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

Adam C Sales

Worcester Polytechnic Institute

Department of Mathematical Sciences

Worcester, Massachusetts, USA

asales@wpi.edu

Abstract

A researcher choosing between models ordered by some criterion may seek the best specification that satisfies a testable assumption. In this scenario, sequential specification tests (SSTs) are hypothesis tests of that assumption for each model in the sequence. We introduce a method using the p-values from SSTs to estimate the point in the sequence where the assumption ceases to hold. Unlike alternative approaches, this method is robust to individual errant p-values and does not require choosing a test level or tuning parameter. We demonstrate the method's properties with a simulation study, and illustrate it by choosing a bandwidth in a regression discontinuity design and a lag order for a time series model.

Keywords— Model Selection; Time Series; Regression Discontinuity Designs

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 $t-p$. To choose the order p, 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 on college students' subsequent grade point averages. 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 (RDD; Thistlethwaite and Campbell, 1960), in which treatment is assigned if a numeric "running variable" R falls below (or above) a pre-specified cutoff c. Since treatment assignment is entirely a function of R and c, researchers can model the relationship between R and an outcome variable Y in order to estimate the effect of the treatment without confounding. A common tool for ensuring that RDD models are well specified is to limit the data analysis sample to subjects with $R \in \{c-\omega, c+\omega\}$, where $\omega > 0$ is a bandwidth selected by the data analyst. One method for choosing ω relies on subjects' baseline covariates: researchers will estimate "effects" of

the treatment on baseline covariates using data from subjects with R within ω of c. Since the treatment cannot possibly have an effect on baseline covariates, any estimated effects are due to model misspecification or an overly-large choice of ω. Following this reasoning, some methodologists recommend testing for effects on covariates using an array of candidate bandwidths, and choosing the largest bandwidth within which the null hypothesis of no effect cannot be rejected. The bandwidth tradeoff is similar to the $AR(p)$ case: if ω is too large, the causal model might be misspecified and the effect estimate will be biased. If ω is too small, there will not be enough data to precisely estimate the effect of interest.

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 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

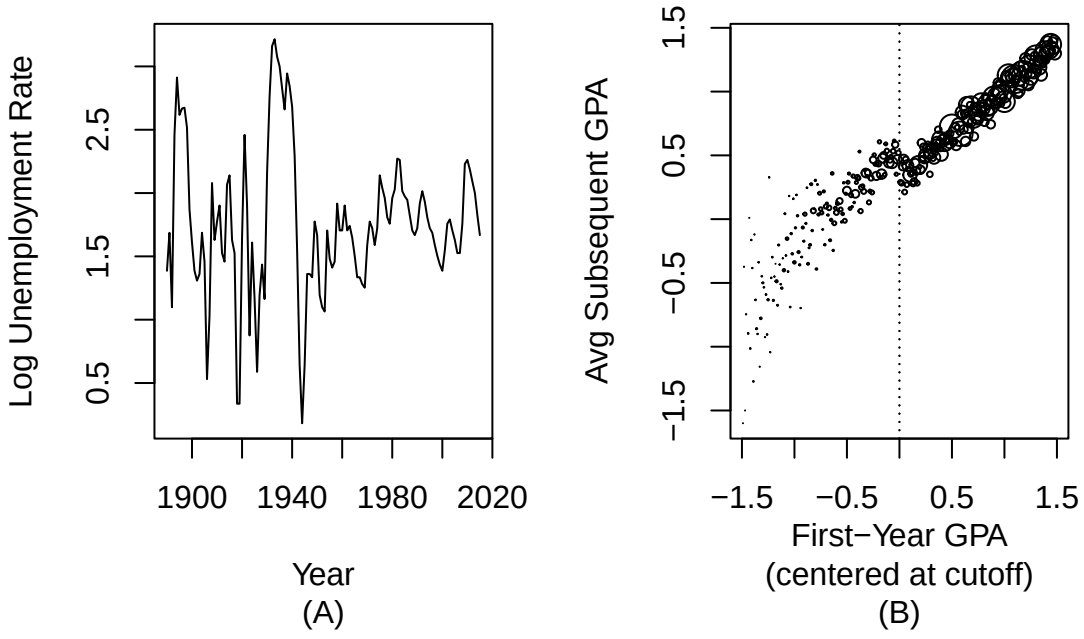

Figure 1: Plot (A) shows a time-series of log annual United States total unemployment from 1890 to 2015. Data were combined from Pfaff (2008) and Bureau of Labor Statistics (2016). Plot (B) shows data from Lindo et al. (2010): average subsequent grade point averages (GPAs), as a function of first-year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

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.

We adapt that idea to the case of SSTs, choosing the change-point in a sequence of models, i.e. the point when models stop being correct. Our change-point estimator is based on the entire sequence of p-values, so that (unlike under current approaches) an individual outlier p-value will not drive its conclusions. 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.

Model selection is a very broad field in statistics with a rich history and—since an appropriate model must be chosen before any data analysis can proceed—it is centrally important. A nice overview can be found in Rao et al. (2001)[1], who state that "[a]lmost all statistical problems can be considered as model selection problems" (p. 3). That monograph focuses on regression problems, and describes model selection based on hypothesis tests (including SSTs) as well as a range of methods that seek to optimize prediction errors, information criteria, or posterior probability for a range of different scenarios. This paper is both more broad and much more restricted in its focus. It is restricted to model selection based on hypothesis tests (more specifically, p-values), and in the particular case in which the researcher is choosing between a sequence of models, ranked by preference. We are not presenting any new hypothesis tests (unlike, e.g. Vuong 1989) or any other way to decide between two competing models (e.g. Schwarz et al. 1978)—instead, we assume that a hypothesis test already exists. Our goal is to provide guidance on how to use p-values from that test to choose between models in a sequence. On the other hand, the method we are introducing will apply in any scenario where such a hypothesis test is available.

[1]This very helpful citation was suggested by an anonymous reviewer.

2 Background

2.1 Sequences of Models and Tests

Say, in specifying a model, a researcher must choose from a discrete set of specifications $\mathcal{S}_d$; $d = 1, 2, \ldots, D$. In this setup, we consider the set of D candidate specifications as fixed, and not dependent on sample size. The models are ordered by preference, subject to a testable assumption operationalized as H_{0d}; that is, if $d > d'$, and H_{0d} holds for both $\mathcal{S}_d$ and $\mathcal{S}_{d'}$, then $\mathcal{S}_d$ is preferable to $\mathcal{S}_{d'}$. The assumption H_{0d} can have the same form for each d (e.g. model $\mathcal{S}_d$ fits the data) or can be relative (e.g. model $\mathcal{S}_d$ fits the data as well as $\mathcal{S}_{d-1}$). Denote the optimal specification choice as $\mathcal{S}_{d^*}$, where

$$d^* = \begin{cases} max\{d \in [1, D] : H_{0d} \text{ is true}\} & \text{if } H_{0d} \text{ is true for some } d \in [1, D] \\ 0 & \text{otherwise} \end{cases}$$

Then we assume the following:

Assumption 1. If $d_2 > d_1$, then H_{0d_2} implies H_{0d_1}

That is, either H_{0d} is false for all $d \in [1, D]$, in which case $d^* = 0$, H_{0d} is true for all $d \in [1, D]$, in which case $d^* = D$, or H_{0d} is true for $1 \leq d \leq d^*$ and H_{0d} is false for $d^* < d \leq D$.

Now assume the researcher has chosen a testing procedure for H_{0d} producing a p-value for each H_{0d}, $p_1, \ldots, p_D$. Let n_d be the sample size included in the test for H_{0d}. This can take a number of forms—for instance, in the unemployment example of Figure 1A, as d increases the model becomes more parsimonious, but the data used to fit the model stays the same, so $n_d = n$ for all d. In contrast, in the RDD example of Figure 1B, a specification with larger d includes more observations, so $d_2 > d_1$ implies $n_2 \geq n_1$. When a dataset has a multilevel or hierarchical structure, the concept of a sample size becomes more subtle. For instance, in a clustered survey sample n_d may refer to the number of clusters included in a hypothesis test, rather than the number of observations. n_d plays no direct role in the computations we will

introduce, so for the sake of generality we can afford to be slightly vague about its meaning. When we wish to emphasize the dependence of p_d on n_d, we write it as p_{nd}. In any event, we assume that the test of each H_{0d} is valid and asymptotically powerful. That is, we assume:

Assumption 2. For each $d = 1, \ldots, D$,

- If $d \leq d^*$, $p_{nd} \sim \mathcal{U}(0,1)$
- If $d > d^*$, $p_{nd} \to_p 0$ as $n_d \to \infty$

so that when H_{0d} is true, the associated p-value is uniformly distributed, and when H_{0d} is false the resulting p-value should be small in large samples. In some cases, it will suffice to substitute the following weaker form of Assumption 2:

Assumption 2′. For each $d = 1, \ldots, D$, as $n_d \to \infty$,

- If H_{0d} is true, $E[p_{nd}] \to 1/2$
- If H_{0d} is false, $p_{nd} \to_p 0$

The goal here is to use $\boldsymbol{p}_D$ to choose a specification $\hat{d}$ that is as large as possible without violating the specification assumption encoded in H_{0d}.

For the methods we describe here and in the following section, it is not necessary for the p-values $\boldsymbol{p}_D$ to be mutually independent; indeed, they typically are not.

2.2 Two Existing Approaches to Sequential Specification Tests

In the contexts of sequential specification tests, Rao et al. (2001) suggests a common approach to choosing a specification: for a pre-specified $\alpha \in (0,1)$, let

$$d_\alpha^{max} \equiv \begin{cases} \max\{d : p_d \geq \alpha\} & \text{if } \max p_d \geq \alpha \\ 0 & \text{otherwise} \end{cases}$$

That is, d_α^{max} is the largest value of d for which H_{0d} cannot be rejected at level α (and 0 if no such d exists). Although it may seem as though multiplicity corrections may be necessary here, 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, and the probability of falsely rejecting any null hypothesis H_{0d} is bounded by α:

Proposition 1 (Hansen and Sales 2015). Under Assumptions 1 and 2, $pr(d_\alpha^{max} < d^*) \leq \alpha$

The proof can be found in Hansen and Sales (2015). Informally, note that (1) if $d^* > 0$, selecting $d_\alpha^{max} < d^*$ entails rejecting H_{0d^*}, (2) that H_{0d^*} is true, and (3) that under Assumption 2, the probability of rejecting a true null at level α is equal to α. If $d^* = 0$, then $d_\alpha^{max} < d^*$ is impossible by definition. d_α^{max} is the specification that would result from testing null hypotheses backwards: for $d' = D, D-1, \ldots, d, \ldots, 1$, test $H_{0d'}$, and stop testing at $d' = d_\alpha^{max} - 1$, the first d' for which $p_{d'} \geq \alpha$.

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 ($d_\alpha^{min} = 0$ if all p-values are $< \alpha$). 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.

Both d_α^{max} and d_α^{min} require the researcher to choose a rejection level α in advance, typically without much guidance or motivation. Also, both procedures are susceptible to outlier p-values: an errant p-value exceeding α for a large value of d will cause d_α^{max} to be too large, and an errant low p-value for a low value of d will cause d_α^{min} to be too small.

In the following section, we will suggest an alternative way to select a specification from

a sequence of p-values that avoids these pitfalls by relying on the logic of estimation instead of the logic of hypothesis testing.

3 The Change Point Estimator

The asymptotic behavior described in Assumption 2′—when $d \leq d^*$, $Ep_{dn} = 1/2$ and when $d > d^*$, $p_{dn} \to_p 0$—suggests a least-squares estimator for d^*:

$$\hat{d}_M \equiv \arg\min_d \sum_{t \leq d} (p_t - 1/2)^2 + \sum_{t > d} p_t^2 = \arg\max_d \sum_{t \leq d} (p_t - 1/4). \tag{1}$$

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.

Because p-values are inherently unpredictable when the null hypothesis is true, even in large samples $\hat{d}_M$ may choose a sub-optimal specification—i.e. $pr(\hat{d}_M < d^*) > 0$ for all n. In particular, note that under Assumption 2, $pr(p_{d^*} < 1/4) = 1/4$ regardless of n_d, and that when $p_{d^*} - 1/4 < 0$, $\hat{d}_M \neq d^*$, because $\sum_{d \leq d^*-1}(p_d - 1/4) > \sum_{d \leq d^*}(p_d - 1/4)$. However, $\hat{d}_M$ is asymptotically conservative:

Proposition 2. Under Assumptions 1 and 2′, $pr(\hat{d}_M > d^*) \to 0$ as $\min_{1 \leq d \leq D} \{n_d\} \to \infty$.

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

If $\hat{d}_M > d^*$, then specification $\mathcal{S}_{\hat{d}_M}$ violates the assumption encoded in H_{0d}; as sample size increases, the probability of this event decreases to zero. The same property holds for d_α^{max}, with $\alpha > 0$ fixed, for the same reason.

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 3. $\hat{d}_M \leq d_{0.25}^{max}$, 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.

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

In finite samples, p-values from tests of false null hypotheses will typically be greater than 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:

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

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.

	$\beta = 0.5$			$\beta = 2$			$\beta = 3$		
Estimator	RMSE	%Opt.	%$> d^*$	RMSE	%Opt.	%$> d^*$	RMSE	%Opt.	%$> d^*$
$d^{max}_{0.05}$	15.3	0	100	0.6	47	50	0.2	82	14
$d^{max}_{0.25}$	8.5	5	93	0.6	60	20	0.6	73	3
$d^{min}_{0.05}$	6.3	6	65	1.8	39	37	1.4	67	11
$d^{min}_{0.25}$	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 $\beta = 0.5, 2, 3$, the root-mean-squared error (RMSE) of each estimator $\left\{\overline{(\hat{d} - d^*)^2}\right\}^{1/2}$ and the percentages each estimator chose the optimal model (%Opt.) or chose a misspecified model (%$> d^*$)

4 A Simulation Study

This section will present a simulation study to compare the behavior of model selectors d^{max}_α, d^{min}_α, 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 following a linear gradual change model (c.f. Vogt and Dette, 2015; Shao et al., 2016). 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 $Z \sim \mathcal{N}\big\{\beta(d-5), 1\big\}$, where the slope parameter β 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 and Table 1 give the results of the simulation study, comparing $\hat{d}_M$ to $d^{min}_{0.05}$, $d^{min}_{0.25}$, $d^{max}_{0.05}$ and $d^{max}_{0.25}$, respectively. Table 1 compares all five model selectors at $\beta = 0.5$, 2, and 3 on three criterion: root mean-squared-error ($RMSE(x) = \left\{\overline{(x - d^*)^2}\right\}^{1/2}$), 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 $d > d^*$ (%$> d^*$).

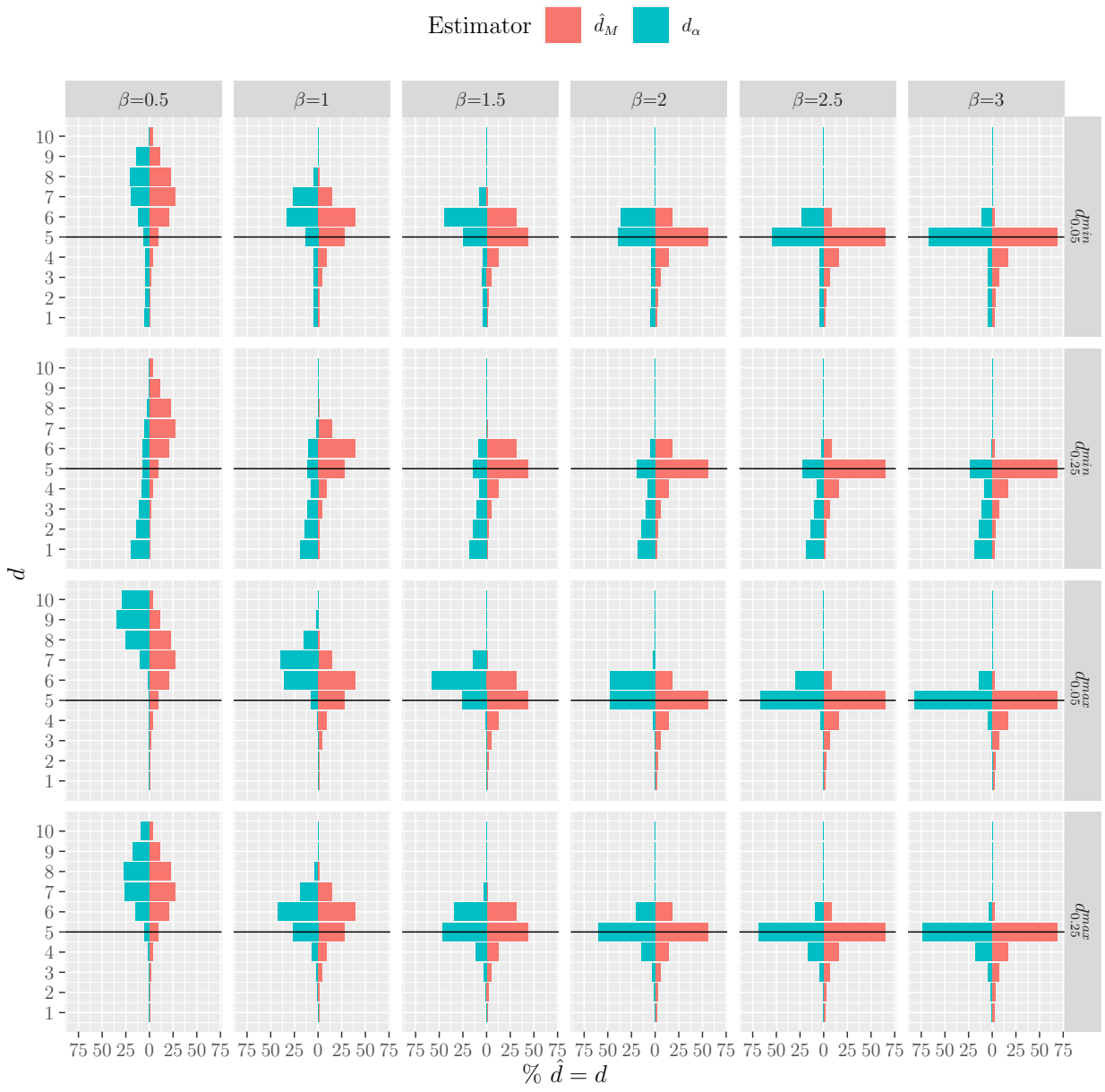

Figure 2: Results from 10^4 simulation runs comparing $\hat{d}_M$ to d^{max} and d^{min} with $\alpha = 0.05$ and 0.25. Each row compares either d^{max} or d^{min} to the same set of $\hat{d}_M$ estimates. Each bar represents the percent of runs in which an estimator selects each possible model, indexed 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.

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 $\beta \in \{0.5, 1, 1.5, 2, 2.5, 3\}$. As β 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 $d^{min}_{0.05}$, $d^{min}_{0.25}$, $d^{max}_{0.05}$, or $d^{max}_{0.25}$.

When $\beta = 0.5$, the power to detect misspecification for models $d > d^*$ is relatively low. $\hat{d}_M$, $d^{min}_{0.05}$, 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^*. $d^{min}_{0.25}$, which is the least likely to recommend a misspecified model, tends to recommend $d = 1$, the smallest possible, least-optimal model.

As β 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 $\beta = 2$, and $\beta = 3$, 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.

The appendix gives a larger version of Table 1 including results from when there are $D = 20$ candidate models or $D = 10$, as here, and when the optimal model $d^* = 2$, $d^* = 5$ (as here), and, when $D = 20$, $d^* = 10$. Broadly speaking, the patterns of performance are similar as D and d^* vary.

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 $2(d+2) - 2log(\hat{L}_d)$, where $\hat{L}_d$ is the maximized likelihood of the $AR(d)$ model, or BIC, which is defined as $log(n)(d+2) - 2log(\hat{L}_d)$. 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 log of the ratio of the 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.

The change-point selectors $\hat{d}_M$ and $\hat{d}^{ab}_M$ 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 $d^{max}_{0.25}$. In contrast, $d^{max}_{0.05}$ 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 and 19, due to the presence of of small p-values of 0.044 and 0.198 at d =16 and 18.

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.

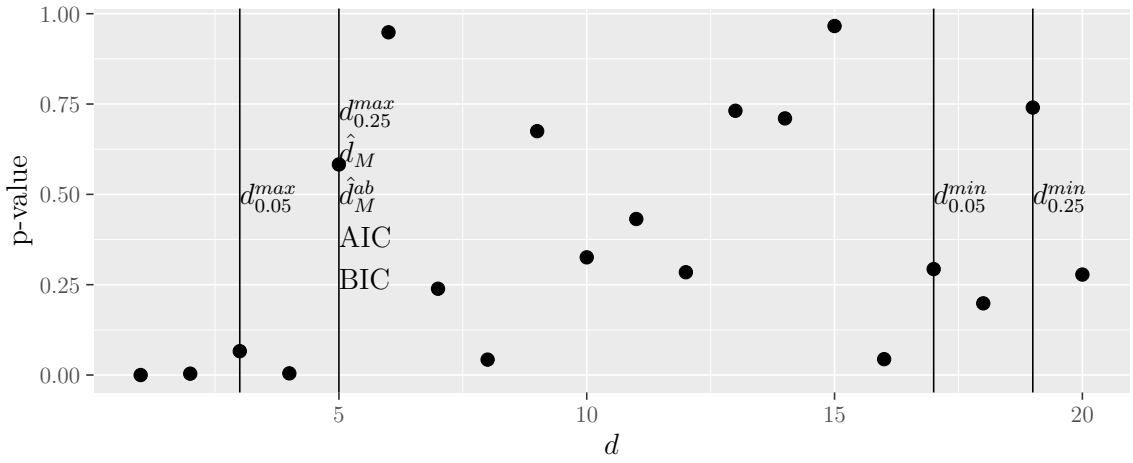

Figure 3: P-values from likelihood-ratio tests of model fit, comparing models AR(d) with AR($d+1$) in the annual total US unemployment rate (logged) time series.

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" $\omega^* > 0$ such that subjects i with $R_i \in (c - \omega^*, c)$ are suitably comparable to subjects with $R_i \in (c, c + \omega^*)$.

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 ω. At a sequence of candidate bandwidths $0 < \omega_1 < \cdots < \omega_d < \cdots < \omega_D$, they recommend testing the equality of covariate means (again, perhaps after adjusting for R) between subjects with $R_i \in (c - \omega_d, c)$ and those with $R_i \in (c, c + \omega_d)$, and choosing a bandwidth $\omega^* = \omega_{d^*}$. 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, 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 ω_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 $R \in (c - \omega_d, c + \omega_d)$. 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 omnibus specification p-value for bandwidth ω_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 $d^{max}_{0.05}$, $d^{max}_{0.25}$, $d^{min}_{0.05}$, $\hat{d}_M$, and $\hat{d}^{ab}_M$. 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 ω_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 ω_1 =0.02, the p-value is p_1 =0.072, and the p-value at the 6th bandwidth, $\omega_6 = 0.07$, another small bandwidth, is $p_6 = 0.009$. After around $\omega = 0.75$, the p-values

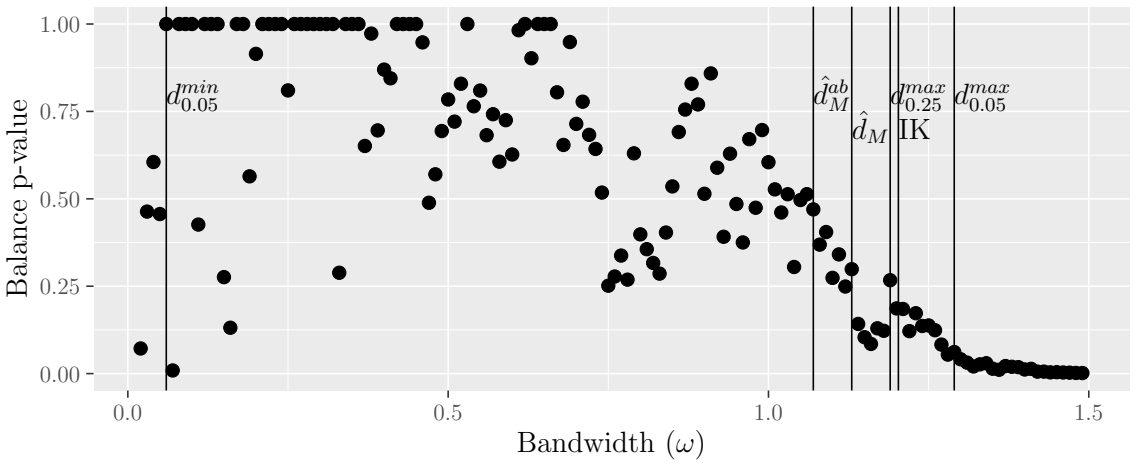

Figure 4: P-values for balance in all seven covariates from the Lindo et al. (2010) analysis, following the method in Sales and Hansen (2020). Vertical lines denote bandwidth choices using different criteria.

begin decreasing, until by $\omega = 1.5$, the p-values are all close to zero.

	$\hat{d}$	Bandwidth	Effect (95% CI)
$d^{max}_{0.05}$	128	1.29	0.22 (0.18,0.26)
$d^{max}_{0.25}$	118	1.19	0.23 (0.18,0.27)
$d^{min}_{0.05}$	5	0.06	0.1 (-0.17,0.37)
$d^{min}_{0.25}$		N/A	N/A
$\hat{d}_M$	112	1.13	0.23 (0.19,0.27)
$\hat{d}^{ab}_M$	106	1.07	0.22 (0.18,0.27)
IK	119	1.2	0.23 (0.19,0.28)

Table 3: Selected regression discontinuity bandwidths ("$\hat{d}$" is the point in the sequence selected, and "Bandwidth" is the actual bandwidth) using covariate balance tests, or using the method described in Imbens and Kalyanaraman (2011), along with their associated estimates for the average treatment effect of academic probation on subsequent GPAs (ATE), with 95% confidence intervals in parentheses.

Since the p-value at the smallest candidate bandwidth, $p_1 = 0.072 < 0.25$, the model selector $d^{min}_{0.25}$ 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 $d^{min}_{0.05}$ 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 $\hat{d}^{ab}_M$,

which recommends bandwidth $\omega_{\hat{d}_M^{ab}}$=1.07 to $d_{0.05}^{max}$, which recommends bandwidth $\omega_{d_{0.05}^{max}}$ =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' subsequent 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 $d_{0.05}^{min}$, all estimated effects are roughly equal, slightly less than 1/4 of a grade point. Actually, the confidence interval corresponding to the $d_{0.05}^{min}$ 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 $d_{0.05}^{min}$ prevents efficient effect estimation; the conservativism of $d_{0.25}^{min}$ 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 threshold. 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_{α}^{min} 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_{α}^{max} does control familywise type-I error rates. However, both threshold-based approaches, d^{min} and d^{max}, require specifying a threshold, and there is rarely any clear guidance on how to do so.

The alternatives introduced here, $\hat{d}_M$ and $\hat{d}_M^{ab}$, 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 threshold to be specified. As shown in the simulation study and the two data examples, they tend to avoid the conservativism and outlier sensitivity of $d_{0.25}^{min}$ and the anti-conservativism of $d_{0.05}^{max}$ (indeed, Proposition 2 states that $\hat{d}_M$ is asymptotically conservative).

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

Model selectors $\hat{d}_M$ and $\hat{d}_M^{ab}$ may be particularly useful in measurement modeling, where model choice based on sequences of p-values are common (such as to select the number of

components in factor or latent class analysis). The encouraging performance of $\hat{d}_M$ when $d^* = 2$ suggests that $\hat{d}_M$ may be appropriate even when the true number of components is small.

References

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

Berger, R. L., Boos, D. D., and Guess, F. M. (1988). Tests and confidence sets for comparing two mean residual life functions. *Biometrics*, pages 103–115.

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

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

Dimmery, D. (2016). *rdd: Regression Discontinuity Estimation*. R package version 0.57.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

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

Imbens, G. and Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, page rdr043.

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

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

Li, F., Mattei, A., Mealli, F., et al. (2015). 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.

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

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

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

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

McQuarrie, A. D. and Tsai, C.-L. (1998). *Regression and time series model selection*. World Scientific.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). 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.

Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R*. Springer, New York, second edition. ISBN 0-387-27960-1.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C., Wu, Y., Konishi, S., and Mukerjee, R. (2001). On model selection. *Lecture Notes-Monograph Series*, pages 1–64.

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

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

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

Shao, C., Li, J., and Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, 81(4):1118–1141.

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

Vogt, M. and Dette, H. (2015). Detecting gradual changes in locally stationary processes. *The Annals of Statistics*, 43(2):713–740.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16.

Zeileis, A., Köll, S., and Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, 95(1):1–36.

A Simulation Results with different D and d^*

Table 4 gives the results of Table 1 for varying values of D (the total number of candidate models) and d^* (the optimal model).

			$\beta = 0.5$			$\beta = 2$			$\beta = 3$		
D	d^*	Est.	RMSE	%Opt.	%$> d^*$	RMSE	%Opt.	%$> d^*$	RMSE	%Opt.	%$> d^*$
10	2	$d^{max}_{0.05}$	21	0.1	100	1	48.6	49	0	81	14.7
		$d^{max}_{0.25}$	10	4.5	94	0	60.4	19	0	73	3.1
		$d^{min}_{0.05}$	8	7.2	83	1	47.6	43	0	77	13.2
		$d^{min}_{0.25}$	2	17.7	39	0	46.4	11	0	55	1.8
		$\hat{d}_M$	6	9.8	85	0	60.5	18	0	73	2.7
	5	$d^{max}_{0.05}$	15	0.2	100	1	47.3	50	0	82	14.2
		$d^{max}_{0.25}$	9	4.9	93	1	60.0	20	1	73	3.0
		$d^{min}_{0.05}$	6	6.0	65	2	39.2	37	1	67	11.1
		$d^{min}_{0.25}$	5	7.1	16	5	19.1	5	5	23	0.7
		$\hat{d}_M$	6	8.9	84	1	56.4	18	1	69	2.7
20	2	$d^{max}_{0.05}$	22	0.1	100	1	48.9	48	0	81	15.0
		$d^{max}_{0.25}$	10	5.1	93	0	61.0	19	0	72	3.5
		$d^{min}_{0.05}$	7	7.3	83	1	47.4	43	0	77	13.5
		$d^{min}_{0.25}$	2	17.5	39	0	45.8	11	0	54	1.9
		$\hat{d}_M$	6	10.0	84	0	61.1	18	0	72	3.1
	5	$d^{max}_{0.05}$	22	0.1	100	1	47.8	50	0	81	14.5
		$d^{max}_{0.25}$	10	4.6	94	1	60.1	20	1	73	3.2
		$d^{min}_{0.05}$	8	6.1	72	2	39.3	38	1	67	11.4
		$d^{min}_{0.25}$	5	7.3	17	5	19.2	5	5	23	0.8
		$\hat{d}_M$	6	8.9	84	1	56.4	18	1	69	2.6
	10	$d^{max}_{0.05}$	22	0.1	100	1	47.8	50	0	81	14.8
		$d^{max}_{0.25}$	10	4.8	94	1	60.1	20	1	73	2.9
		$d^{min}_{0.05}$	18	4.7	54	13	31.1	29	12	51	8.6
		$d^{min}_{0.25}$	35	1.9	4	35	4.5	1	34	6	0.2
		$\hat{d}_M$	6	9.4	84	2	56.4	17	2	68	2.5

Table 4: Simulation runs with the total number of models to compare $D \in \{10, 20\}$ and the optimal model $d^* \in \{2, 5, 10\}$ 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 (%$> d^*$)