Sequential Specification Tests to Choose a Model: A Change-Point Approach

March 10, 2021

1 Introduction

Null hypothesis tests and p-values play a central role in model checking. In this context, the null hypothesis may be that that the data are drawn from a distribution contained in the the model under study, or it may be derived from an underlying assumption. Typically, researchers use these specification tests to check the fit of a model chosen by other means, but in some cases hypothesis tests form the basis of a model selection procedure. In these cases, researchers construct a sequence of model specifications, ordered by preferability, and test each one. The best model whose assumptions "pass" the hypothesis test is chosen.

For example, take the datasets displayed in Figure 1, which will be discussed in more detail in Section 5. Figure 1A shows the annual total unemployment rate in the United States from 1890 to 2015. One of the simpler models for time series such as these is an order p autoregression, or $AR(p)$ under which the value of the time series at point t may depend on its historical values at $t-1, ..., t-p$ but, conditional on those, is independent of values at points before $t-p$. SSTs can be useful here, too: researchers may test model fit for a sequence of lag orders p, and choose the smallest p that the tests fail to reject. Here a smaller lag orders p are preferable because they lead to more parsimonious models and more precise

estimates.

Figure 1B plots data that Lindo et al. [2010] used to estimate the effect of academic probation. University students were put on academic probation if their first-year cumulative grade point averages fell below a cutoff. This is an example of a regression discontinuity design [Thistlethwaite and Campbell, 1960], in which treatment is assigned if a numeric "running variable" R falls below (or above) a pre-specified cutoff c. Typically [e.g. Imbens and Lemieux, 2008, Lee and Lemieux, 2010], regression is used to model regression discontinuity designs, but Cattaneo et al. [2015], models Y in a small bandwidth around c as independent of R and other covariates, as if generated by a randomized experiment. Cattaneo et al. [2015] suggest choosing the bandwidth via sequential specification tests: for a sequence of nested windows around c, test whether, for subjects with R in the window, covariates are independent of treatment assignment. Then, choose the largest bandwidth for which the independence assumption cannot be reject at a pre-specified level.

These are both examples of the use of sequential specification tests (SSTs) to choose a model. SSTs are also used in covariate selection for regression models [Greene, 2003], selecting the number of components in mixture models, latent class analysis, and factor analysis [Nylund et al., 2007] and in propensity-score matching [Hansen and Sales, 2015].

Do hypothesis tests make any sense in model selection? The results of a null hypothesis test, of course, are never evidence in favor of a null hypothesis; null hypotheses can only be rejected, not accepted. Along similar lines, the logic of controlling type-I error rates seems backwards when it comes to model selection, in which accepting a problematic specification—a type II error—is the major concern. These issues have prompted some methodologists [e.g. Cattaneo et al., 2015] to propose adjusting the size of specification tests to a value higher than the conventional $\alpha = 0.05$. However, the appropriate value for α, and the criteria for selecting α, remain unclear.

On the other hand, a conceptually-sound model-selection method based on SSTs would be particularly useful; specification tests already exist for most common models, and they

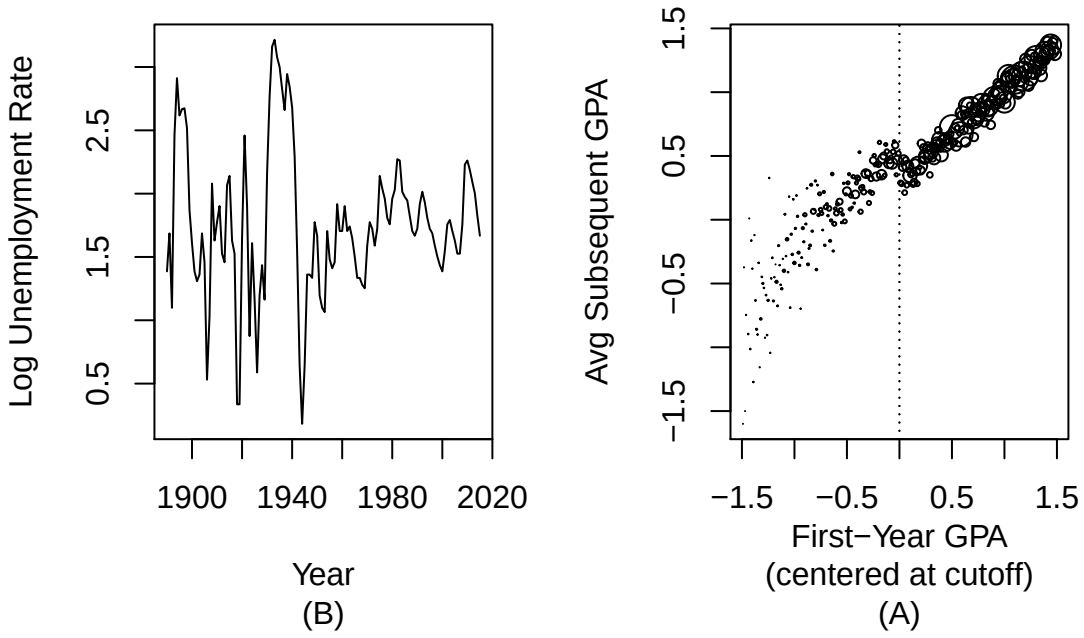

Figure 1: Plot (A) shows a time-series of log annual United States total unemployment from 1890 to 2015. Data were combined from Pfaff [2008] and Bureau of Labor Statistics [2016]. Plot (B) shows data from Lindo et al. [2010]: average subsequent grade point averages (GPAs), as a function of first-year GPAs, centered at the academic probation cutoff (dotted line).The points are sized proportionally to the number of students with each first-year GPA.

are regularly taught in introductory quantitative methods classes.

This paper develops such a method, based on a clever idea in change-point or threshold estimation. Mallik et al. [2011] points out that in a process with a change point, the p-values from a sequence of tests of a null regression function are uniformly-distributed as long as the regression function is correct, but asymptotically zero when the function is not correct. They use this dichotomous behavior to construct a simple, consistent estimator of the change-point, that is, the point at which the null model stops being correct.

In the same way, their estimator can choose the change-point in a sequence of models, when models stop being correct. Unlike under current approaches, an individual test result will itself not drive the change-point estimator, which is instead based on the entire sequence of p-values. What's more, unlike other SST model selectors, the change-point approach does

not require the researcher to specify a level α or any other tuning parameter. This approach shifts the model selection rationale away from the logic of hypothesis testing, based on type-I and type-II error rates, and towards the logic of estimation.

2 The Setup, in General

2.1 Sequences of Models and Tests

Say, in specifying a model, a researcher must choose from a discrete, ordered, set of specifications $d = 1, 2, \ldots, D$. The resulting model must satisfy testable assumption $\mathcal{A}$. Assume that either $\mathcal{A}$ is false for all d, or that for some $1 \leq d^* \leq D$, $\mathcal{A}$ is true for $d \leq d^*$ and false for all $d > d^*$. Further assume that if d^* exists, it is the optimal choice; for instance, it is the smallest model, or the biggest dataset, that satisfies $\mathcal{A}$. Finally, assume the researcher has chosen a valid, unbiased test of $\mathcal{A}$ and calculated p-values for each d: $\boldsymbol{p}_D = p_1, \ldots, p_d, \ldots, p_D$. The goal here is to use $\boldsymbol{p}_D$ to choose a specification $\hat{d}$ that is as large as possible without violating $\mathcal{A}$.

A common choice for d in this scenario relies on the logic of null hypothesis testing: for a pre-specified $\alpha \in (0, 1)$, let

$$d_\alpha^{max} \equiv max\{d : p_d > \alpha\}.$$

That is, d_α^{max} is the largest value of d for which the null hypothesis that $\mathcal{A}$ is true for $d \leq d_\alpha^{max}$ cannot be rejected at level α. Although it may seem as though the multiplicity of tests involved in this procedure invalidates the null hypothesis framework, it turns out that this is not the case: the "stepwise intersection-union principle" Berger et al. [1988], Rosenbaum [2008], Hansen and Sales [2015] insures that the family-wise error rate is maintained. That is, the probability of falsely rejecting the null and choosing $d_\alpha^{max} < d^*$, is bounded by α. d_α^{max} is the specification that would result from testing null hypotheses backwards: for $d' = D, D-1, \ldots, d, \ldots, 1$, test $H_{0d'} : \mathcal{A}$ is true for $d \leq d'$. Then, stop testing at $d' = d_\alpha^{max} - 1$,

the first d' for which $p_{d'} \geq \alpha$; reject all null hypotheses $H_{0d'}$ for which $d' \geq d_\alpha^{max}$, and fail to reject the rest. This protects the family-wise error rate of α because rejecting any true null implies rejecting the first true null, a probability α event.

Another common choice for $\hat{d}$ [e.g. Lütkepohl, 2005] does not have this property. Let

$$d_\alpha^{min} \equiv min\{d : p_d < \alpha\} - 1$$

d_α^{min} selects $\hat{d}$ to be the largest value of d before the first significant p-value. This is equivalent to the opposite procedure as d_α^{max}: start with the $d' = 1$ and test sequentially for larger values of d' until the first rejection, at d_α^{min}, then stop; reject all null hypotheses $H_{0d'}$ for $d' \geq d_\alpha^{min}$ and fail to reject the rest. This procedure does not control family-wise error rates, so it is likely to reject more than $100\alpha\%$ valid specifications.

2.2 Model Selection and the Logic of Null Hypothesis Testing

Typically, researchers use hypothesis tests to reject a null hypothesis that they consider uninteresting, such as a model parameter equal to zero, and interpret rejection as evidence in favor of an interesting alternative hypothesis. Null hypothesis tests cap the probability of a type-I error, and, given that constraint, seek to minimize the probability of a type-II error.

Sequential specification tests reverse some of these elements; most importantly, their is to identify specifications in which an assumption $\mathcal{A}$ is plausible, rather than to identify a true alternative hypothesis. In the same vein, type-II errors are typically of more concern for sequential specification tests than for typical null hypothesis tests, and type-I errors are somewhat less problematic. For that reason, some methodologists recommend setting α substantially higher for specification tests than for tests in outcome analyses. Still, the hypothesis testing framework, in the case of point null hypotheses, does not allow a researcher to fix the type-II error rate at a pre-specified value, and then optimize the type-I error rate, though that might be ideal for specification tests. In fact, in continuous data models with

continuous parameter spaces, no hypothesis test can provide any evidence in favor of a point null hypothesis. A common Bayesian argument (e.g. Kadane, 2011, p. 439; Gelman, 2004) states that, theoretically, nearly all null hypotheses are false anyway, so testing them makes little sense. In the case of specification tests, that means that an assumption $\mathcal{A}$ can be assumed to be false for all d without even conducting a test; in other words, "all models are wrong" [Box, 1979, p. 2]. That said, it may make sense to identify a set of specifications d for which $\mathcal{A}$ is plausible, or approximately true, and sequential specification tests can be useful in this regard.

In many scenarios the choice of d involves a bias-variance trade-off: if $d > d^*$, then $\mathcal{A}$ is false and the resulting analysis will be biased. On the other hand, a sub-optimal choice for d often means a high-variance estimate. For instance, in the regression discontinuity bandwidth case, choosing $d > d^*$ might mean fitting a misspecified model to Y and R, but choosing $d << d^*$ means discarding data that can boost precision. Rather than choosing a criterion, such as mean-squared-error, that balances bias and variance, the specification testing approach attempts to hold bias at approximately zero, and minimize variance under that constraint. However, doing so requires researchers to choose a test-level α. While using tuning parameters to mediate the bias-variance trade-off is not uncommon in statistics, the level α is a particularly hard parameter to choose.

More broadly, perhaps, one might argue that null hypothesis tests are design to rule out hypotheses that are inconsistent with the data, not to estimate parameters. However, as Hodges Jr and Lehmann [1963] showed, these aims are not contradictory: tests that rule out implausible hypothesis may also point researchers towards the correct answer. Moving from rejecting implausible specifications to estimating optimal specifications requires a theory, or at least a reasonable heuristic. The following section will suggest one.

3 Finding the Change-Point

3.1 The Change Point Estimator

In the context of change point estimation, Mallik et al. [2011] suggests such a heuristic. They discuss a random variable x_t, whose distribution is a function of a continuous covariate t. For $t < d^*$, $E(x_t) = \tau_0$, a constant; for $t > d^*$, $E(x_t) > \tau_0$. They propose an estimate of d_0 based on p-values p_t testing the hypotheses $H_{0t} : E(x_t) = \tau_0$. They note that for $t < d^*$, the null hypotheses are true, so $p_t \sim U(0, 1)$, and $E(p_t) = 1/2$; when $t > d^*$, the null hypotheses are false, and the p-values converge in probability to zero. That fact leads them to the following least-squares estimator for d^*:

$$\hat{d}_M \equiv arg\min_d \sum_{t \leq d} (p_t - 1/2)^2 + \sum_{t > d} p_t^2.$$

In other words, the estimate $\hat{d}_M$ is the point at which the p-values cease behaving as p-values testing a true null, with mean 1/2, and instead are drawn from a distribution with a lower mean. It turns out that an equivalent expression for $\hat{d}_M$ is:

$$\hat{d}_M = arg\max_d \sum_{t \leq d} (p_t - 1/4). \tag{1}$$

Mallik et al. [2011] shows that as n_t, the number of data points at each value t, and the number of sampled values of t increase, $\hat{d}_M$ converges in probability to d^*.

The same broad logic applies to p-values from sequential specification tests. Some differences in the details, though, lead to differences in $\hat{d}_M$'s behavior. For instance:

Proposition 1. *If indeed $p_d \to_p 0$ for $d > d^*$, as $n \to \infty$ then $\hat{d}_M$ is asymptotically conservative: $pr(\hat{d}_M > d^*) \to 0$.*

Proof. For each d, $pr(p_d - 1/4 > 0) \to 0$, implying that for all d', $pr(\sum_{d^* < t \leq d'} (p_t - 1/4) > 0) \to 0$. Therefore, for $d^* < d \leq D$, $pr(\sum_{t \leq d} (p_t - 1/4) > \sum_{t \leq d^*} (p_t - 1/4)) \to 0$. □

That is, as sample size increases, the probability that $\hat{d}_M$ suggests a model that violates assumption $\mathcal{A}$ decreases to zero. The same property holds for d_α^{max}, with $\alpha > 0$ fixed, for the same reason.

On the other hand, even with an infinite sample $\hat{d}_M$ may choose a sub-optimal model, $\hat{d}_M < d^*$. As sample size grows, the distribution of p_d, $d \leq d^*$ remains stable at $U(0,1)$. When $p_d^* - 1/4 < 0$, $\hat{d}_M \neq d^*$, because $\sum_{d \leq d^*-1}(p_d - 1/4) > \sum_{d \leq d^*}(p_d - 1/4)$. Because $pr(p_d^* - 1/4 < 0) = 1/4$ regardless of sample size, $\hat{d}_M$ will be conservative in large samples. The difference between sequential specification tests and the change-point case in Mallik et al. [2011] is that the latter case relies on a continuous covariate that may be sampled from any point on the unit interval, whereas in the SST case the choice set $d = 1, \ldots, D$ is discrete and held fixed in the asymptotics.

In a way, $\hat{d}_M$ is similar to $d_{0.25}^{max}$, the largest d for which $p_d > \alpha = 0.25$, because both penalize p-values lower than 0.25. However, they are not equivalent, as the following proposition shows:

Proposition 2. *$\hat{d}_M \leq d_{0.25}^{max}$, with $pr(\hat{d}_M < d_{0.25}^{max}) > 0$.*

Proof. By definition, $p_d < 0.25$ for all $d > d_{0.25}^{max}$. Therefore, $\sum_{t=d_{0.25}^{max}+1}^{d'}(p_t - 1/4) < 0$ for all $d' \geq d_{0.25}^{max} + 1$, which in turn implies that $\sum_{t \leq d_{0.25}^{max}}(p_t - 1/4) > \sum_{t \leq d'}(p_t - 1/4)$, proving that $\hat{d}_M \leq d_{0.25}^{max}$. On the other hand, if, say, $p_{d_{0.25}^{max}-1} + p_{d_{0.25}^{max}} < 1/2$, or, more generally, $\sum_{t=d'}^{d_{0.25}^{max}}(p_t - 1/4) < 0$, then $\hat{d}_M < d_{0.25}^{max}$. □

In general, the difference between d_α^{max} and $\hat{d}_M$ will be most pronounced when the distributions of p-values for $d > d^*$ are not monotonically decreasing in probability. In such a scenario, it is most probable that an errant p-value for $d >> d^*$ will be greater than α; one p-value determines d_α^{max}, but $\hat{d}_M$ relies on the entire set of p-values.

3.2 A More Flexible $\hat{d}_M$

In finite samples, p-values from tests of false null hypotheses will not always be zero. Similarly, many hypothesis tests are asymptotic and may not yield uniformly-distributed p-values in finite samples. Still, p-values from sequential specification tests may exhibit something similar to the dichotomous behavior that motivates $\hat{d}_M$, in which p-values for $d \leq d^*$ are distributed differently than p-values for $d > d^*$. For this reason, Mallik et al. [2011] suggested a more flexible estimate:

$$\hat{d}_M^{ab} \equiv arg \min_{d;0<b<a<1} \sum_{t \leq d} (p_t - a)^2 + \sum_{t > d} (p_t - b)^2$$

Like $\hat{d}_M$, model selector $\hat{d}_M^{ab}$ looks for behavior that differs between p-values testing true and false null hypotheses. Unlike $\hat{d}_M$, it does not depend on theoretically established distributions for these p-values, but searches over a grid for their location parameters. $\hat{d}_M^{ab}$ will be more computationally expensive to compute than $\hat{d}_M$, but will may yield better results, especially in small samples.

4 A Simulation Study

This section will present a simulation study to compare the behavior of model selectors d_α^{max}, d_α^{min}, and $\hat{d}_M$. The simulation imagines a sequence of 10 models, ordered from least to most preferable. The first 5 models are well specified; thereafter the models are increasingly misspecified. Each model is assessed with a Z-test. For models $d = 1, \ldots, 5$, the test statistic $Z_d \sim \mathcal{N}(0, 1)$, the standard normal distribution. For models $d = 6, \ldots, 10$, the test statistic is distributed as $Z \sim \mathcal{N}\big\{b(d-5), 1\big\}$, where the slope parameter b controls the power of specification tests for these misspecified models, which increases with d for values of $d > 5$. Specification p-values are generated by comparing all of these simulated test statistics against the null distribution $\mathcal{N}(0, 1)$.

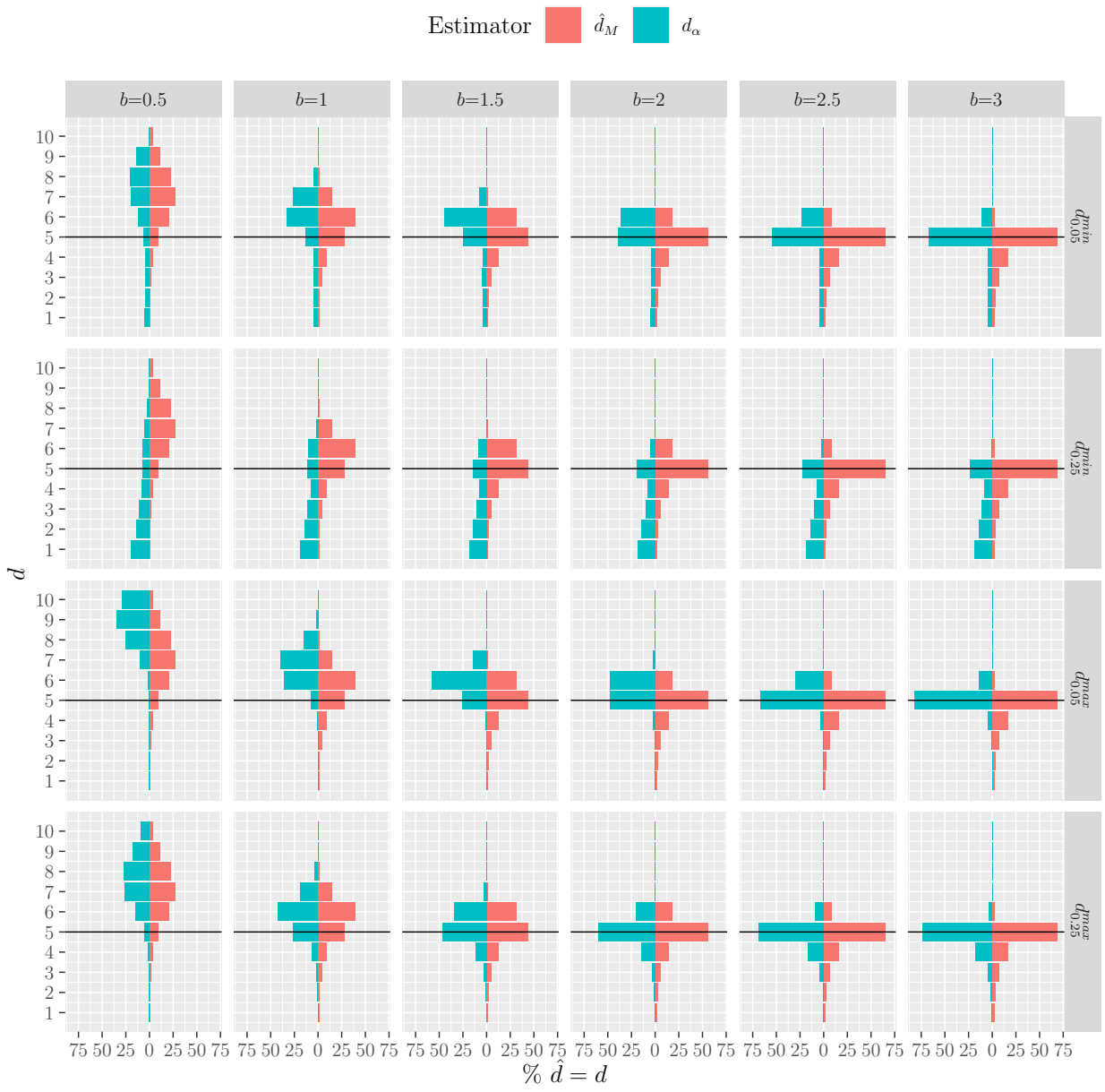

Figure 2: Results from 10^4 simulation runs comparing $\hat{d}_M$ to d^{max} and d^{min} with $\alpha = 0.05$ and 0.25. Each row compares either d^{max} or d^{min} to the same set of $\hat{d}_M$ estimates. Each bar represents the percent of runs in which an estimator selects each possible model, indexed as $d = 1, \ldots, 10$. Model d =5 (indicated with a horizontal line) is the optimal model, with models d >5 misspecified, and models d <5 well-specified but suboptimal.

	b=0.5			b=2			b=3		
Estimator	RMSE	%Opt.	%$> d^*$	RMSE	%Opt.	%$> d^*$	RMSE	%Opt.	%$> d^*$
$d_{0.05}^{max}$	15.3	0	100	0.6	47	50	0.2	82	14
$d_{0.25}^{max}$	8.5	5	93	0.6	60	20	0.6	73	3
$d_{0.05}^{min}$	6.3	6	65	1.8	39	37	1.4	67	11
$d_{0.25}^{min}$	5.4	7	16	4.7	19	5	4.7	23	1
$\hat{d}_M$	6.1	9	84	1.1	56	18	1.0	69	3

Table 1: Some results from 10^4 simulation runs comparing $\hat{d}_M$ to d^{max} and d^{min} with $\alpha = 0.05$ and 0.25. For $b = 0.5, 2, 3$, the root-mean-squared error (RMSE) of each estimator $\overline{(\hat{d} - d^*)^2}^{0.5}$ and the percentages each estimator chose the optimal model (%Opt.) or chose a misspecified model (%$> d^*$)

Figure 2 and Table 1 give the results of the simulation study, comparing $\hat{d}_M$ to $d_{0.05}^{min}$, $d_{0.25}^{min}$, $d_{0.05}^{max}$ and $d_{0.25}^{max}$, respectively. Table 1 compares all five model selectors at $b = 0.5$, 2, and 3 on three criterion: root mean-squared-error ($RMSE(x) = \overline{(x - d^*)^2}^{0.5}$), a measure of how close, in general, the estimator is to the optimal value, the percentage of runs in which it chose the optimal value d^* (%Opt.) and the percentage of runs in which it chose a misspecified model, i.e. chose $d > d^*$ (%$> d^*$).

In Figure 2, each bar represents the percentage of the times each model selector chose model d, with $d = 1, \ldots, 10$. Model d =5 (indicated with a horizontal line) is the optimal model, with models d >5 misspecified, and models d <5 well-specified but suboptimal. Each column of Figure 2 corresponds to a different value for the slope parameter $b \in \{0.5, 1, 1.5, 2, 2.5, 3\}$. As b increases, so does the power of the specification test, allowing the test to reject misspecified models at smaller values of $d > d^*$. Each row compares the same set of $\hat{d}_M$ to either $d_{0.05}^{min}$, $d_{0.25}^{min}$, $d_{0.05}^{max}$, or $d_{0.25}^{max}$.

When $b = 0.5$, the power to detect misspecification for models $d > d^*$ is relatively low. $\hat{d}_M$, $d_{0.05}^{min}$, and both d^{max} model selectors tend to choose models that are too big. That said, of those four estimators, $\hat{d}_M$ has the smallest root mean-squared-error (RMSE) and $\hat{d}_M$ is most likely of all model selectors to choose the optimal model d^*. $d_{0.25}^{min}$, which is the least likely to recommend a misspecified model, tends to recommend $d = 1$, the smallest possible,

least-optimal model.

As b increases, the performance of all five model selectors improves. Throughout, $\hat{d}_M$ is competitive in all three criteria and, arguably, balances them the best. At $b = 2$, and $b = 3$, the d^{max} estimators have better RMSE and tend to pick the optimal model slightly more often than $\hat{d}_M$, but are more likely to pick misspecified models. $d^{min}_{0.25}$ is the least likely to pick misspecified models, but the models it does pick tend to be much too small.

In general, $\hat{d}_M$ tends to be more conservative than d^{max} or d^{min} with $\alpha = 0.05$ but much less conservative than $d^{min}_{0.25}$. Its performance is most similar to $d^{max}_{0.25}$, while being slightly more conservative.

5 Two Data Examples

5.1 Lag Order in Autoregression Models: US Total Unemployment

Figure 1B shows the natural logarithm of the United States total unemployment rate from 1890 to 2016. The data were combined from the "Nelson & Plosser extended data set" provided in the `urca` library in `R` [Pfaff, 2008, R Core Team, 2016], which covers years 1890–1988, and a downloadable dataset from the United States Bureau of Labor Statistics, itself derived from the Current Population Survey, which covers years 1947–2015 [Bureau of Labor Statistics, 2016]. The two datasets agree on the overlapping years.

Assume that the time series follows an "AR(d)" model; that is,

$$unemp_t = \mu + \sum_{i=1}^{d} \phi_i unemp_{t-i} + \epsilon_t \tag{2}$$

where μ and $\{\phi_i\}_{i=1}^{d}$ are parameters to be estimated and ϵ_t is white noise. In this model, the unemployment in one year is a function of unemployment rates in the previous d years, but conditionally independent of even earlier measurements.

Having settled on model (2), the analyst must choose d, the lag order. sequential specification tests can be useful here [e.g. Ivanov et al., 2005]. Consider the null hypothesis $H_d : \phi_i = 0$ for all $i > d$; a researcher could test a sequence of such null hypotheses, for a set of plausible values of d, and choose the d based on the results. Other options for choosing d include optimizing information criteria [Akaike, 1969, Schwarz et al., 1978]. For instance, choosing the model that minimizes AIC, defined as $2d - 2log(\hat{L}_d)$, where $\hat{L}_d$ is the maximized likelihood of the $AR(d)$ model, or BIC, which is defined as $log(n)d - 2 - 2log(\hat{L}_d)$. Pötscher 1991 points out that differences in AIC or BIC are essentially likelihood ratio test statistics. Sequential specification tests can assist a modeler to choose the smallest model that is still approximately correct, as opposed to the model that maximizes predictive accuracy as measured by, say, mean squared error. A large literature surrounds this important question [See, e.g. McQuarrie and Tsai, 1998, Liew, 2004, and the citations therein]. This section is not meant as a complete treatment, or even an overview, of lag order selection, but as an illustration of sequential specification tests in a well-known area.

Figure 3 gives the p-values from a sequence likelihood ratio tests, as described in Pfaff [2008, Ch.1], which discussed a similar dataset. For each candidate lag order d, the likelihood ratio test compares twice the ratio of the log likelihoods of $AR(d+1)$ and $AR(d)$ models to a χ^2_1 distribution. If the $AR(d+1)$ model fits much better than the $AR(d)$ model, a lag order of d may not be sufficient. The p-values follow a stark pattern: for $d < 5$, they are close to zero, while for $d \geq 5$, they appear roughly uniformly distributed.

Table 2, and vertical lines in Figure 3, show the lag order choices from d^{max}_{α}, d^{min}_{α}, $\hat{d}_M$, and $\hat{d}^{ab}_M$, which are based on the p-values, and the lag orders that minimize AIC and BIC, based directly on the models' likelihood and numbers of parameters. Here, smaller models are preferable to larger models, so d^* is the smallest acceptable value for d. This is the opposite of the regression discontinuity case, which attempted to find the largest dataset on which to fit the model.

The change-point selectors $\hat{d}_M$ and $\hat{d}^{ab}_M$ both selected a lag order of 5, consistent with

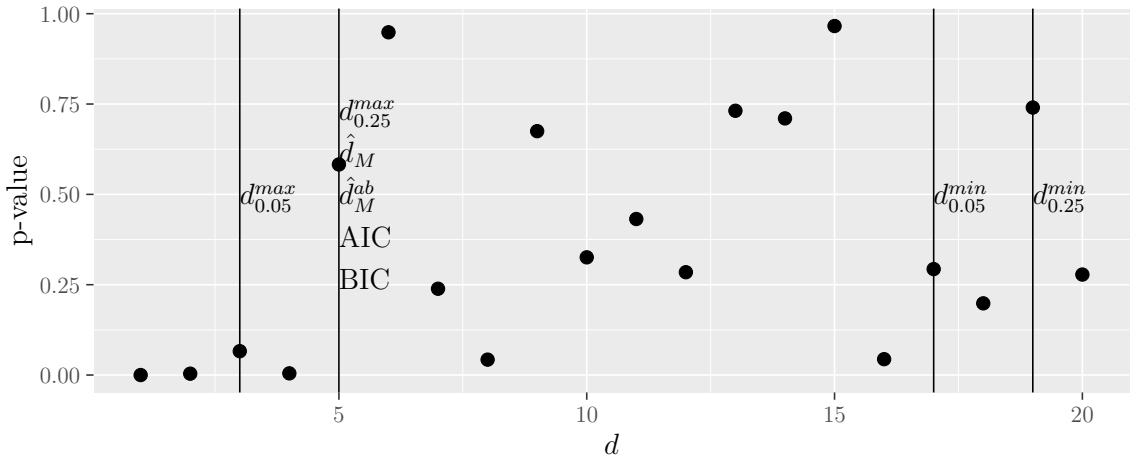

Figure 3: P-values from likelihood-ratio tests of model fit, comparing models AR(d) with AR($d+1$) in the annual total US unemployment rate (logged) time series.

the casual observation that p-values for lags less than this value are very small, while those greater appear approximately uniform. Incidentally, the two information criteria considered, AIC and BIC, agreed with this choice, as did $d^{max}_{0.15}$. In contradistinction, $d^{max}_{0.05}$ chose a smaller lag order of 3, because the corresponding p-value of 0.066 slightly exceeds the threshold of 0.05.

At the other extreme, the d^{min}_{α} selectors both chose very large models with d =17, due to the presence of of a small p-value of 0.044 at d =16.

This example illustrates how considering the entire distribution of p-values, as $\hat{d}_M$ does, can lead to better model selection than considering only the small (as in d^{min}_{α}) or large (d^{max}_{α}) values.

	$d^{max}_{0.05}$	$d^{max}_{0.25}$	$d^{min}_{0.05}$	$d^{min}_{0.25}$	$\hat{d}_M$	$\hat{d}^{ab}_M$	AIC	BIC
Lag Order	3	5	17	19	5	5	5	5

Table 2: Lag order selections for an $AR(d)$ model of the US unemployment time series.

5.2 Sequential specification tests in Regression Discontinuity Bandwidth Selection: Estimating the Effect of Academic Probation on College GPAs

At many universities, students who fail to achieve a minimum GPA c are put on academic probation. Lindo et al. [2010] recognized that academic probation can form a regression discontinuity design, in which treatment is a function of a "running variable" with a predetermined cutoff. Specifically, probation Z is a function of a "running variable" R, students' GPAs: students with $R < c$ are put on probation—$Z = 1$—and students with $R > c$ are not, $Z = 0$. That being the case, students with GPAs just below c may be comparable to students with GPAs just above c, so comparing these two sets of students allows researchers to estimate the effect of probation on outcomes Y (perhaps after adjusting for Y's relationship with R). The challenge becomes defining "just above" and "just below"—that is, selecting a "bandwidth" $bw^* > 0$ such that subjects i with $R_i \in (c - bw^*, c)$ are suitably comparable to subjects with $R_i \in (c, c + bw^*)$.

A number of authors [Lee and Lemieux, 2010, Cattaneo et al., 2015, Li et al., 2015, e.g.] recommend sequential specification tests, using baseline covariates X, as part of the procedure for choosing bw. At a sequence of candidate bandwidths $0 < bw_1 < \cdots < bw_d < \cdots < bw_D$, they recommend testing the equality of covariate means (again, perhaps after adjusting for R) between subjects with $R_i \in (c - bw_d, c)$ and those with $R_i \in (c, c + bw_d)$, and choosing a bandwidth $bw^* = bw_{d^*}$. These are essentially placebo tests—since the treatment cannot affect baseline covariates, differences in covariate means between treated and untreated subjects must be an indicator of incomparability between the groups, or model misspecification.

In a secondary analysis of the academic probation dataset, Sales and Hansen [2020] chose an RDD bandwidth using a set of seven baseline covariates: students' high-school GPA (expressed in percentiles), age at college matriculation, number of attempted credits, gender,

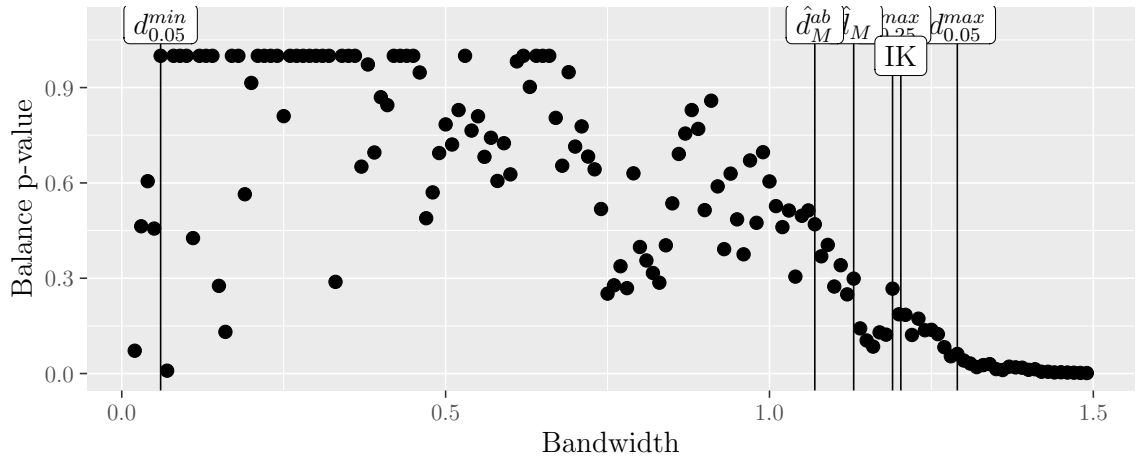

Figure 4: P-values from for balance in all seven covariates from the Lindo et al. [2010] analysis, following the method in Sales and Hansen [2020]. Vertical lines denote bandwidth choices using different criteria.

native language (English or other), birth place (North America or other) and university campus (the university consisted of three campuses). For each covariate X_k and for each candidate bandwidth bw_d, they let p_{kd} be the p-value corresponding the coefficient on Z from the regression of X_k on R and Z, fit to the subset of students with $R \in (c - bw_d, c + bw_d)$. These regression models were linear for continuous covariates and logistic for binary covariates, with heteroskedasticity-consistent sandwich standard errors [Zeileis et al., 2020, Zeileis, 2004, 2006]. Then, the ommibus specification p-value for bandwidth bw_d was $p_d = min\{1, 7p_{1d}, \ldots, 7p_{7d}\}$, the minimum of the Bonferroni-adjusted p-values p_{kd}.

The resulting p-values are plotted in Figure 4, with bandwidth selections corresponding to $d_{0.05}^{max}$, $d_{0.25}^{max}$, $d_{0.05}^{min}$, $\hat{d}_M$, $\hat{d}_M^{ab}$, and IK. Also plotted is the more conventional bandwidth recommended by Imbens and Kalyanaraman [2011], denoted IK, which is based on non-parametric estimates of the curvature of the regression function of Y on R, rather than covariate placebo tests. These bandwidth selections are also listed in Table 3.

For most small bandwidths bw_d, p_d is fairly large, and in many cases equal to 1. This apparent super-uniform distribution is probably due to the conservative Bonferroni correction applied to the p-values from individual covariates. On the other hand, at the smallest

candidate bandwidth bw_1 =0.02, the p-value is p_1 =0.072, and the p-value at the 6th bandwidth, $bw_6 = 0.07$, another small bandwidth, is $p_6 = 0.009$. After around $bw = 0.75$, the p-values begin decreasing, until by $bw = 1.5$, the p-values are all close to zero.

	$\hat{d}$	Bandwidth	Effect (95% CI)
$d_{0.05}^{max}$	128.00	1.29	0.22 (0.18,0.26)
$d_{0.25}^{max}$	118.00	1.19	0.23 (0.18,0.27)
$d_{0.05}^{min}$	5.00	0.06	0.1 (-0.17,0.37)
$d_{0.25}^{min}$		N/A	N/A
$\hat{d}_M$	112.00	1.13	0.23 (0.19,0.27)
$\hat{d}_M^{ab}$	106.00	1.07	0.22 (0.18,0.27)
IK	119.00	1.2	0.23 (0.19,0.28)

Table 3: Selected regression discontinuity bandwidths ('d' is the point in the sequence selected, and 'bw' is the bandwidth) using covariate balance tests for all the covariates in Lindo et al. [2010] and only high school GPA, respectively, along with their assosicated estimates for the average treatment effect of academic probation (ATE), with standard errors.

Since the p-value at the smallest candidate bandwidth, $p_1 = 0.072 < 0.25$, the model selector $d_{0.25}^{min}$ does not select anything—there is no d' small enough so that $p_d < 0.25$ for all $d \leq d'$. Similarly, the very low p-value at the 6th bandwidth causes $d_{0.05}^{min}$ to select a relatively small bandwidth of 0.07. This illustrates the sensitivity of d^{min} to outlier p-values at small d.

The remaining selectors all recommend bandwidths greater than 1, ranging from $\hat{d}_M^{ab}$, which recommends bandwidth $bw_{\hat{d}_M^{ab}}$=1.07 to $d_{0.05}^{max}$, which recommends bandwidth $bw_{d_{0.05}^{max}}$ =1.29. As in the simulation and the unemployment example, $\hat{d}_M$ and $d_{0.25}^{max}$ are quite close to each other. The similarity of the IK bandwidth of 1.2 to the bandwidths selected by d^{max} and $\hat{d}_M$ suggests an encouraging agreement, in this example, between covariate-based bandwidth selection and the more conventional RDD approach.

It is worth noting that super-uniformity of the p-values for small bandwidths inflates the sums $\sum_{t \leq d}(p_t - 1/4)$ in (1). Therefore, in this case $\hat{d}_M^{ab}$, which relies less on the uniform model for p-values under H_0, may be a more appropriate choice than $\hat{d}_M$.

Table 3 also lists estimated treatment effects of academic probation on students' sub-

sequent GPAs, along with 95% confidence intervals. The effects were estimated following the method described in Sales and Hansen [2020], with the exception of the estimate for the IK bandwidth which used local linear regression, as implemented in the `R` package `rdd` [Dimmery, 2016]. With the exception of the effect corresponding to $d^{min}_{0.05}$, all estimated effects are roughly equal, about grade points. Actually, the confidence interval corresponding to the $d^{min}_{0.05}$ bandwidth, (-0.17,0.37) is wide enough to contain both a negative academic probation effect along with all of the other estimated effects and confidence intervals. The conservativism of $d^{min}_{0.05}$ prevents efficient effect estimation; the conservativism of $d^{min}_{0.25}$ prevents estimation altogether.

6 Discussion

As long as data analysts use specification tests and p-values to select their models, decision rules translating a sequence of p-values to a model choice will be necessary. Currently, the most common approach compares the p-values to a pre-specified threshhold. This approach turns the logic of null hypothesis testing on its head, using p-values to identify well-specified models—i.e. true null hypotheses–rather than to reject misspecified models. Moreover, the d^{min}_{α} approach, by failing to control the familywise type-I error rate, can be extremely conservative, for instance recommending very high lag order in the unemployment example of Section 5.1 and very small bandwidths (or none at all) in the academic probation example of Section 5.2. In contrast, d^{max}_{α} does control familywise type-I error rates. However, both threshhold-based approaches, d^{min} and d^{max}, require specifying a threshhold, and there is rarely any clear guidence on how to do so.

The alternatives introduced here, $\hat{d}_M$ and $\hat{d}^{ab}_M$, drawn from the change-point literature, skirt these issues entirely. Rather than using p-values to reject (or accept) null hypotheses, they examine the full distribution of p-values. They require no arbitrary threshhold to be specified. As shown in the simulation study and the two data examples, they tend to avoid

the conservativism and outlier sensitivity of $d^{min}_{0.25}$ and the anti-conservativism of $d^{max}_{0.05}$ (indeed, Proposition 1 states that $\hat{d}_M$ is asymptotically conservative).

Actually, the simulation results and examples show that $\hat{d}_M$ tends to agree with $d^{max}_{0.25}$, which itself performs rather well. This suggests that d^{max} with a default threshhold of 0.25 may be a good option for data analysts who wish to continue using threshhold-based approaches.

There are several open questions regarding $\hat{d}_M$'s behavior and use. First, it is unclear whether or when the more flexible version $\hat{d}^{ab}_M$ should be preferred to $\hat{d}_M$; there is good reason to expect it to perform better when sample sizes are small, but is there a cost associated with using $\hat{d}^{ab}_M$ in larger samples? Further, there may be ways to construct sequential specification tests in a way that improves $\hat{d}_M$'s performance. How to best construct specification tests for different model selectors is a topic for future research.

Ultimately, the goal of model selection is to produce parameter estimates or predictions with desired properties. Ideally, researchers would select a model with this end in mind; however, the effect of model selection on final estimates or predictions depends heavily on specific circumstances. That said, a careful study of the effect of $\hat{d}_M$ on estimates and predictions in a wide range of cases could be useful.

Researchers who want to use hypothesis tests to choose from a sequence of models may feel uneasy about the statistical validity of their procedure or their choice of α. This paper will hopefully show how to choose a model using a sequence of p-values in a way that is coherent and does not require arbitrary cutoffs or tuning parameters.

References

Hirotugu Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.

Roger L Berger, Dennis D Boos, and Frank M Guess. Tests and confidence sets for comparing

two mean residual life functions. *Biometrics*, pages 103–115, 1988.

George EP Box. Robustness in the strategy of scientific model building. *Robustness in statistics*, 1:201–236, 1979.

United State Department of Labor Bureau of Labor Statistics. Employment status of the civilian noninstitutional population, 1946 to date. `https://www.bls.gov/cps/cpsaat01.xlsx` (accessed 1/25/2017), 2016.

Matias D Cattaneo, Brigham R Frandsen, and Rocio Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference*, 3(1):1–24, 2015.

Drew Dimmery. *rdd: Regression Discontinuity Estimation*, 2016. URL `https://CRAN.R-project.org/package=rdd`. R package version 0.57.

Andrew Gelman. Type 1, type 2, type s, and type m errors. `http://andrewgelman.com/2004/12/29/type_1_type_2_t/`, (accessed 2/9/17), 2004.

William H Greene. *Econometric analysis*. Pearson Education India, 2003.

Ben B Hansen and Adam Sales. Comment on cochran's "observational studies". *Introduction to Observational Studies and the Reprint of Cochran's paper "Observational Studies" and Comments*, page 184, 2015.

Joseph L Hodges Jr and Erich L Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, pages 598–611, 1963.

Guido Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, page rdr043, 2011.

Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635, 2008.

Ventzislav Ivanov, Lutz Kilian, et al. A practitioner's guide to lag order selection for var impulse response analysis. *Studies in Nonlinear Dynamics and Econometrics*, 9(1):1–34, 2005.

Joseph B Kadane. *Principles of uncertainty*. CRC Press, 2011.

David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

Fan Li, Alessandra Mattei, Fabrizia Mealli, et al. Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 9(4):1906–1931, 2015.

Venus Liew. Which lag length selection criteria should we employ? *Economics Bulletin*, 3 (33):1–9, 2004.

Jason M Lindo, Nicholas J Sanders, and Philip Oreopoulos. Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics*, 2(2):95–117, 2010.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

A Mallik, B Sen, M Banerjee, and G Michailidis. Threshold estimation based on a p-value framework in dose-response and regression settings. *Biometrika*, 98(4):887, 2011.

Allan DR McQuarrie and Chih-Ling Tsai. *Regression and time series model selection*. World Scientific, 1998.

Karen L Nylund, Tihomir Asparouhov, and Bengt O Muthén. Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural equation modeling: A multidisciplinary Journal*, 14(4):535–569, 2007.

B. Pfaff. *Analysis of Integrated and Cointegrated Time Series with R.* Springer, New York, second edition, 2008. URL `http://www.pfaffikus.de`. ISBN 0-387-27960-1.

Benedikt M Pötscher. Effects of model selection on inference. *Econometric Theory*, pages 163–185, 1991.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

Paul R Rosenbaum. Testing hypotheses in order. *Biometrika*, 95(1):248–252, 2008.

Adam C Sales and Ben B Hansen. Limitless regression discontinuity. *Journal of Educational and Behavioral Statistics*, 45(2):143–174, 2020.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.

Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6): 309–317, 1960.

Achim Zeileis. Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004. doi: 10.18637/jss.v011.i10.

Achim Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16, 2006. doi: 10.18637/jss.v016.i09.

Achim Zeileis, Susanne Köll, and Nathaniel Graham. Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, 95 (1):1–36, 2020. doi: 10.18637/jss.v095.i01.