). Suppose $X_1, \ldots, X_n$ are independent and identically distributed random vectors in $\mathbb{R}^d$ with mean μ. Consider the selection of a coordinate among $j = 1, \ldots, d$ with the largest mean. In this case, the universe $\mathcal{Q}$ is $\{1, \ldots, d\}$ and the event $\widehat{q} = q$ can be written as

$$\{\widehat{q} = q\} = \left\{ e_j^\top \bar{X}_n \leqslant e_q^\top \bar{X}_n, \qquad j = 1, \ldots, d \right\} = \{A_q \bar{X}_n \preceq 0\},$$

where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $A_q \in \mathbb{R}^{(d-1)\times d}$ is a matrix with rows $\{e_j^\top - e_q^\top : j \neq q\}$. Hence, assumption **(A1)**. is satisfied with $D_{n,q} = \sqrt{n}A_q\bar{X}_n$. Note that $\mathbb{P}(D_{n,q} \preceq 0) = \mathbb{P}[\sqrt{n}(A_q\bar{X}_n - A_q\mu) \preceq -\sqrt{n}A_q\mu]$. Define $\Sigma = \mathrm{Var}(\sqrt{n}\bar{X}_n)$. If $A_q\Sigma A_q^\top$ is non-singular, then by the Berry–Esseen bound for all rectangles in Chernozhukov et al. (2020), we get that

$$\left|\mathbb{P}(D_{n,q} \in \mathcal{S}_{n,q}) - \mathbb{P}(N(0, A_q\Sigma A_q^\top) \preceq -\sqrt{n}A_q\mu)\right| \leqslant \frac{\mathfrak{C}_{X,q}(d)}{\sqrt{n}}, \qquad 22.$$

for a constant $\mathfrak{C}_{X,q}(d)$ depending on the distribution of X, q, and also the dimension d. Hence, inequality 19. holds if $\mathbb{P}[N(0, A_q\Sigma A_q^\top) \preceq -\sqrt{n}A_q\mu]$ stays away from zero as $n \to \infty$. This cannot hold if $-A_q\mu \preceq 0$ and $\|A_q\mu\|_2 = O(1)$ as $n \to \infty$. Assumption **(A3)**. is readily satisfied using the central limit theorem. Here Σ_{qq} is the q-th diagonal element of Σ. Assumption **(A4)**. also holds by replacing Σ in Ω_q by the sample covariance matrix of $X_1, \ldots, X_n$.

3.3.2. Lasso selection. This example was discussed in Lee et al. (2016) and Tibshirani et al. (2018), among others. The lasso selection procedure of Tibshirani (1996) selects a subset of covariates via the optimization problem:

$$\widehat{\beta}^{\texttt{lasso}} \;:=\; \underset{\beta\in\mathbb{R}^p}{\arg\min}\ \frac{1}{2}\sum_{i=1}^n (Y_i - X_i^\top\beta)^2 + \lambda\|\beta\|_1,$$

based on regression data $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$. The lasso estimator $\widehat{\beta}^{\texttt{lasso}}$ has some coefficients that are exactly zero, so the covariates selected are $\widehat{\mathrm{M}} = \{j : \widehat{\beta}_j \neq 0\}$. For a data-independent λ, we consider the selection as selecting covariates and also the signs of the lasso coefficients, which are included for easier expression for the selection event. Thus $\widehat{q} = (\widehat{\mathrm{M}}, \widehat{s})$, where $\widehat{s}$ is the vector of signs of $\widehat{\beta}^{\texttt{Lasso}}$ and $\widehat{\theta}_{\widehat{q}}$ is the ordinary least squares linear regression estimator $\widehat{\beta}_{\mathrm{M}}$ defined in 2.. The analysis of Lee et al. (2016, Theorem 4.3), Markovic et al. (2017, Section 3), and McCloskey (2020, Section 5) shows that the event $\{\widehat{q} = q\} = \{\widehat{\mathrm{M}} = \mathrm{M}, \widehat{s} = s\}$ can be written as $\{A_qD'_{n,q} \preceq \widehat{a}_{n,q}\} = \{A_qD'_{n,q} - \widehat{a}_{n,q} \preceq 0\}$, where

$$A_q := \begin{pmatrix} -\mathrm{diag}(s_{\mathrm{M}}) & 0 \\ 0 & I_{p-|\mathrm{M}|} \\ 0 & -I_{p-|\mathrm{M}|} \end{pmatrix}, \quad D'_{n,q} := \begin{pmatrix} n^{1/2}(\mathbf{X}_{\mathrm{M}}^\top\mathbf{X}_{\mathrm{M}})^{-1}(\mathbf{X}_{\mathrm{M}}^\top\mathbf{Y}) \\ n^{-1/2}\mathbf{X}_{-\mathrm{M}}^\top(I_p - \mathbf{X}_{\mathrm{M}}(\mathbf{X}_{\mathrm{M}}^\top\mathbf{X}_{\mathrm{M}})^{-1}\mathbf{X}_{\mathrm{M}}^\top)\mathbf{Y} \end{pmatrix},$$

and

$$\widehat{a}_{n,q} := \begin{pmatrix} -\lambda n^{1/2}\mathrm{diag}(s_{\mathrm{M}})(\mathbf{X}_{\mathrm{M}}^\top\mathbf{X}_{\mathrm{M}})^{-1}s_{\mathrm{M}} \\ \lambda n^{-1/2}(\mathbf{1}_{p-|\mathrm{M}|} - \mathbf{X}_{-\mathrm{M}}^\top\mathbf{X}_{\mathrm{M}}(\mathbf{X}_{\mathrm{M}}^\top\mathbf{X}_{\mathrm{M}})^{-1}s_{\mathrm{M}}) \\ \lambda n^{-1/2}(\mathbf{1}_{p-|\mathrm{M}|} + \mathbf{X}_{-\mathrm{M}}^\top\mathbf{X}_{\mathrm{M}}(\mathbf{X}_{\mathrm{M}}^\top\mathbf{X}_{\mathrm{M}})^{-1}s_{\mathrm{M}}) \end{pmatrix}.$$

Hence, **(A1)**. holds with $D_{n,q} = A_qD'_{n,q} - \widehat{a}_{n,q}$. It is easy to find $a_{n,q}$ such that $\widehat{a}_{n,q} - a_{n,q}$ converges in probability to zero (McCloskey 2020, Section 5). Assumptions **(A3)**. and **(A4)**. follow readily from moment assumptions and bootstrap/subsampling results. Assumption **(A2)** can be verified using the distributional convergence result. Once again, this assumption may fail.

Lee et al. (2016) consider the problem under a homoscedastic Gaussian model for the response vector $\mathbf{Y}$ and fixed covariates. The analyses in Markovic et al. (2017, Section 3) and McCloskey (2020) allow random covariates and do not require Gaussianity of $\mathbf{Y}$.

Several other covariate selection strategies can be covered under assumptions **(A1)**.–**(A4)**.; see Markovic et al. (2017) and Tibshirani et al. (2018, Lemma 3). These works cover methods such as the cross-validated lasso, forward stepwise regression, least angle regression (LAR), Akaike's information criterion, and the randomized lasso.

3.3.3. Conditional selective inference methodology. Under assumptions **(A1)**.–**(A4)**., a confidence interval satisfying the asymptotic conditional coverage condition 18. can be obtained following Algorithm 2. Under assumptions **(A1)**.–**(A4)**., the confidence interval

---

Algorithm 2: Conditional Selective Inference under Polyhedral Selection

Input: Estimator $\widehat{\theta}_q$, consistent estimator $\widehat{\Omega}_q$, coverage probability $1-\alpha$.

Output: Conditional confidence intervals $\widehat{\mathrm{CI}}_q^{\mathrm{cond}}$ satisfying 18..

Define $\widehat{\Gamma}_q = \widehat{\Omega}_{D\theta}/\widehat{\omega}_q^2$ and $N_{n,q} = D_{n,q} - \sqrt{n}\widehat{\Gamma}_q\widehat{\theta}_q$.

Define

$$\mathcal{V}^- = \max_{j:\,\widehat{\Gamma}_{q,j}<0} \frac{-N_{n,q,j}}{\widehat{\Gamma}_{q,j}}, \qquad \mathcal{V}^+ = \min_{j:\,\widehat{\Gamma}_{q,j}>0} \frac{-N_{n,q,j}}{\widehat{\Gamma}_{q,j}}.$$

Here $N_{n,q,j}$ and $\widehat{\Gamma}_{q,j}$ refer to the j-th coordinate of $N_{n,q}$ and $\widehat{\Gamma}_q$.

Set $F(\cdot;\mu,\sigma^2,\mathcal{L},\mathcal{U})$ to be the cumulative distribution function of a normal distribution with mean μ, variance σ^2 conditional on belonging to $[\mathcal{L},\mathcal{U}]$.

Define $\widehat{L}_{q,\alpha}$ and $\widehat{U}_{q,\alpha}$, respectively, as solutions (in θ) to the equations

$$F(\sqrt{n}\theta;\sqrt{n}\widehat{\theta}_q,\widehat{\omega}_q^2,\mathcal{V}^-,\mathcal{V}^+) = \frac{\alpha}{2}, \qquad F(\sqrt{n}\theta;\sqrt{n}\widehat{\theta}_q,\widehat{\omega}_q^2,\mathcal{V}^-,\mathcal{V}^+) = 1-\frac{\alpha}{2}.$$

return the confidence interval $\widehat{\mathrm{CI}}_q^{\mathrm{cond}} := [\widehat{L}_{q,\alpha},\widehat{U}_{q,\alpha}]$.

---

returned by Algorithm 2. has asymptotic coverage $1-\alpha$; see Tian and Taylor (2017), Markovic et al. (2017). The proof is based on an asymptotic version of a "polyhedral lemma" (Lee et al. 2016). Proposition 1 of McCloskey (2020) (with $\gamma = 0$) provides an alternative coverage guarantee without requiring assumption **(A2)**..

Variations of the conditional selective inference method appear in the literature. The vanilla version described in Examples 3.3.1 and 3.3.2 that considers selection on the whole data without randomization can lead to much wider confidence intervals than the sample splitting and simultaneous approaches. Kivaranovic and Leeb (2018) proved that the vanilla version may yield confidence intervals with infinite width, prompting several modifications that either consider selection based on a part of the data or by explicitly adding randomization to $D_{n,q}$ in selection. This is called data carving (Fithian et al. 2014, Tian and Taylor 2018) and is related to adaptive data analysis in machine learning and computer science. Data carving can be regarded as a combination of sample splitting and vanilla selective inference. Model selection in data carving differs from that in the vanilla version. Kivaranovic and Leeb (2020) prove that, in contrast to the vanilla version, randomized selective inference yields confidence intervals with bounded expected length. Andrews et al. (2019) and McCloskey (2020) combine simultaneous and selective inference; their approach conditions on the event that θ_q lies in a simultaneous confidence interval as well as on the event $\{\widehat{q} = q\}$. This additional conditioning implies that the combined confidence interval will be smaller than the simultaneous confidence interval; see McCloskey (2020) for more details. Finally, there is an approach to conditional selective inference from the Bayesian perspective (Panigrahi et al. 2016). Also, there exist selective inference approaches that can account for convex selection methods (Tian et al. 2016).

3.3.4. Advantages. Conditional selective inference allows for selection based on the whole data, similarly to simultaneous inference and in contrast to sample splitting. It is also computationally more similar to sample splitting than to simultaneous inference. With a good choice of the selective inference method, the resulting selective confidence intervals can vary between the naive unadjusted confidence intervals and the sample splitting confidence intervals; see Figure 4 of Fithian et al. (2014). If the selection event $\{\widehat{q} = q\}$ holds with probability close to one (asymptotically), then there is no need to adjust the naive confidence interval 17.. Unlike both sample splitting and simultaneous inference, the selective inference approach accounts for the specific selection methodology employed by the practitioner.

3.3.5. Disadvantages. The selective inference approach relies heavily on the specific selection methodology used prior to inference. This limits its applicability in practice, and explains why the existence of a general theory of conditional selective inference, which applies beyond the specialized settings where it has been studied, is open. This can be understood from assumption **(A1)**.. Although **(A1)**. holds for several covariate selection methods, it does not accommodate variable transformation and other exploration methods involving graphical tools. Applying the conditional approach to a new selection method requires new theoretical analysis to ensure validity of assumptions; Algorithm 1 and sample splitting can be employed for any selection method and selection universe $\mathcal{Q}$. Also, as mentioned before, the vanilla version of the method can yield much wider confidence intervals than sample splitting and simultaneous inference.

3.3.6. Selective Inference Applied to the Boston Housing Data. When the data follow a Gaussian distribution, then the resulting procedure provides tests with the correct type I error in finite samples; otherwise the guarantees are asymptotic. We begin with an application to stepwise regression. This procedure sequentially adds variables, with the next variable in each case chosen to maximize the increase in the regression sum of squares. This is equivalent to using AIC to select the next variable, but in this case stopping only after examining a certain number of larger models to avoid premature stopping. The p-values and confidence intervals adjusted for stepwise selection are given in Table 3. The forward stepwise implementation in this package selected all covariates instead of 10 variables obtained via the `step` function.

Table 3 shows confidence intervals for linear parameters that are wider than the näive intervals, to correctly allow for the effect of selection.

Selection bias associated with overfitting, as is a well-known problem when selecting variables using AIC, can adversely affect post-selection uncertainty assessments, yielding post-selection predictive and confidence intervals which tend to undercover if selection is not accounted for; see Hong et al. (2018).

One might also consider application of the Lasso. First, apply cross-validation to minimize squared error. Tibshirani et al. (2019) recommend applying the Lasso to centered and scaled covariates. Results are in Table 4.

Effect	Adjusted p value	Lower Bound	Upper Bound
lstat	0.33550	-0.05246	0.053042
ptratio	0.24041	-0.04493	0.042796
crim	0.34545	-0.03413	0.025595
rm	0.32349	-0.24014	0.380779
dis	0.65175	-0.05189	0.301466
nox	0.09967	-4.78435	0.460184
black	0.62735	-Inf	0.007394
rad	0.44520	-0.03969	0.017964
tax	0.04415	-0.00113	-0.000023
chas	0.00387	0.16362	Inf
zn	0.25555	-0.00164	0.002741
indus	0.29829	-0.00957	0.020518
age	0.54697	-0.00418	0.002789

Table 3: Selective Inference applied to the Boston Housing Data. Units are given in the text following Table **??**

Order Entered	Variable	Adjusted p value	Lower Bound	Upper Bound
1	crim	2.49e-11	-0.012091	-0.006957
2	zn	2.72e-01	-0.000901	0.001283
3	age	1.40e-03	0.000578	0.002370
4	rad	5.28e-05	0.009575	0.020953
5	tax	3.19e-05	-0.000995	-0.000443
6	ptratio	0.00e+00	-0.050948	-0.031245
7	black	2.84e-04	0.000181	0.000605
8	lstat	1.35e-60	-0.040679	-0.034145

Table 4: Lasso applied to the Boston Housing Data

4. Honesty and uniform validity

In all the methods discussed in Section 3, we have discussed pointwise (asymptotic) validity, i.e., validity of coverage is required and provided for a given probability distribution of the data that is fixed as the sample size changes. In the context of data exploration, such pointwise asymptotics are known to be misleading, as discussed by Leeb and Pötscher (2005). The requirement of honesty or uniform validity for conditional and unconditional post-selection inference (respectively) can be described as

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}^{\otimes n}} \mathbb{P}\left(\theta_{\hat{q}} \in \widehat{\mathrm{CI}}_{\hat{q}}\right) \geqslant 1-\alpha \quad \text{and} \quad \liminf_{n\to\infty} \inf_{P\in\mathcal{P}^{\otimes n}} \mathbb{P}\left(\theta_{\hat{q}} \in \widehat{\mathrm{CI}}_{\hat{q}} \mid \hat{q} = q\right) \geqslant 1-\alpha. \quad 23.$$

Here $\mathcal{P}^{\otimes n}$ is a subset of all probability distributions for a sample of n observations, often satisfying certain moment conditions and $P \in \mathcal{P}^{\otimes n}$ represents the true distribution of the data. For all the methods described in Section 3, uniform validity holds under regularity conditions on $\mathcal{P}^{\otimes n}$. For sample splitting and the simultaneous approach, uniform validity (first part of 23.) follows from Berry–Esseen bounds, e.g. 22. (Belloni et al. 2018, Rinaldo et al. 2019, Bachoc et al. 2020, Kuchibhotla et al. 2021). For the selective inference approach, uniform validity (second part of 23.) was proved in Tibshirani et al. (2018), Andrews et al. (2019), and McCloskey (2020).

The impossibility results of Leeb and Pötscher (2006, 2008) seem to be at odds with uniform validity of the simultaneous and selective approaches. Before we explain the discrepancy, we describe these impossibility results. Let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ satisfy the linear model $Y = X^\top \beta_0 + \xi$ for $\xi \sim N(0, 1)$, and let $\widehat{\mathrm{M}}$ be a subset of covariates chosen using the data. We have the least squares estimator $\widehat{\beta}_{\widehat{\mathrm{M}}} \in \mathbb{R}^{|\widehat{\mathrm{M}}|}$ from 2.. Define $\widetilde{\beta}_{\widehat{\mathrm{M}}} \in \mathbb{R}^d$ as the augmentation of $\widehat{\beta}_{\widehat{\mathrm{M}}}$ with zeroes for components corresponding to non-selected covariates. In Leeb and Pötscher (2006, 2008), the authors consider estimating $G(t \mid \widehat{\mathrm{M}}) = \mathbb{P}(\sqrt{n}A(\widetilde{\beta}_{\widehat{\mathrm{M}}} - \beta_0) \preceq t \mid \widehat{\mathrm{M}} = \mathrm{M})$ and $G(t) = \mathbb{P}(\sqrt{n}A(\widetilde{\beta}_{\widehat{\mathrm{M}}} - \beta_0) \preceq t)$, respectively, for a given non-random $A \in \mathbb{R}^{s \times d}$ and $t \in \mathbb{R}^s$. Their results imply that no estimator of $G(t \mid \widehat{\mathrm{M}})$ and $G(t)$ can be consistent uniformly over all β_0 satisfying $\|\beta_0 - \beta^\star\|_2 \leqslant Cn^{-1/2}$ (for any fixed $\beta^\star \in \mathbb{R}^d$); note that the data generating distributions in this case are indexed by β_0. As shown in Leeb and Pötscher (2006, Section 2.2), it is possible to construct estimators that are consistent for each $\beta_0 \in \mathbb{R}^d$ (fixed as $n \to \infty$), but the impossibility refers to uniform consistency over all β_0 (in a shrinking neighborhood). With this understanding of the impossibility results, the discrepancy with uniform validity of simultaneous and selective inference can be explained rather easily. The target we use for VIDE differs for different selected models. For instance, in linear regression, our target is defined as $\beta_{\widehat{\mathrm{M}}}$, which is β_{M} in 1. evaluated at $\mathrm{M} = \widehat{\mathrm{M}}$. If $\widehat{\mathrm{M}}_1 = \{1, 2\}$ and $\widehat{\mathrm{M}}_2 = \{1, 3\}$, then the first coordinate of $\beta_{\widehat{\mathrm{M}}_1}$ can be different from that of $\beta_{\widehat{\mathrm{M}}_2}$. They are both coefficients of covariate X_1 but in two different models, as described in Berk et al. (2013). In contrast, the target in Leeb and Pötscher (2006, 2008) is the coefficient vector β_0 in a well-specified full model. This difference in targets is also described in Bachoc et al. (2019), where the VIDE target $\theta_{\widehat{q}}$ is called a non-standard target. This difference is the main cause of impossibility results. Furthermore, the results of Leeb and Pötscher (2006, 2008) only refer to the estimator $\widetilde{\beta}_{\widehat{\mathrm{M}}}$ in the selected model. It is possible to define other estimators for the full model parameter β_0 that use a model selection procedure (such as lasso) while also providing uniformly valid inference; see Belloni et al. (2015, 2016) and Chernozhukov et al. (2015).

These considerations of uniformity are important. Procedures that provide only approximate pointwise error control potentially break down in contexts involving more complex universes of models, and may fail to hold at more difficult parameter values for a fixed model. More difficult here refers to parameter settings where model selection procedures lead to high variability in selection; for example, in a linear regression model with true parameter values around $1/\sqrt{n}$. See Leeb and Pötscher (2005) for a detailed discussion on uniform validity in the context of model selection.

5. What are the implications for statistical practice?

Our current understanding of the scope of the problems caused by selection on subsequent inferences is limited. It is easy to understand why using the data for both selection and inference *may* invalidate subsequent inference methods which pretend that no selection took place, and many papers contain simple simulation experiments to illustrate that naive inference after selection can be misleading or incorrect; see, e.g., Freedman (1983), Freedman (2009, Chapter 5), and Austin et al. (2006). However, there has been little effort to demonstrate that failing to account for selection can have negative effects in high-stakes decisions.

As a community, statisticians need to provide more practical guidance about when it is truly important to account for selection, and when it is likely to make little difference. With all three approaches we presented, there are significant challenges to implementation even in relatively simple linear regression problems with popular variable selection procedures. Researchers in this area have a virtually endless horizon of open problems, as all existing data exploration techniques could be studied again within the post-selection framework, from the perspective of inference, prediction, classification, or other statistical decisions. The mathematical frameworks of both simultaneous and conditional selective inference prohibit their employment in practice, because practical data analysis often tends to be dynamic, with future exploration methods dictated by past explorations of the same data. See, for example, the analysis of the realtor data in Pardoe (2008), or Gelman et al. (2020). Neither the selection universe nor the method of selection is decided before analyzing the data; the data dictate both. Sample splitting is the only general practical solution allowing such dynamic data analysis, but it requires splitting the data only once at the beginning, and only applies to independent data; a general solution for time series or other dependent data is yet to emerge.

If one wants to employ the simultaneous inference techniques discussed in Section 3 in data analysis, then decisions about either the universe or method of selection must be made in advance. This is much like writing a protocol and sticking to it. Even if the protocol is complicated, a selection universe can be created and the simultaneous inference approach 3.2 applies. This yields better model selection (because more data is used) than sample splitting and also provides valid inference.

ACKNOWLEDGMENTS

Thanks to an anonymous referee for constructive comments which improved the presentation. We also thank the participants at the first four Workshops on Higher-Order Asymptotics and Post-Selection Inference (WHOA-PSI) for their interesting ideas and captivating discussions. The second author was partially funded by NSF DMS 1712839. The third author was partially funded by NSF DMS 1712940.

LITERATURE CITED

Andrews, I., Kitagawa, T., and McCloskey, A. (2019). Inference on winners. Technical report, National Bureau of Economic Research.

Arcones, M. A. (2005). Convergence of the optimal M-estimator over a parametric family of M-estimators. *Test*, 14(1):281–315.

Austin, P. C., Mamdani, M. M., Juurlink, D. N., and Hux, J. E. (2006). Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *Journal of clinical epidemiology*, 59(9):964–969.

Bachoc, F., Leeb, H., and Pötscher, B. M. (2019). Valid confidence intervals for post-model-selection predictors. *Ann. Statist.*, 47(3):1475–1504.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2020). Uniformly valid confidence intervals post-model-selection. *Ann. Statist.*, 48(1):440–463.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018). High-dimensional econometrics and regularized GMM. *arXiv preprint arXiv:1806.01888*.

Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika*, 102(1):77–94.

Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619.

Benjamini, Y., Heller, R., and Yekutieli, D. (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4255–4271.

Beran, R. J. (1988). Balanced simultaneous confidence sets. *J. Am. Stat. Assoc.*, 83(403):679–686.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.*, 41(2):802–837.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754.

Cai, J. (2020+). *tmax: Valid post-selection inference under misspecified linear model.* R package version 1.0.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, 45(4):2309–2352.

Chernozhukov, V., Chetverikov, D., Kato, K., and Koike, Y. (2019). Improved central limit theorem and bootstrap approximations in high dimensions. *arXiv preprint arXiv:1912.10529*.

Chernozhukov, V., Chetverikov, D., and Koike, Y. (2020). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *arXiv preprint arXiv:2012.09513*.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7(1):649–688.

Cole, J. H. (2020). Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. *Neurobiology of Aging*, 92:34–42.

Dodge, Y. and Jurevckova, J. (2000). *Adaptive Regression*. Springer-Verlag, New York.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, (1):1–26.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Freedman, D. A. (1983). A note on screening regression equations. *Amer. Statist.*, 37(2):152–155.

Freedman, D. A. (2009). *Statistical models: theory and practice*. Cambridge university press.

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6):460–465.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

Giessing, A. and Fan, J. (2020). Bootstrapping ℓ_p-statistics in high dimensions. *arXiv preprint arXiv:2006.13099*.

Harrison, Jr., D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manage.*, 5(1).

Hong, L., Kuffner, T. A., and Martin, R. (2018). On overfitting and post-selection uncertainty assessments. *Biometrika*, 105(1):221–224.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *The Annals of Mathematical Statistics*, 11(3):271–283.

Kafadar, K. (2021). Editorial: Statistical significance, p-values, and replicability. *Annals of Applied Statistics*, 15(3):To appear.

Kivaranovic, D. and Leeb, H. (2018). Expected length of post-model-selection confidence intervals conditional on polyhedral constraints. *ArXiv e-prints:1803.01665*.

Kivaranovic, D. and Leeb, H. (2020). A (tight) upper bound for the length of confidence intervals with conditional coverage. *arXiv preprint arXiv:2007.12448*.

Kuchibhotla, A. K. (2018). Deterministic inequalities for smooth m-estimators. *arXiv preprint arXiv:1809.05172*.

Kuchibhotla, A. K. (2020). *Unified Framework for Post-Selection Inference*. PhD thesis, University of Pennsylvania.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2020). Valid post-selection inference in model-free linear regression. *Ann. Statist.*, 48:2953 – 2981.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2021). A Model Free Perspective for Linear Regression: Uniform-in-model Bounds for Post Selection Inference. *Econometric Theory*.

Kuchibhotla, A. K., Mukherjee, S., and Banerjee, D. (2021). High-dimensional CLT: Improvements, non-uniform extensions and large deviations. *Bernoulli*, 27(1):192–217.

Kuchibhotla, A. K. and Rinaldo, A. (2020). High-dimensional CLT for sums of non-degenerate random vectors: $n^{-1/2}$-rate. *arXiv preprint arXiv:2009.13673*.

Kuffner, T. A. and Young, G. A. (2018). Principled statistical inference in data science. In Adams, N., Cohen, E., and Guo, Y.-K., editors, *Statistical Data Science*, pages 21–36. World Scientific, Singapore.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution ofpost-model-selection estimators? *Ann. Statist.*, 34(5):2554–2591.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

Liquet, B. and Commenges, D. (2001). Correction of the p-value after multiple coding of an explanatory variable in logistic regression. *Statistics in Medicine*, 20(19):2815–2826.

Liquet, B. and Riou, J. (2013). Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 13(1):1–10.

Liquet, B. and Riou, J. (2019). CPMCGLM: an R package for p-value adjustment when looking for an optimal transformation of a single explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 19(1):1–8.

Lunde, R. (2019). Sample splitting and weak assumption inference for time series. *arXiv preprint arXiv:1902.07425*.

Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, 93(4):439–455.

Markovic, J., Xia, L., and Taylor, J. (2017). Unifying approach to selective inference with applications to cross-validation. *arXiv preprint arXiv:1703.06559*.

McCloskey, A. (2020). Hybrid confidence intervals for informative uniform asymptotic inference after model selection. *arXiv preprint arXiv:2011.12873*.

Moore, D. and McCabe, G. (1998). *Introduction to the Practice of Statistics*. W. H. Freeman, New York.

Norvaiša, R. and Paulauskas, V. (1991). Rate of convergence in the central limit theorem for empirical processes. *Journal of Theoretical Probability*, 4(3):511–534.

Panigrahi, S., Taylor, J., and Weinstein, A. (2016). Integrative methods for post-selection inference under convex constraints. *arXiv preprint arXiv:1605.08824*.

Pardoe, I. (2008). Modeling home prices using realtor data. *Journal of Statistics Education*, 16(2).

Paulauskas, V. and Račkauskas, A. (1989). *Approximation Theory in the Central Limit Theorem*. Kluwer Academic Publishers Group, Dordrecht.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.*, 22(4):2031–2050.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer-Verlag, New York.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 33(2):659–680.

Rasines, D. G. and Young, G. A. (2020). Bayesian selective inference: sampling models and non-informative priors. *arXiv preprint arXiv:2008.04584*.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.*, 47(6):3438–3469.

Sampson, A. R. and Sill, M. W. (2005). Drop-the-losers design: normal case. *Biometrical Journal*, 47(3):257–268.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2):87–110.

Sill, M. W. and Sampson, A. R. (2009). Drop-the-losers design: Binomial case. *Computational Statistics & Data Analysis*, 53(3):586–595.

Stine, R. and Foster, D. (2013). *Statistics for Business: Decision Making and Analysis*. Pearson Education, New York.

Tian, X., Bi, N., and Taylor, J. (2016). MAGIC: a general, powerful and tractable method for selective inference. *ArXiv e-prints*.

Tian, X. and Taylor, J. (2017). Asymptotics of selective inference. *Scand. J. Stat.*, 44(2):480–499.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *Ann. Statist.*, 46(2):679–710.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, 58(1):267–288.

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., Reid, S., and Markovic, J. (2019). *selectiveInference: Tools for Post-Selection Inference*. R package version 1.2.5.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Ann. Statist.*, 46(3):1255–1287.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

Tukey, J. W. (1953). The problem of multiple comparisons: Introduction and parts a, b, and c. Photocopy of typescript from Princeton University eventually published in the Tukey collected works in 1993 by Chapman and Hall.

Tullock, G. (2001). A comment on Daniel Klein's "a plea to economists who favor liberty". *Eastern Economic Journal*, 27(2):203–207.

Weisberg, S. (2005). *Applied Linear Regression*. John Wiley &; Sons, New York, third edition.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., and Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, 75(5):1182–1189.

Yekutieli, D. (2012). Adjusted Bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):515–541.

Zhang, K. (2012). *Valid Post-selection Inference*. PhD thesis, University of Pennsylvania, Publicly Accessible Penn Dissertations. 598. https://repository.upenn.edu/edissertations/598.

Zhang, X. and Cheng, G. (2014). Bootstrapping high dimensional time series. *arXiv preprint arXiv:1406.1037*.

Zhang, X. and Cheng, G. (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli*, 24(4A):2640–2675.

Post-Selection Inference Computational Supplement

Arun K. Kuchibhotla,[1] John E. Kolassa,[2] and Todd A. Kuffner[3]

[1]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, United States, 15232; email: arunku@cmu.edu
[2]Department of Statistics, Rutgers University, Piscataway, United States, 08854, email: kolassa@stat.rutgers.edu. Supported in part by NSF DMS-1712839.
[3]Department of Mathematics & Statistics, Washington University, St. Louis, United States, 63130, email: kuffner@wustl.edu. Supported in part by NSF DMS-1712940.

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–7

https://doi.org/10.1146/((please add article doi))

Keywords

Post-selection inference, Exploratory data analysis, Model selection, Data transformations, Selective inference, Sample splitting.

Abstract

This document supplements our review of VIDE techniques, by providing R code for performing calculations.

1. Introduction

Several approaches attempt to provide solutions to VIDE. They can be characterized by the following terms, to be explained below:

1. Sample Splitting,
2. Simultaneous Inference, and,
3. Conditional Selective Inference.

To illustrate the approaches, we use the Boston housing data available in `R` package `MASS`. This data set was introduced in Harrison and Rubinfeld (1978) to understand the impact of air pollution (measured as concentration of nitrogen oxide `NOX`) on the median value (`MEDV`) of houses in different census tracts in Boston. This effect is estimated when adjusting with other covariates including crime rate (`CRIM`), proportion of land zoned for lots (`ZN`), vicinity of Charles river (`CHAS`), number of rooms (`RM`), proportion of non-retail business acres per town (`INDUS`), proportion of owner-occupied units built prior to 1940 (`AGE`), weighted distances to five Boston employment centres (`DIS`), index of accessibility to radial highways (`RAD`), full-value property-tax rate per $10,000 (`TAX`), pupil-teacher ratio by town (`PTRATIO`), proportion of African-Americans (`Black`), and % lower status of the population (`LSTAT`).

2. Sample splitting

The package `caTools` Tuszynski (2021) includes a function for random sample splitting. In the interest of reproducibility, the random seed is fixed.

```
data(Boston, package="MASS")
library(caTools)
set.seed(27032021)
split_indicator <- sample.split(Boston$medv, SplitRatio = 1/2)
## Initial fit
lm0 <- lm(log(medv) ~ ., data = Boston)

## forward stepwise fixing 'nox' on split_indicator data
stepout <- step(lm(log(medv) ~ nox, data = Boston, subset = split_indicator),
            scope = formula(lm0), direction = "forward", trace = 0)

## p-values from stepout model using training data ignoring selection
print(summary(stepout)$coeff[-1, 4], digits=3)

##      nox    lstat  ptratio      dis     crim       rm     chas    black
## 2.69e-05 1.31e-28 2.42e-11 3.48e-07 1.41e-08 7.63e-05 4.20e-03 1.31e-02
##      rad      tax
## 1.96e-04 1.34e-03

## p-values from stepout model using test data
print(summary(lm(formula(stepout), data = Boston,
subset = !split_indicator))$coeff[-1, 4], digits = 3)

##      nox    lstat  ptratio      dis     crim       rm     chas    black
## 9.63e-03 2.77e-14 2.44e-05 1.28e-04 1.74e-06 4.65e-05 1.62e-01 3.91e-03
```

```
##      rad      tax
## 3.39e-04 3.52e-02
```

Similar splitting tools exist in other packages including `splithalfr` Pronk (2020).

3. Simultaneous inference approach to VIDE

As noted before, the methods of Tukey (1949, 1953) and Scheffé (1953) are appropriate for inference on contrasts; they are available in R, via the functions `TukeyHSD` and `scheffe.test` or `ScheffeTest` respectively. The latter functions require the packages by de Mendiburu (2020) and Andri et al. (2021), respectively. In the Boston housing data, the variable `rad` is a categorical variable taking 9 different values. One might simultaneously bound all mean valuation differences for houses with differing accessibility to radial highways. For simplicity, all values of `rad` above 5 are set to 5.

```
Boston_dat <- Boston
Boston_dat$rad[Boston_dat$rad >= 5] <- 5
table(Boston_dat$rad)

##
##   1   2   3   4   5
##  20  24  38 110 314

print(TukeyHSD(aov(log(medv) ~ as.factor(rad), data = Boston_dat)),
        digits = 3)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log(medv) ~ as.factor(rad), data = Boston_dat)
##
## $`as.factor(rad)`
##        diff    lwr     upr p adj
## 2-1  0.1020 -0.229  0.4328 0.917
## 3-1  0.1379 -0.164  0.4398 0.721
## 4-1 -0.1316 -0.397  0.1340 0.656
## 5-1 -0.1621 -0.414  0.0899 0.398
## 3-2  0.0360 -0.249  0.3209 0.997
## 4-2 -0.2336 -0.480  0.0126 0.072
## 5-2 -0.2641 -0.496 -0.0327 0.016
## 4-3 -0.2696 -0.475 -0.0640 0.003
## 5-3 -0.3001 -0.488 -0.1124 0.000
## 5-4 -0.0305 -0.152  0.0906 0.959
```

The output shows the difference in sample means for each pair of values of `rad`. These contrasts can also be tested using `scheffe.test`, but because it provides simultaneous inference over all contrasts (not just pairwise differences) it tends to be less powerful compared to `TukeyHSD` for pairwise differences.

```
library(agricolae)
print(scheffe.test(aov(log(medv) ~ rad, data = Boston_dat), "rad",
        group = FALSE)$comparison[,-3], digits=3)
```

```
##       Difference pvalue      LCL   UCL
## 1 - 2    -0.1020 0.9508 -0.47795 0.274
## 1 - 3    -0.1379 0.8184 -0.48099 0.205
## 1 - 4     0.1316 0.7691 -0.17022 0.434
## 1 - 5     0.1621 0.5478 -0.12425 0.449
## 2 - 3    -0.0360 0.9983 -0.35975 0.288
## 2 - 4     0.2336 0.1565 -0.04615 0.513
## 2 - 5     0.2641 0.0484  0.00111 0.527
## 3 - 4     0.2696 0.0135  0.03592 0.503
## 3 - 5     0.3001 0.0009  0.08679 0.513
## 4 - 5     0.0305 0.9764 -0.10710 0.168
```

Note that the p-values here are larger than those obtained from **TukeyHSD** function. The function **ScheffeTest** in the DescTools package offers more flexibility in specifying contrasts (i.e., more general contrasts than pairwise means).

Algorithm 1 for covariate selection under a well-specified linear model is implemented in the package PoSI (Buja and Zhang 2020).

```
library("PoSI")
summary(PoSI(Boston[,-14], verbose = 0))
```

```
##     K.PoSI K.Bonferroni K.Scheffe
## 95%  3.591        4.904     4.729
## 99%  4.075        5.211     5.262
```

The output shows the $\widehat{K}_\alpha$ values to be used. We only need to provide the covariate matrix to `PoSI`, because of the Gaussian linear model assumption. The PoSI constant shown above under the column `K.PoSI` is the smallest. Note that the Scheffe's constant is also shown under `K.Scheffe`. Without specifying other arguments, the output of `PoSI` provides adjustments for the universe of selection $\mathcal{Q} = \{(j, \mathrm{M}) : j \in \mathrm{M}, \mathrm{M} \subseteq \{1, 2, \ldots, p\}\}$. The argument `ModelSZ` in `PoSI` can be used to reduce the universe, which also reduces the computational complexity.

To go beyond the linear model assumptions and use Algorithm 1 , we can use the tmax package (in development) Cai (2020).

```
#install.packages("tmax_1.0.tar.gz", repos=NULL, dependencies=T)
library("tmax")

XX <- cbind(1, as.matrix(Boston[,-14]))
YY <- as.vector(Boston[,14])
tmp <- maxt_posi(XX, YY, sandwich = TRUE, alpha = 0.05, Nboot = 200)

## The PoSI constant under potential misspecification is
tmp$k
```

```
##       95%
## 4.624482

## If Mhat is the selected model, then adjusted confidence intervals are
Mhat <- c(1, 2, 5)
posi.PoSI_Berk(tmp, Mhat)$intervals.M

##            Lower     Upper
##       21.6190440 25.609018
## crim -0.7167384 -0.095231
## chas -3.8644235 15.019861
```

The function `maxt_posi` with `sandwich = TRUE` implements exactly Algorithm 1 with $\widehat{\Psi}_{n,q}$ equal to the sandwich variance estimator. Setting `sandwich = FALSE` will use the linear model based variance estimator. Case studies involving covariate selection and also transformation selection can be found in the package website of Cai (2020). Finally, max-t style correction in other VIDE problems including optimal cut-off detection and transformations are discussed in Liquet and Commenges (2001), Liquet and Riou (2013, 2019). These works also include illustrations of the R package CPMCGLM (Riou and Liquet 2017).

4. Conditional selective inference

The primary R package for conditional selective inference is selectiveInference (Tibshirani et al. 2019). The package selectiveInference provides valid post-selection inference after using common model-selection techniques including forward stepwise regression and the lasso. The package includes functions that perform model selection, sometimes using criteria different from similarly-named criteria implemented by other packages. These techniques include groupwise techniques such as group lasso. When the data follow a Gaussian distribution, then the resulting procedure provides tests with the correct type I error in finite samples, otherwise the guarantees are asymptotic. We begin with an application to stepwise regression.

```
library("selectiveInference")
## Center the covariates, but do not standardize
scaled_x <- scale(Boston[,-14], TRUE, FALSE)
fsout <- fs(scaled_x, log(Boston[,14]))
```

This procedure sequentially adds variables, with the next variable in each case chosen to maximize the increase in the regression sum of squares. This is equivalent to using AIC to select the next variable, but in this case stopping only after examining a certain number of larger models to avoid premature stopping, and gives the same model. The p-values and confidence intervals adjusted to stepwise selection can be obtained using

```
fsinf_out <- fsInf(fsout, type = "aic")
mat <- cbind(fsinf_out$pv, fsinf_out$ci)
dimnames(mat) <- list(colnames(Boston[,-14])[fsinf_out$vars],
                      c("p adj", "lwr", "upr"))
print(mat, digits = 3)
```

```
##           p adj      lwr       upr
## lstat   0.33550 -0.05246  0.053042
## ptratio 0.24041 -0.04493  0.042796
## crim    0.34545 -0.03413  0.025595
## rm      0.32349 -0.24014  0.380779
## dis     0.65175 -0.05189  0.301466
## nox     0.09967 -4.78435  0.460184
## black   0.62735     -Inf  0.007394
## rad     0.44520 -0.03969  0.017964
## tax     0.04415 -0.00113 -0.000023
## chas    0.00387  0.16362       Inf
## zn      0.25555 -0.00164  0.002741
## indus   0.29829 -0.00957  0.020518
## age     0.54697 -0.00418  0.002789
```

Note that the forward stepwise implementation in this package selected all covariates instead of 10 variables obtained via the `step` function. The output shows that confidence intervals for linear parameters are wider, to correctly allow for the effect of selection.

Apply the Lasso by first using package `glmnet` to apply cross-validation to minimize squared error. The package authors (Tibshirani et al. 2019) recommend applying the Lasso to centered and scaled covariates.

```
library("glmnet")
cv_out <- cv.glmnet(scaled_x, log(Boston[,14]), standardize = FALSE,
                    grouped = FALSE)
n <- dim(scaled_x)[1]
lambda <- 0.1
beta <- coef(cv_out$glmnet.fit, x = scaled_x,
             y = log(Boston[,14]), s = lambda, exact = TRUE)[-1]
sigmahat <- estimateSigma(scaled_x, log(Boston[,14]))$sigmahat
# Note that lambda is scaled differently in glmnet
# and selectiveInference packages.
# For the fixedLassoInf function, one CANNOT use cross-validated
# lambda value.
lasso_out <- fixedLassoInf(scaled_x, log(Boston[,14]),
                           beta, n*lambda, alpha = 0.05)

print(data.frame(vars = names(lasso_out$vars),
          p_adj = lasso_out$pv, lwr = lasso_out$ci[,1],
          upr = lasso_out$ci[,2]), digits = 3)

##      vars    p_adj       lwr       upr
## 1    crim 2.49e-11 -0.012091 -0.006957
## 2      zn 2.72e-01 -0.000901  0.001283
## 3     age 1.40e-03  0.000578  0.002370
## 4     rad 5.28e-05  0.009575  0.020953
## 5     tax 3.19e-05 -0.000995 -0.000443
```

```
## 6 ptratio 0.00e+00 -0.050948 -0.031245
## 7   black 2.84e-04  0.000181  0.000605
## 8   lstat 1.35e-60 -0.040679 -0.034145
```

LITERATURE CITED

Andri et al., S. (2021). *DescTools: Tools for Descriptive Statistics*. R package version 0.99.40.

Buja, A. and Zhang, K. (2020). *PoSI: Valid Post-Selection Inference for Linear LS Regression*. R package version 1.1.

Cai, J. (2020+). *tmax: Valid post-selection inference under misspecified linear model*. R package version 1.0.

de Mendiburu, F. (2020). *agricolae: Statistical Procedures for Agricultural Research*. R package version 1.3-3.

Harrison, Jr., D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manage.*, 5(1).

Liquet, B. and Commenges, D. (2001). Correction of the p-value after multiple coding of an explanatory variable in logistic regression. *Statistics in Medicine*, 20(19):2815–2826.

Liquet, B. and Riou, J. (2013). Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 13(1):1–10.

Liquet, B. and Riou, J. (2019). CPMCGLM: an R package for p-value adjustment when looking for an optimal transformation of a single explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 19(1):1–8.

Pronk, T. (2020). *splithalfr: Extensible Bootstrapped Split-Half Reliabilities*. R package version 2.0.12.

Riou, J. and Liquet, B. (2017). *CPMCGLM: Correction of the P-Value after Multiple Coding in Generalized Linear Models*. R package version 1.2.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2):87–110.

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., Reid, S., and Markovic, J. (2019). *selectiveInference: Tools for Post-Selection Inference*. R package version 1.2.5.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

Tukey, J. W. (1953). The problem of multiple comparisons: Introduction and parts a, b, and c. Photocopy of typescript from Princeton University eventually published in the Tukey collected works in 1993 by Chapman and Hall.

Tuszynski, J. (2021). *caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc.* R package version 1.18.1.