Submitted to Statistical Science

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

Adam C Sales*

The University of Texas at Austin

Abstract. Researchers faced with a sequence of candidate model specifications must often choose the best specification that does not violate a testable identification assumption. One option in this scenario is sequential specification tests (SSTs): hypothesis tests of the identification assumption over the sequence. Borrowing an idea from the change-point literature, this paper shows how to use the distribution of p-values from SSTs to estimate the point in the sequence where the identification assumption ceases to hold. Unlike current approaches, this method is robust to individual errant p-values and does not require choosing a test level or tuning parameter. This paper demonstrates the method's properties with a simulation study, and illustrates it by application to the problems of choosing a bandwidth in a regression discontinuity design while maintaining covariate balance and of choosing a lag order for a time series model.

1. INTRODUCTION

One of many mis-matches between best practices recommended by statisticians and practice in quantitative research regards model selection. Statisticians conceptualize model selection as a tradeoff between bias and variance. Many quantitative researchers adopt a different outlook, thinking about model selec-

*Thanks George Michailidis, Ben B Hansen and Brian Junker for helpful comments. This work was partially supported by IES Grant R305B1000012

tion as choosing the best model that satisfies the assumptions of their intended statistical test or estimator—essentially minimizing variance while constraining bias at zero. This latter outlook leads researchers towards hypothesis tests of model assumptions; in particular, "sequential specification tests" (SSTs): a sequence of hypothesis tests, for a sequence of models, ordered by preferability. The best model whose assumptions "pass" a hypothesis test is chosen.

Do hypothesis tests make any sense in model selection? For one, "all models are wrong" [Box, 1979, p. 2] and "there's no such thing as unbiased estimation" [Gelman, 2004b], so the search for a correct model might be hopeless, and therefore pointless. Further, the logic of null-hypothesis testing seems incompatible with this framework. 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.

On the other hand, "some models are useful," and depending on their intended use, their usefulness may depend on *approximate* correctness. If so, hypothesis tests may have a role to play. Specification tests already exist for most common models, and they are regularly taught in introductory quantitative methods classes. If their use in model selection could be made conceptually sound, they are likely to be actually used—and maybe even correctly.

This paper will apply a clever idea from change-point or threshold estimation to the more general problem of model selection from SSTs. 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—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. In doing so, it shifts the model selection rationale away from the logic of hypothesis testing and towards the logic of

estimation. In the tradition of constructing confidence intervals from hypothesis tests and Hodges Jr and Lehmann [1963], their estimator exploits the behavior of hypothesis tests to estimate quantities of interest. Further, as opposed to model selectors based on strict hypothesis-testing logic, an individual test result will itself not drive the change-point estimator, which is instead based on the entire sequence of p-values. Thus, the change-point view of model selection is arguably conceptually more satisfying and practically more reliable than the conventional test-based approach. 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.

The following sub-section will briefly introduce two running examples of sequential specification tests: choosing a bandwidth for a regression discontinuity design and choosing a lag order for a time-series model. Next, Section 2 will review the formalism of SSTs and discuss common SST-based model selectors. Section 3 will introduce the new method, Section 4 will demonstrate some of its properties in a simulation study, 5 will apply it to the running examples, and 6 will conclude.

1.1 SSTs in Regression Discontinuity and Time Series

Figure 1 displays two datasets that will serve as illustrations of SSTs. Section 5 will discuss both of these examples in more detail. The brief overview here will be helpful to fix ideas.

Figure 1A plots data that Lindo et al. [2010] used to estimate the effect of academic probation. Students at an unnamed large Canadian University were put on academic probation—simultaneously given extra help and threatened with suspension—if their first-year cumulative grade point averages (GPAs) fell below a cutoff. This is an example of a regression discontinuity design (RDD) [Thistlethwaite and Campbell, 1960], in which treatment (in this case academic probation) is assigned if a numeric "running variable" R (first-year GPA) falls below (or above) a pre-specified cutoff c. Typically [e.g. Imbens and Lemieux, 2008, Angrist and Lavy, 1999, Lee and Lemieuxa, 2010] analysts will fit regression models

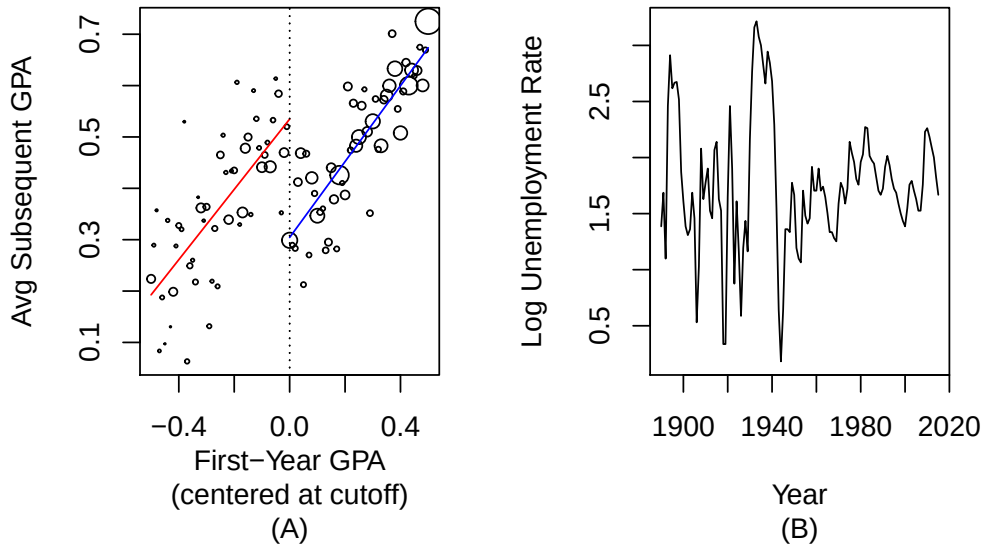

FIG 1. *Two data examples for SSTs. Plot (A) shows data from Lindo et al. [2010]—subsequent grade point averages (GPAs) for students at a large Canadian university, as a function of first-year GPAs. Subsequent GPAs are averaged by first-year GPAs, which are centered at the academic probation cutoff (dotted line), and the sizes of the plotted points are proportional to the number of students with each first-year GPA. The red and blue lines are linear least-squares fits on either side of the cutoff. Students with first-year GPAs to the left of the cutoff are put on probation. Plot (B) 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].*

$Y = f_1(R) + \epsilon$ and $Y = f_2(R) + \epsilon$ to data on either side of the cutoff, modeling the relationship between R and an outcome of interest Y. The difference between the models' predictions when R is set equal to the c is interpreted as a "local average treatment effect," roughly speaking the treatment effect when the running variable is equal to the cutoff [Hahn et al., 2001]. Figure 1A shows one of the outcomes Lindo et al. [2010] considered, students' subsequent GPAs, along with linear regression models below (in red) and above (blue) c, which is signified with a dotted line. A simpler alternative approach, suggested in Cattaneo et al. [2015], models the relationships between Y and R on either side of c as constant, and treats the data as if they were generated by a randomized experiment.

Of course, misspecified regression models will lead to biased treatment effect estimators. To minimize the influence of model misspecification, researchers will typically fit the regression models using only subjects for whom $R \in \mathcal{W}_b \equiv (c - b, c+b)$ for some bandwidth $b > 0$. A number of options exist for choosing the RDD bandwidth, including cross-validation [Ludwig and Miller, 2005] and asymptotic minimization of mean-squared-error [Imbens and Kalyanaraman, 2011]. Cattaneo et al. [2015] and others Sales and Hansen [2014], Angrist and Rokkanen [2015] suggest SSTs of covariate balance—at a sequence of candidate bandwidths b, test for the presence of a "treatment effect" on a pre-treatment covariate X, referred to as covariate imbalance. A window choice $\mathcal{W}_b$ that, paradoxically, leads to a statistically significant treatment effect on a covariate is unacceptable; on the other hand, larger windows include larger data samples, yielding higher precision. Therefore, SSTs could be used to choose the largest b for which a hypothesis test fails to reject the hypothesis of covariate balance.

Figure 1B 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 since they lead to more parsimonious models and more precise estimates.

2. THE SETUP, IN GENERAL

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, 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 procedure 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

$$\bar{d}_\alpha \equiv max\{d : p_d > \alpha\}.$$

That is, $\bar{d}_\alpha$ is the largest value of d for which the null hypothesis that $\mathcal{A}$ is true for $d \leq \bar{d}_\alpha$ 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—choosing $\bar{d}_\alpha < d^*$, is bounded by α. $\bar{d}_\alpha$ 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 $d' = \bar{d}_\alpha - 1$—the first d' for which $p_{d'} \geq \alpha$; reject all null hypotheses $H_{0d'}$ for which $d' \geq \bar{d}_\alpha$, and fail to reject the rest. This protects the family-wise error rate of α since rejecting *any* true null implies rejecting the first true null—a probability α event.

Another common choice for $\hat{d}$ [e.g. Lütkepohl, 2005], say $\underline{d}_\alpha$, does not have this

property. Let

$$\underline{d}_\alpha \equiv min\{d : p_d < \alpha\} - 1 \tag{1}$$

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

2.1 Model Selection and the Logic of Null Hypothesis Testing

In order to avoid certain methodological mistakes, it may be helpful to clarify some of the conceptual distinctions between SSTs and conventional null hypothesis tests (NHTs). The logic of NHTs is familiar to anyone who has taken (and understood) even the most basic college statistics course; nonetheless we restate it here to distinguish it from the logic of SSTs. Typically, researchers use NHTs to reject a null hypothesis that they consider uninteresting—most of the time, that a model parameter is equal to zero—and interpret rejection as evidence in favor of an interesting alternative hypothesis. NHTs cap the probability of a type-I error—falsely rejecting a true null hypothesis—and, given that constraint, seek to minimize the probability of a type-II error, failing to reject a false null hypothesis.

SSTs reverse some of these elements; most importantly, the goal of SSTs is to identify specifications in which an assumption $\mathcal{A}$ is plausible, rather than to identify true alternative hypothesis. In the same vein, type-II errors are typically of more concern for SSTs than for typical NHTs, and type-I errors are less problematic. In fact, a type-II error from a specification test could lead a researcher to fit a misspecified model, which in turn may inflate the probability of a type-I error in her final outcome analysis. For that reason, some methodologists recommend setting α substantially higher for specification tests than for NHTs 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. For instance, take the common $H_0 : \theta = 0$, for some parameter $\theta \in \mathbb{R}$. In finite samples, for any type-I or type-II error rate, there will always be some plausible alternative hypothesis $H_a\theta = \epsilon \neq 0$. Further, in these situations, finite sample estimates $\hat{\theta}$ will almost surely be non-zero. This is important to state to avoid misinterpretations of SST procedures as providing evidence, or showing, that an assumption $\mathcal{A}$ is true for certain specifications d. A common Bayesian argument (e.g. Kadane, 2011, p. 439; Gelman, 2004a) 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].

"But some are useful." In practice there is much to be gained by considering assumptions such as $\mathcal{A}$. In this framework, it may indeed make sense to identify a set of specifications d for which $\mathcal{A}$ is plausible, or approximately true, and SSTs can be useful in this regard—as long as they are understood correctly, and not as providing evidence *for* $\mathcal{A}$.

In many scenarios the choice of d involves a bias-variance tradeoff: 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 RDD 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 SST approach may be seen as an attempt to hold bias at approximately zero, and minimize variance under that constraint. Granted, this is an overly-optimistic take on model fitting; still, SSTs hope to

constrain bias to be approximately zero, and from there minimize variance.

2.2 More Reservations with Null Hypothesis Testing for Model Selection

Applying a strict hypothesis-testing framework to SSTs for model selection has some additional drawbacks. First, it requires researchers to choose a test-level α. While using tuning parameters to mediate the bias-variance tradeoff is not uncommon in statistics, the level α is a particularly hard parameter to choose.

Granger et al. [1995] poses an additional problem with the use of hypothesis tests to choose a model: the need to specify a null hypothesis. In their words (p. 179),

> Whenever a hypothesis test is used to choose between two models, one model must be selected as a null hypothesis. In most instances, this is usually the more parsimonious model and typically a nested test is applied. Often it is difficult to distinguish between the two models because of data quality (multicollinearity, near-identification, or the models being very similar such as in testing for integration). In such cases, the model chosen to be the null hypothesis is unfairly favored.

In other words, because of the structure of null hypothesis tests, which constrain the type-I error rate, the null model is unfairly favored. In our terminology, $\hat{d}$ is likely to be too small, perhaps $\mathbb{E}\hat{d} < d^*$. However, such a bias (if it indeed exists) needn't doom SSTs—an underestimated $\hat{d}$ is merely sub-optimal. In our setup, choosing $\hat{d}$ to be too low will yield and inefficient, but still valid, model. Would that every statistical model were valid yet suboptimal!

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

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 $t < d^*$, $\mathbb{E}x_t = \tau_0$, a constant; for $t > d^*$, $\mathbb{E}x_t > \tau_0$. They propose an estimate of d_0 based on p-values p_t testing the hypotheses $H_{0t} : \mathbb{E}x_t = \tau_0$. They note that for $t < d^*$, the null hypotheses are true, so $p_t \sim U(0,1)$, and $\mathbb{E}p_t = 1/2$; 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^*:

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

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:

$$\hat{d}_M = arg\max_{d} \sum_{t\leq d}(p_t - 1/4). \tag{2}$$

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 any set of p-values from sequential tests: $\hat{d}_M = argmax_d \sum_{t\leq d}(p_t - 1/4)$ may be considered an estimate of d^*. In the case of SSTs, for $d \leq d^*$, p-values p_d are draws from a $U(0,1)$ distribution, and hence have mean 1/2, and, as n_d or N increase, $p_d \to_p 0$ for $d > d^*$. 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 $\bar{d}_\alpha$, 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 $p_d^* - 1/4 < 0$, $\hat{d}_M \neq d^*$, since $\sum_{d \leq d^*-1}(p_d - 1/4) > \sum_{d \leq d^*}(p_d - 1/4)$. Since $Pr(p_d^* - 1/4 < 0) = 1/4$ regardless of sample size, $\hat{d}_M$ will be conservative in large samples. The difference between the SST case discussed here 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 $d = 1, 2, \ldots, D$ is discrete and held fixed in the asymptotics.

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

PROPOSITION 2. *$\hat{d}_M \leq \bar{d}_{0.25}$, with $Pr(\hat{d}_M < \bar{d}_{0.25}) > 0$.*

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

In general, the difference between $\bar{d}_\alpha$ 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 $\bar{d}_\alpha$, but $\hat{d}_M$ relies on the entire set of p-values.

3.1 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 SSTs 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:

$$\hat{d}_M^{ab} \equiv arg \min_{\hat{d} \in \mathbb{N}; 0<b<a<1} \sum_{d \leq \hat{d}} (p_d - a)^2 + \sum_{d > \hat{d}} (p_d - b)^2 \tag{3}$$

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 small simulation study to compare the behavior of model selectors $\bar{d}_\alpha$, $\underline{d}_\alpha$, $\hat{d}_M$, and $\hat{d}_M^{ab}$ in finite samples.

In the simulation, a researcher tests a sequence of hypotheses

$$H_d : \mathbb{E}[X| - d < r < 0] = \mathbb{E}[X|0 < r < d] \tag{4}$$

for a covariate X and a sequencing variable $r \in [-30, 30]$. This might arise if the researcher wants to identify the largest possible region around the origin in which X is balanced around 0.

For every run of the simulation, H_d is true for $d \leq 10$, and, with one exception, false for $d > 10$. In other words, the optimal choice for d—the largest d satisfying (4)—is $d^* = 10$. The simulation runs differ via two factors: the first factor is sample size, the number of samples available at each value of r: $n = \{10, 50, 100\}$. The second factor concerns departures from H_d for $d > 10$, and is illustrated in Figure 2: in simulations with linear imbalance, departures from H_d for $r > 10$ are linear in d, with $\mathbb{E}[X|r] - \mathbb{E}[X| - r] = 2\beta(r - 10)$, where β =0.1. In simulations with sinusoidal imbalance, departures from H_d for $r > 10$ are sinusoidal. In the latter case, H_d is actually true for $d = 30$, which will give rise to radically different behavior between $\underline{d}_\alpha$ and the other model selectors.

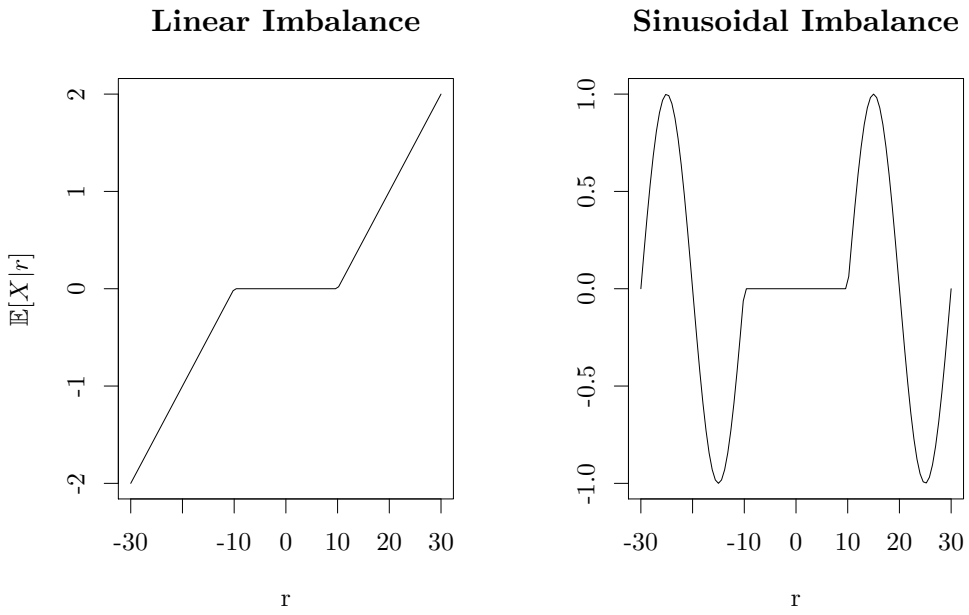

FIG 2. *Two designs for the simulation study.*

In each dataset, the researcher tests H_d, $d = 1, ..., 30$ via t-tests and records the p-values. These p-values, in turn, give rise to eight different choices for $\hat{d}$: $\bar{d}_\alpha$ and $\underline{d}_\alpha$, with $\alpha = 0.5, 0.15, 0.25$, $\hat{d}_M$, and $\hat{d}_M^{ab}$.

Figure 3 shows the results of 1000 replications under linear imbalance. Table 1 shows some statistics that may not be apparent from the figure. For all three sample sizes, the model selectors tended to select samples that were too large. However, as the sample size increased, and with it the power to reject H_b for smaller values of b, the performance of the selectors improved. $\bar{d}_\alpha$ had the smallest variance across the board, but at the price of often choosing the larger datasets. On the other extreme, $\underline{d}_\alpha$ picked very small datasets—for all three sample sizes, $\underline{d}_{0.25} \leq 3$ in over half of the simulation runs.

$\hat{d}_M$ and $\hat{d}_M^{ab}$, on average, performed the best across runs—their average choices were close to $d^* = 10$ in all three runs. However, they were more variable than $\bar{d}_\alpha$, and occasionally chose very small $\hat{d}$. The flexible estimate $\hat{d}_M^{ab}$ was less than or equal to $\hat{d}_M$ in every case, though the two estimates coincided about 70% of the time.

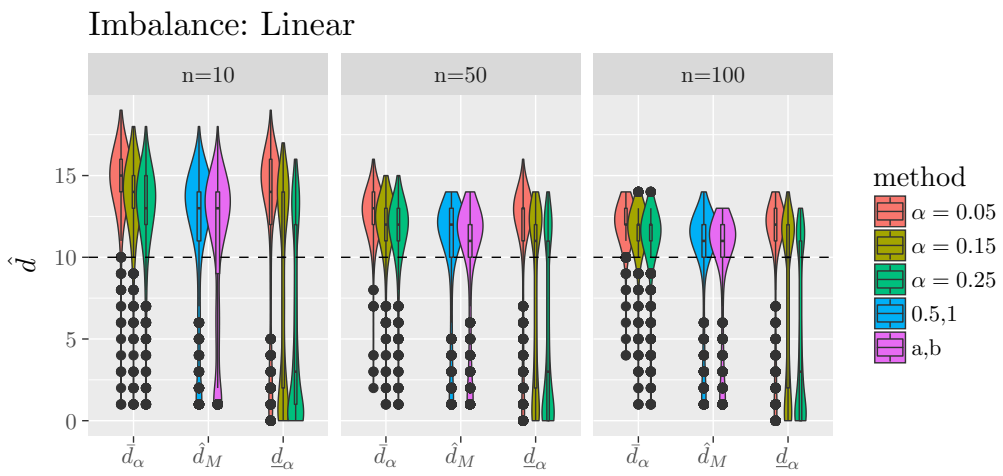

FIG 3. *Simulation results for simulations featuring linear imbalance. The parameter n controls the sample size at each value of r, from -30 to 30. Eight different model selectors are shown for each sample size via overlaid violin and box-plots*

	$n = 10$				$n = 50$				$n = 100$			
	avg	sd	≤ 3	> 13	avg	sd	≤ 3	> 13	avg	sd	≤ 3	> 13
$\bar{d}_{0.05}$	14.7	2.2	0.01	0.79	12.7	1.4	0.00	0.26	12.1	1.1	0.00	0.04
$\bar{d}_{0.15}$	13.5	2.7	0.02	0.60	12.0	1.9	0.01	0.12	11.5	1.7	0.01	0.02
$\bar{d}_{0.25}$	12.8	2.9	0.02	0.47	11.5	2.2	0.02	0.07	11.1	2.0	0.02	0.01
$\underline{d}_{0.05}$	12.3	5.2	0.14	0.65	10.9	4.2	0.13	0.20	10.3	4.0	0.14	0.02
$\underline{d}_{0.15}$	8.1	6.0	0.36	0.29	7.6	5.2	0.34	0.03	7.2	5.1	0.36	0.00
$\underline{d}_{0.25}$	5.7	5.6	0.52	0.14	5.3	5.1	0.52	0.00	5.1	4.9	0.52	0.00
$\hat{d}_M$	11.7	4.3	0.10	0.43	10.4	3.6	0.10	0.05	10.0	3.5	0.12	0.00
$\hat{d}_M^{ab}$	10.8	4.7	0.14	0.34	9.8	4.0	0.15	0.03	9.4	3.8	0.15	0.00

TABLE 1

The average and standard deviation of each selection rule $\hat{d}$, as well as the proportions of runs each method selected $d \leq 3$ or $d > 13$, for $n = 10, 50, 100$ with linear covariate imbalance

When departures from H_d are not monotonic in d, as in the sinusoidal simulation, differences between model selectors can be starker. Figure 4 shows the results from that simulation. Most striking is that $\bar{d}_\alpha$ chooses the maximum $d = 30$ rather frequently. This may not actually be a bad thing—indeed, the mean of the covariate X is equal between positive and negative r when $d = 30$. However, depending on the application, large differences in the mean of X for particular values of r may indicate other departures from model assumptions. Whether the behavior exhibited by $\bar{d}_\alpha$, which occasionally chooses the largest possible d, is preferable to that of $\hat{d}_M$, which rarely does, will depend on the specific data scenario. That said, the results in Figure 4 illustrate the difference between a procedure like $\bar{d}_\alpha$ whose decision may be driven by one individual p-value, and one like $\hat{d}_M$ which

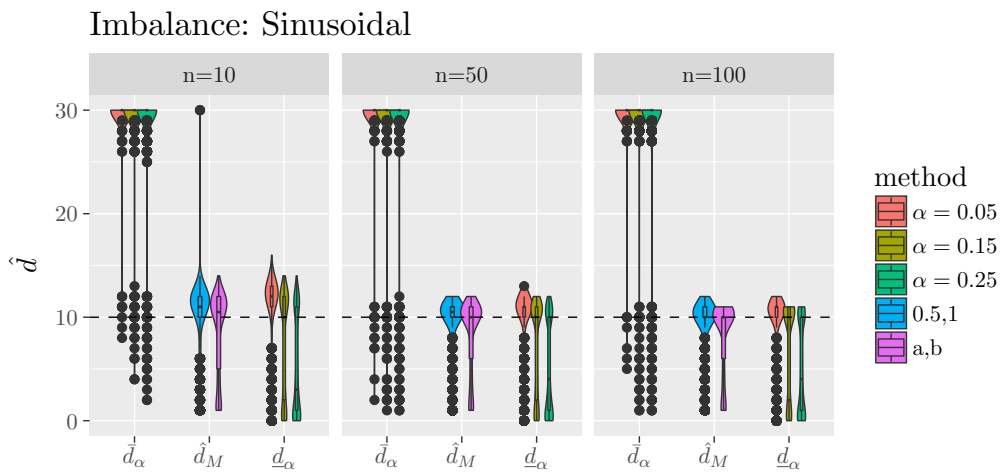

FIG 4. *Simulation results for simulations featuring sinusoidal imbalance. The parameter n controls the sample size at each value of r, from -30 to 30. Eight different model selectors are shown for each sample size via overlaid violin and box-plots*

is driven by the entire distribution of p-values.

5. TWO DATA EXAMPLES

This section will present a more complete treatment of the two examples from Section 1.1: choosing a bandwidth in a regression discontinuity design, and choosing a lag order for an autoregressive model. The two examples correspond to two broad categories of specification: selecting data to analyze and selecting a model specification.

5.1 SSTs 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 are put on academic probation (AP) [See, e.g. Tovar and Simon, 2006]. This provides them access to a set of resources designed to address personal issues that may be hindering their performance. Perhaps more importantly, AP is a threat—students on AP who do not improve are subject to disciplinary measures such as suspension. Lindo et al. [2010] recognized that AP can form a regression discontinuity design (RDD), in which treatment is a function of a "running variable" with a predetermined cutoff. Specifically the treatment Z, students' AP status, is (almost)

a deterministic function of a "running variable" R, students' grade-point-averages (GPAs). Students with a GPA below a pre-determined cutoff, $R < c$, are put on AP. That being the case, students with GPAs just below c may be comparable to students with GPAs just above c—comparing these two sets of students allows researchers to estimate the effect of AP on outcomes Y. The challenge becomes defining "just above" and "just below"; SSTs may be able to play a role here.

For example, Cattaneo et al. [2015] suggests directly comparing the outcomes of subjects with R very close to c, say with $R \in [c - bw, c + bw]$ for some bandwidth $bw > 0$. To choose bw, they use pre-treatment covariates $\boldsymbol{X}$, and covariate balance tests range of candidate bandwidths. For each possible bw, they test the hypothesis that the covariates are balanced:

$$\boldsymbol{X} \perp\!\!\!\perp Z | R \in [c - bw, c + bw] \tag{5}$$

and choose the largest bandwidth in which (5) cannot be rejected[1].

Bandwidth selection for RDDs, and the role of covariate balance tests, encompasses a growing literature. As its name suggests, regression discontinuity typically relies on regression modeling: the goal is to model Y as a function of R on either side of c to estimate the average treatment effect for subjects with R in an infinitesimally-small interval around the cutoff c [See Imbens and Lemieux, 2008]. In contrast, Cattaneo et al. [2015] dispenses with regression altogether. One popular way to ensure robustness to model misspecification is to fit the regression models to a subset of the data with R in a window around c. A number of methods exist to choose an optimal bandwidth bw—the width of the window—that is both large enough to allow for precise effect estimation but small enough to ensure robustness. Imbens and Kalyanaraman [2011] suggest using non-parametric estimates of the curvature of the regression function of Y on R, combined with local linear regression, to choose a bw that minimizes mean-squared-error. However, other authors have suggested choosing bw (or an analogous quantity) based on SSTs, including Li et al. [2015], which presents a Bayesian approach analogous to

[1]Equation (5) is a simplification of Assumption 4 in Cattaneo et al. [2015], which treats x as fixed, not random.

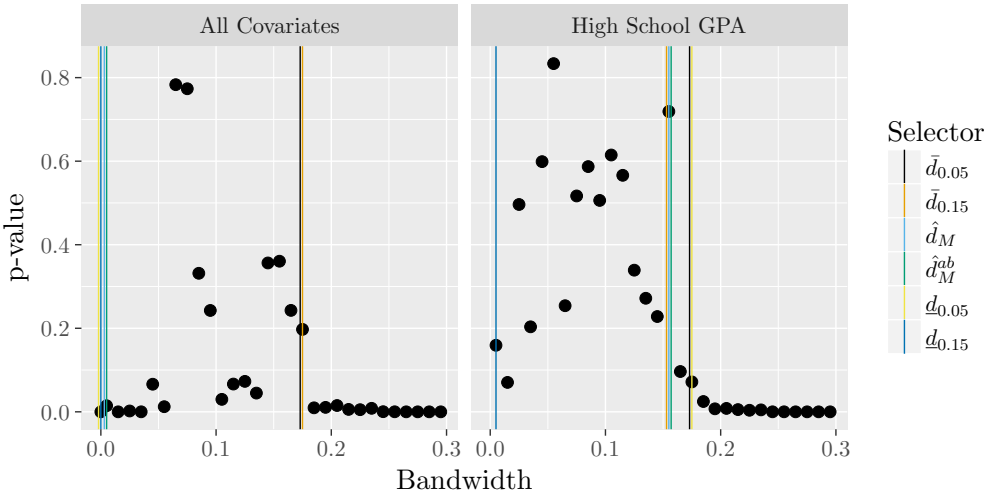

FIG 5. *P-values from for balance in all seven covariates from the Lindo et al. [2010] analysis and only high school GPA, respectively. Vertical lines denote bandwidth choices using different criteria.*

Cattaneo et al. [2015], Sales and Hansen [2014], which discusses the use of robust regression models, and Angrist and Rokkanen [2015], which proposes a method to estimate effects for subjects with R farther from c. In the latter paper, SSTs do not test covariate balance, but the irrelevance of R conditional on covariates X, for subjects in a given bandwidth.

This section will illustrate several approaches to SSTs in the context of estimating the effect of AP for first year college students on subsequent GPAs. For the sake of simplicity, the discussion will be limited to the methods in Cattaneo et al. [2015]; however, many of the SST methods can be extended to other RDD analyses. In their analysis, Lindo et al. [2010] considered a set of seven covariates: students' high-school GPA (expressed in percentiles), age at college matriculation, number of attempted credits, gender, native language (English or other), birth place (North America or other) and university campus (the university consisted of three campuses). A version of Hotellings T^2 test that models treatment assignment Z, and not X, as random [Hansen and Bowers, 2008] is used to test balance. The resulting p-values are plotted in the left panel of Figure 5. Various bandwidth selections are plotted as vertical lines in the figure, and enumerated in Table 2.

	All Covariates			HS GPA		
	d	bw	ATE	d	bw	ATE
$\bar{d}_{0.05}$	18	0.175	0.11 (0.028)	18	0.175	0.11 (0.028)
$\bar{d}_{0.15}$	18	0.175	0.11 (0.028)	16	0.155	0.13 (0.03)
$\underline{d}_{0.05}$	0	N/A	NA (NA)	18	0.175	0.11 (0.028)
$\underline{d}_{0.15}$	0	N/A	NA (NA)	1	0.005	0.22 (0.122)
$\hat{d}_M$	1	0.005	0.22 (0.122)	16	0.155	0.13 (0.03)
$\hat{d}_M^{ab}$	1	0.005	0.22 (0.122)	16	0.155	0.13 (0.03)

TABLE 2

Selected RDD 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.

The p-values from testing balance on all covariates do not follow a monotonic pattern—they are near zero for very small bandwidths, larger for bandwidths greater than 0.035, and close to zero again for bandwidths greater than 0.175. Here, as in the sinusoidal simulation in Section 4, social scientists may disagree about appropriate bandwidth selection. On the one hand, covariates appear to be approximately balanced (or, at least, there is no evidence to the contrary) for a range of bandwidths. On the other hand, the apparent covariate imbalance for small bandwidths is worrying: perhaps it suggests deeper problems with this design—suggesting either using regression to adjust for trends in the running variable, or abandoning the design altogether.

This ambivalence is reflected in the various bandwidth selectors. $\underline{d}_{0.05}$ and $\underline{d}_{0.15}$ reject every possible bandwidth, while both $\hat{d}_M$ and $\hat{d}_M^{a,b}$ select the lowest possible bandwidth of 0. According to these methods, the Cattaneo et al. [2015] method is unsuitable for this dataset. However, the scattered large p-values at some bandwidths lead $\bar{d}_\alpha$ to select larger bandwidths.

To better illustrate differences between the window selection strategies, we consider the covariate high school GPA alone. Since the outcome of interest is itself a GPA, prior measures of GPA are arguably the most relevant and important to control. P-values from tests of balance in high school GPA are displayed in the right panel of Figure 5. Fortunately for the illustration here, high school GPA may be balanced for small bandwidths. These p-values from are more nearly monotonic, appearing roughly uniformly distributed for smaller bandwidths and

close to zero at high bandwidths. The behavior of bandwidth selectors $\hat{d}$, shown with vertical lines in the figure and in Table 2, reflects this feature. The first p-value below 0.05 occurs at a bandwidth of 0.185; accordingly, both $\bar{d}_{0.05}$ and $\underline{d}_{0.05}$ select a bandwidth of 0.175. On the other hand, the p-value of 0.0704 at 0.015 caused $\underline{d}_{0.15}$ to select the smallest possible bandwidth of 0. The change-point selectors, $\hat{d}_M$ and $\hat{d}^{ab}_M$, along with $\bar{d}_{0.15}$, chose a bandwidth of 0.155. In fact, a close inspection of Figure 5 reveals that beginning at a bandwidth of 0.1, the p-values seem to be decreasing—suggesting, perhaps, slight violations of the assumption in (5). $\hat{d}_M$ and $\hat{d}^{ab}_M$ chose a higher bandwidth than this due, in part, to the high p-value at 0.15, which broke the trend, but otherwise may have chosen an even smaller bandwidth. In contradistinction, selectors such as $\bar{d}_\alpha$ and $\underline{d}_\alpha$ that are based entirely on individual extreme p-values cannot account for such patterns.

Table 2 also gives the estimated average treatment effect for each bandwidth. At very small bandwidths, the estimated effect is larger, though with a large standard error as well. The choice between bandwidths at the higher end does not make a large difference in the estimated effects.

5.2 Lag Order in AR(p) 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,

$$unemp_t = \mu + \sum_{i=1}^{d} \phi_i unemp_{t-i} + \epsilon_t \tag{6}$$

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. More generally, we may write (6) as

$$unemp_t = \mu + \sum_{i=1}^{\infty} \phi_i unemp_{t-i} + \epsilon_t \tag{7}$$

with $\phi_i = 0$ for $i > d$.

Having settled on model (6), the analyst must choose d, the lag order. SSTs 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. Of course, there are other options for choosing d, including substantive theory or optimizing information criteria, like AIC or BIC (Akaike 1969, Schwarz et al. 1978; Pötscher 1991 points out that differences in AIC or BIC are essentially likelihood ratio test statistics). In the absence of substantive theory, SSTs 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 SSTs in a well-known area.

Figure 6 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 3, and vertical lines in Figure 6, show the lag order choices from $\bar{d}_\alpha$, $\underline{d}_\alpha$, $\hat{d}_M$, 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 RDD case, which attempted to

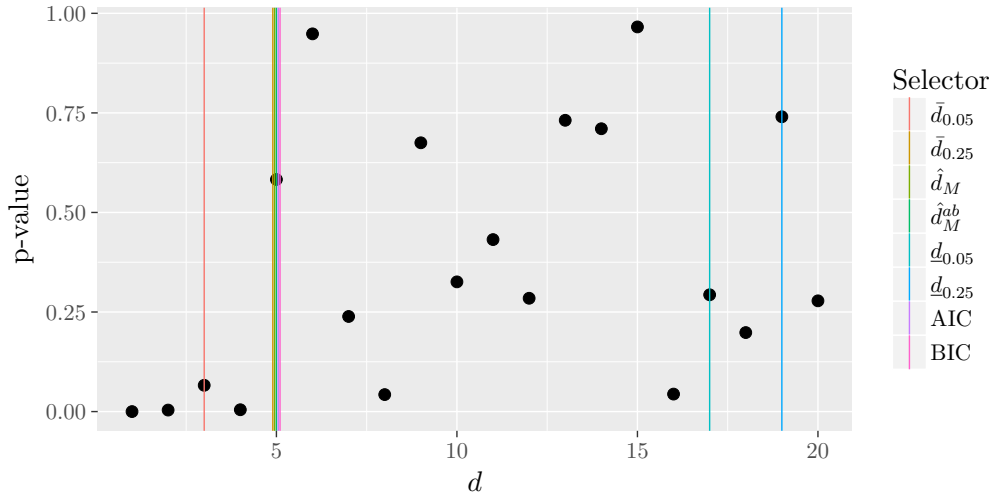

FIG 6. *P-values from sequential likelihood-ratio tests of model fit, comparing models AR(d) with AR(p + 1)o in the annual total US unemployment rate (logged) time series.*

find the *largest* dataset on which to fit the model.

The change-point selectors $\hat{d}_M$ and $\hat{d}_M^{ab}$ both selected a lag order of 5, consistent with 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 $\bar{d}_{0.15}$. In contradistinction, $\bar{d}_{0.05}$ chose a smaller lag order of 3, since the corresponding p-value of 0.066 slightly exceeds the threshold of 0.05.

At the other extreme, the $\underline{d}_\alpha$ 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 $\underline{d}_\alpha$) or large ($\bar{d}_\alpha$) values.

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

TABLE 3
Lag order selections for an AR(d) model of the US unemployment time series.

6. DISCUSSION

The simple intuition behind the use of SSTs is that there exists a set of candidate models that are approximately correct, and another set that are false. By examining the results of specification tests, the thinking goes, a researcher may pick the best model from the former category. However, the negative logic of specification tests makes them poorly suited to finding the boundary between the two groups. Further, they require the choice of a tuning parameter—the level α—and there is little guidance as to how to choose it. Finally, common procedures for choosing a model based on a sequence of hypothesis tests are driven entirely by individual extreme test results, and may, therefore, be unstable.

This paper argues that a clever idea from the change point literature can solve all of these problems. Mallik et al. [2011] suggests that the full distribution of p-values from SSTs may be used to construct an estimator of the optimal true model in the sequence. The estimator they suggest does not require researchers to specify α and is not typically driven by individual extreme p-values. Just as Hodges Jr and Lehmann [1963] showed how to estimate a parameter from a sequence of p-values from hypothesis tests, Mallik et al. [2011] shows how to select an optimal model from a sequence of p-values from specification tests.

This paper examined the performance of the selector from Mallik et al. [2011] and other SST-based selectors via simulation and example, and showed that, in certain ways, it tends to perform better, if only marginally so. In the simulation studies, its average value was consistently closer to the correct value than the other methods. It successfully compromised between the two dominant strategies, denoted here as $\bar{d}_\alpha$ and $\underline{d}_\alpha$, typically avoiding the over-aggressiveness of $\bar{d}_\alpha$ while also avoiding the occasionally extreme conservativeness of $\underline{d}_\alpha$. Its advantage came from its ability to overlook intermittent anomalous p-values that distract $\bar{d}_\alpha$ or $\underline{d}_\alpha$ from an overall pattern.

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 $\hat{d}_M^{ab}$ in larger samples?

Further, there may be ways to construct SSTs in a way that improves $\hat{d}_M$'s performance. For instance, tests that focus on the difference between two successive models, as opposed to each model's overall quality, may, in some circumstances, have higher power to detect departures from assumptions. Such tests would yield p-values with less statistical dependence, which would provide another advantage to $\hat{d}_M$. How to construct such tests, and under what circumstances, if any, they improve model selection, is a topic for future research.

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.

Joshua D Angrist and Victor Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575, 1999.

Joshua D Angrist and Miikka Rokkanen. Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344, 2015.

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.

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), 2004a.

Andrew Gelman. Type 1, type 2, type s, and type m errors. http://andrewgelman.com/2015/05/11/theres-no-such-thing-as-unbiased-estimation-and-its-a-good-thing-too/, (accessed 2/9/17), 2004b.

Clive WJ Granger, Maxwell L King, and Halbert White. Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics*, 67(1):173–187, 1995.

Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

Ben B Hansen and Jake Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, pages 219–236, 2008.

Ben B Hansen and Adam Sales. Comment on cochrans observational studies. *Introduction to Observational Studies and the Reprint of Cochrans 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 practitioners 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 Lemieuxa. 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.

Jens Ludwig and Douglas L Miller. Does head start improve children's life chances? evidence from a regression discontinuity design. Technical report, National Bureau of Economic Research, 2005.

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.

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. *arXiv preprint arXiv:1403.5478*, 2014.

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.

Esau Tovar and Merril A Simon. Academic probation as a dangerous opportunity: Factors influencing diverse college students' success. *Community College Journal of Research and Practice*, 30(7):547–564, 2006.