Qual Quant
DOI 10.1007/s11135-013-9884-7

An unbiased model comparison test using cross-validation

Bruce A. Desmarais · Jeffrey J. Harden

Abstract Social scientists often consider multiple empirical models of the same process. When these models are parametric and non-nested, the null hypothesis that two models fit the data equally well is commonly tested using methods introduced by Vuong (Econometrica 57(2):307–333, 1989) and Clarke (Am J Political Sci 45(3):724–744, 2001; J Confl Resolut 47(1):72–93, 2003; Political Anal 15(3):347–363, 2007). The objective of each is to compare the Kullback–Leibler Divergence (KLD) of the two models from the true model that generated the data. Here we show that both of these tests are based upon a biased estimator of the KLD, the individual log-likelihood contributions, and that the Clarke test is not proven to be consistent for the difference in KLDs. As a solution, we derive a test based upon *cross-validated* log-likelihood contributions, which represent an unbiased KLD estimate. We demonstrate the CVDM test's superior performance via simulation, then apply it to two empirical examples from political science. We find that the test's selection can diverge from those of the Vuong and Clarke tests and that this can ultimately lead to differences in substantive conclusions.

Keywords Model selection · Cross-validation · Kullback–Leibler Divergence · Vuong test · Clarke test

B. A. Desmarais
Department of Political Science, University of Massachusetts—Amherst, 420 Thompson Hall,
200 Hicks Way, Amherst, MA 01003, USA
e-mail: desmarais@polsci.umass.edu

J. J. Harden (✉)
Department of Political Science, University of Colorado Boulder, 136 Ketchum, UCB 333,
Boulder, CO 80309, USA
e-mail: jeffrey.harden@colorado.edu

Published online: 27 June 2013

1 Introduction

When social scientists are faced with two equally plausible theoretical accounts of a process under study, empirical evidence is commonly used to discriminate between theoretical models. More often than not, the assessment of two models is done by comparing their relative abilities in fitting and/or predicting observations of the process under study. Common methods employed for this purpose include those proposed by Vuong (1989) and Clarke (2001, 2003, 2007)—the Vuong and Clarke tests, respectively. Analysts often use these methods to test the hypothesis that two competing models are equally close to the true data generating process (i.e., fit the data equally well).[1] This is a more principled approach than adding all relevant variables and parameters into the same "garbage can" model and allowing them to compete through significance tests (see Achen 2005).[2] However, we show here that, due to their use of the same data to estimate parameters and assess model fit, both the Vuong and Clarke tests are based on biased measures of the difference in the fit of two models. We describe these issues in detail below and propose leave-one-out cross-validation, an out-of-sample method for computing model fit, as a solution.

When models are fit to data, three distinct steps are involved: (1) specification, (2) parameter estimation, and (3) assessment of explanatory power. The motivation for the development of the Vuong and Clarke tests, in addition to countless other fit diagnostics, is to compare the explanatory power of two or more models. Consider two possible specifications for a true model, f_t, named g_1 and g_2. The parameter estimates from g_1 and g_2, along with the sample data, are used to estimate the explanatory powers of g_1 and g_2 for data drawn from f_t. A critical step in calculating the explanatory power of a model is determining what data should be used to estimate parameters and what data should be used to estimate model fit, given the parameters. No matter the choice regarding the data to be used for parameter and explanatory power estimation, because only a sample is available, the explanatory power of a model for data drawn from f_t will be estimated with error. An important question to ask is whether this error is unbiased. If the same data are used to estimate the model and to compute model fit, explanation of some noise in the data will be falsely attributed to explanatory power, and this "double-dipping" in the data will lead to an overly optimistic assessment of explanatory power. In other words, using the same data for parameter estimation and fit assessment leads to a biased (upward) estimate of explanatory power in a finite sample (Ward et al. 2010).

We show below that both the Vuong and Clarke tests suffer from this problem. Both methods are based on finite-sample biased measures of the quantity they are designed to test—the difference in the accuracies of g_1 and g_2 in approximating f_t. This "accuracy" is given by the (negated) Kullback–Leibler Divergence (KLD) from f_t, a quantity we discuss in detail below. The bias results from using the same data to estimate parameters and assess fit. Moreover, we show that the Clarke test is not proven to be consistent for the difference in the fits of g_1 and g_2.

We also demonstrate here that cross-validation can be an effective solution to the problems of the Vuong and Clarke tests. Diebold and Mariano (2002) show that cross-validation can be used in conjunction with any particular measure of model fit chosen by the analyst (e.g., R^2, log-likelihoods) to establish a test of the hypothesis that two models fit the data equally well.

[1] According to Google Scholar, Vuong (1989) has been cited approximately 2,400 times and the relatively more recent work by Clarke (2001, 2003, 2007) has garnered a combined 229 citations.

[2] Moreover, unlike other information-theoretic model comparison criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), the Vuong and Clarke tests can be used to test hypotheses about the equivalence of model fit in the same way one would use the F or likelihood ratio tests with nested models.

This method is widely applicable in that the models under comparison need not be parametric. We show here that in cases where the measure of fit is the log-likelihood of a parametric model, cross-validation can be used to establish a finite-sample unbiased alternative to the Vuong and Clarke tests. Thus, our central contribution is a new non-nested model selection test based on leave-one-out cross-validation of the log-likelihood. This test, which we call the *Cross-Validated Difference in Means* (CVDM) test, has the same limiting distribution as the Vuong test, but is unbiased in finite samples. Below we discuss the problems with the Vuong and Clarke tests in more detail and describe how our CVDM test constitutes a solution to these problems.

2 The problems with in-sample model comparison

In this section we motivate the need for our CVDM test with a discussion of the problems with the Vuong and Clarke tests. First, we describe the relationship between the KLD, which is the measure of distance between two models that serves as the *theoretical* basis for the Vuong and Clarke tests, and the observed value of the log-likelihood function, which is the quantity used to perform the Vuong and Clarke tests. Second, we show (1) that in finite samples the Vuong test does not constitute a test of the null hypothesis that the KLDs of two approximating models from the true model are equal, and (2) that the Clarke test does not test the null of equal KLDs in finite samples, and is not proven to constitute such a test in the limit.

2.1 The Kullback–Leibler Divergence and the log-likelihood

The KLD between two distributions, f and g, denoted $D_{KL}(f||g)$ gives the information lost, in terms of entropy, when f is approximated with g (Kullback and Leibler 1951). The KLD is a concept central to the practice of information-theoretic model selection. Many familiar model selection criteria, including the AIC (Akaike 1974), the Takeuchi Information Criterion (Konishi and Kitagawa 1996), the Generalized Information Criterion (Konishi and Kitagawa 1996), and the cross-validated log-likelihood (CVLL) (Smyth 2000), are designed to provide an unbiased estimate of the KLD of the approximating model from the true model. In addition to the research that uses the KLD as the basis for the derivation of information criteria, numerous other studies have noted the theoretical optimality of KLD minimization as a strategy in classical model selection and estimation (e.g., Amaral and Dunsmore 1980; Hall 1987; Gilula and Haberman 2000). Therefore, it is not surprising that the motivation for both the Vuong and Clarke tests is to determine whether both models under consideration (g_1 and g_2) are, on average, equally close to the true model (f_t) in terms of KLD. Formally, this is a test of whether $D_{KL}(f_t||g_1) = D_{KL}(f_t||g_2)$.

In the case that the two distributions, f and g, are continuous probability density functions, the KLD is given by

$$\begin{aligned} D_{KL}(f||g) &= \int_{-\infty}^{\infty} \ln\left[\frac{f(y)}{g(y)}\right] f(y)dy, \\ &= \int_{-\infty}^{\infty} \ln\left[f(y)\right] f(y)dy - \int_{-\infty}^{\infty} \ln\left[g(y)\right] f(y)dy \\ &= E_f\{\ln[f(y)]\} - E_f\{\ln[g(y)]\}. \end{aligned} \tag{1}$$

The first term on the right hand side of Eq. 1 is the Information (Shannon) entropy of f (Shannon 1948). The second term is the expected value of the log-likelihood of model g evaluated with regard to data drawn from f. The information entropy does not depend on g, and is canceled out in comparison with another model by taking the difference. Moving back to the example with f_t as the true distribution, this is given as $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2)$. Thus, the analyst does not require knowledge of the true distribution (f_t).

Let $\hat{\boldsymbol{\theta}}$ be the parameter estimate and $h_{\hat{\boldsymbol{\theta}}}$ be the sampling distribution of $\hat{\boldsymbol{\theta}}$ when the data $\boldsymbol{y}$ composed of n independent observations are drawn from f. Then

$$\begin{aligned} E_f\{\ln[g(\boldsymbol{y})]\} &= \frac{1}{n}\int\limits_{\mathbb{R}^n}\int\limits_{-\infty}^{\infty} \ln[g(\boldsymbol{y}|\hat{\boldsymbol{\theta}})]f(\boldsymbol{y})h_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}})d\boldsymbol{y}d\hat{\boldsymbol{\theta}} \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n}\ln[g(y_i|\hat{\boldsymbol{\theta}})]\right], \end{aligned} \tag{2}$$

where $y_i \sim f$ and $\hat{\boldsymbol{\theta}} \sim h_{\hat{\boldsymbol{\theta}}}$ and $\hat{\boldsymbol{\theta}}$ is *independent* of y_i $\forall i \in \{1, 2, \ldots, n\}$. The term inside of the expectation in Eq. 2 appears as if it is the average value of the log-likelihood function evaluated at the maximum likelihood estimate (MLE) of g over samples from f. This is not the case. The $\boldsymbol{y}$ at which g is evaluated is independent of the $\hat{\boldsymbol{\theta}}$ input to g. If $\hat{\boldsymbol{\theta}}$ were the MLE, then $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{y}$ would not be independent—$\hat{\boldsymbol{\theta}}$ would be the value of the parameter that maximizes g given the sample of data $\boldsymbol{y}$. Due to the fact that, in the process of ML estimation, $\hat{\boldsymbol{\theta}}$ is always the g-maximizing estimate given the sample of data drawn from f, in finite samples the average value of the observed log-likelihood is greater than (i.e., positively biased) the expected log-likelihood given in Eq. 2 (Akaike 1974).

The correction of this positive bias in the observed log-likelihood as an estimator of the expected log-likelihood is the motivation for the derivation of all of the information-theoretic model selection criteria mentioned above. Correcting it permits the unbiased estimation of $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2)$, which in turn allows selection of the model with the smallest KLD from the true model. The correction applied to arrive at the AIC removes an asymptotic approximation to the bias when g contains f as a special case (Akaike 1974). The corrections applied to arrive at the TIC (for nonparametric estimators) and GIC remove the asymptotic bias, relaxing the assumption that g contains f (Konishi and Kitagawa 1996). The CVLL, which serves as the basis for the CVDM test we propose, is a finite-sample unbiased estimator of the expected log-likelihood (Smyth 2000), and can therefore be used to directly estimate $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2)$. Prior to deriving our test, we next turn to the biases inherent in the Vuong and Clarke tests.

2.2 Finite-sample bias of the log-likelihood and the Vuong test

The Vuong test is a paired z-test of the null hypothesis that the mean difference in individual log-likelihood contributions to g_1 and g_2 is zero (Vuong 1989). Let the ML estimates of the parameters of g_1 and g_2 be $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ respectively. Also, let the difference in log-likelihoods be expressed as $l_i^{(d)} = \ln[g_1(y_i|\hat{\boldsymbol{\theta}}_1)] - \ln[g_2(y_i|\hat{\boldsymbol{\theta}}_2)]$. The Vuong test statistic is the conventional z-statistic computed on the $l^{(d)}$s: $\frac{\sqrt{n}\bar{l}^{(d)}}{\text{sd}(l^{(d)})}$. Since $l_i^{(d)}$ is a consistent (albeit biased) estimator of the difference in expected log-likelihoods, and therefore the difference in KLDs, it can be used to test the null hypothesis that g_1 and g_2 are equally close to f in terms of the KLD. Being one-to-one transformations of independent and identically distributed observations ($\boldsymbol{y}$), the $l_i^{(d)}$ are themselves i.i.d., and thus by the central limit theorem, the Vuong test statistic has a limiting standard normal distribution under the null hypothesis that $E[l^{(d)}] = D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2) = 0$. This *large sample size* result serves as the justification for the Vuong test (Vuong 1989).

In finite samples, the Vuong test constitutes a test for the average difference in the observed log-likelihoods of g_1 and g_2. However, due to the bias in the observed log-likelihoods as estimators of the expected log-likelihoods, it *does not* constitute a direct test of the null hypothesis that $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2) = 0$.[3] However, if the sign of the average difference in observed log-likelihoods is the same as the sign of the difference in expected log-likelihoods, the Vuong test would constitute an indirect test of the null hypothesis that $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2) = 0$. This is surely the case in some instances, but to our knowledge there are no results indicating it is safe to assume that $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2) > 0 \Leftrightarrow E[l^{(d)}] < 0$ in any general class of applications. If there were a correspondence between the signs of the average difference of expected and observed log-likelihoods, there would be no motivation to derive unbiased information criteria, since selecting the model with the maximum log-likelihood would be a method for selecting the model with the minimum KLD.

Consider the following analytic, finite-sample example where $E[l^{(d)}]$ and the difference in expected log-likelihoods $E[l_e^{(d)}]$ are differently signed. The sample of size n is drawn from f, a normal distribution with zero mean and variance τ^2. The two approximating models, g_1 and g_2, are also normal distributions, where the means are estimated and the variances are fixed at σ_1^2 and σ_2^2 ($\sigma_1^2 < \sigma_2^2$). Thus, g_1 and g_2 are reduced to single-parameter distributions where the mean is the single parameter estimated by ML. This example is not particularly applicable, but in most realistic settings there is no closed form for the average expected and/or observed log-likelihoods (Konishi and Kitagawa 1996), so this instance is convenient in its analytical tractability. This example serves as a proof by contradiction that $E[l^{(d)}] > 0 \not\Leftrightarrow E[l_e^{(d)}] > 0$. Specifically, if

$$\frac{\ln\left(\sigma_2^2\right) - \ln\left(\sigma_1^2\right)}{\left[1 + \frac{1}{n}\right]\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)} < \tau^2 < \frac{\ln\left(\sigma_2^2\right) - \ln\left(\sigma_1^2\right)}{\frac{n-1}{n}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)}, \tag{3}$$

then $E[l^{(d)}] > 0$ and $E[l_e^{(d)}] < 0$. See the Appendix for the full proof. To reiterate, although in the limit a test of the null hypothesis that $E[l^{(d)}] = 0$ is equivalent to testing that $D_{KL}(f_t||g_1) = D_{KL}(f_t||g_2)$, this cannot be assumed to be the case with a finite sample.

2.3 Finite-sample model comparison and the Clarke test

Clarke (2007) offers an alternative to the Vuong test that is non-parametric with respect to the distribution of $l_i^{(d)}$. One major advantage of the Clarke test is that its sampling distribution under the null hypothesis is known *in finite samples* regardless of the forms of f_t, g_1, and g_2. The Clarke test constitutes a test of the null hypothesis that the median difference in individual log-likelihood contributions is zero. The Clarke test statistic is

$$\sum_{i=1}^{n} \mathbf{1}\left[l_i^{(d)} > 0\right]. \tag{4}$$

[3] It should be noted that Vuong (1989) is very clear that all of his results are in the limit, focusing on consistency rather than bias.

Under the null hypothesis that the median of $l_i^{(d)}$, denoted $\tilde{\mu}(l_i^{(d)})$, is zero, the Clarke test statistic has a binomial distribution with n trials and a probability of success of 0.5. In the case where there is no difference between the median and mean of $l_i^{(d)}$ (i.e., $l_i^{(d)}$ is symmetrically distributed and has a finite first moment), the Clarke test inherits the finite sample bias that plagues the Vuong test due to its use of the observed log-likelihoods. There is no reason to expect that a test of the null hypothesis that the $\tilde{\mu}(l_i^{(d)}) = 0$ is equivalent to a test of the null that $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2) = 0$, even if the $\tilde{\mu}(l_i^{(d)}) = E[l_i^{(d)}]$.

2.3.1 The influence of skew in the individual log-likelihood differences

The more interesting case to consider is when $\tilde{\mu}(l_i^{(d)}) \neq E[l_i^{(d)}]$ (i.e., the distribution of $l_i^{(d)}$ is skewed). In particular, if the signs of the median of $l_i^{(d)}$ and $E[l_i^{(d)}]$ are different, then the Clarke test will lead to the choice of the model with the larger KLD, even in large samples. An important question, then, is whether it is plausible that $\tilde{\mu}(l_i^{(d)}) \neq E[l_i^{(d)}]$. Note a unique property of the task at hand. When confronted with skewed data, analysts often use a measure of central tendency other than the mean, since resistant measures such as the median will provide an estimate of the center of the distribution that is more typical of the data under study. In the current case, however, since the mean is the *only* measure of central tendency that bears a direct relationship with the quantity of interest as a consistent estimator of $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2)$, there is no theoretical interest in other measures of central tendency.

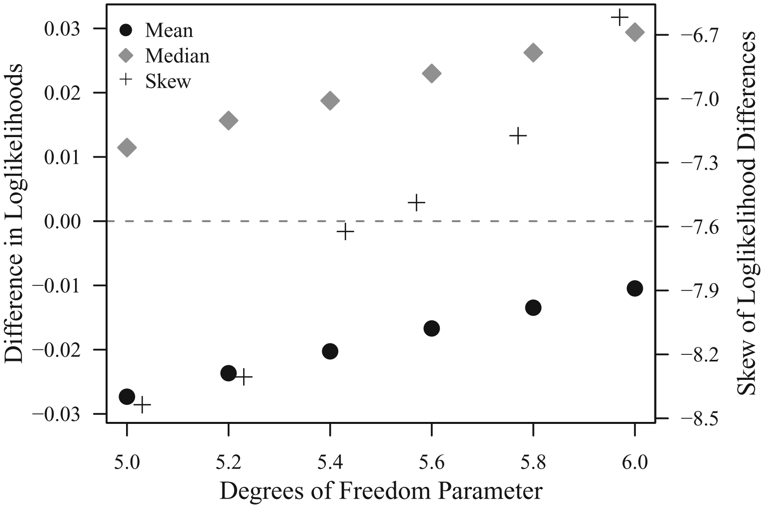
As evidence of this claim, we use simulation to provide an example where $l_i^{(d)}$ has a skewed distribution and the signs of $\tilde{\mu}(l_i^{(d)})$ and $E[l_i^{(d)}]$ are not equal.[4] The design of the simulation is as follows. The model comparison is applied to a regression problem with a continuous and unbounded outcome variable and a single continuous covariate. The sample size is 100, the covariate has a standard normal distribution, and the corresponding regression coefficient is 1 with an intercept of 0. The error term is drawn from a Student's t distribution. The degrees of freedom (df) parameter is varied from 5 to 6 in increments of 0.2. Note that none of these inputs contain any skew.[5] A total of 10,000 replications are performed at each value of df. The two models considered, both estimated by ML, are ordinary least squares (OLS, regression with a normally distributed error term), and regression with a Laplace (or double-exponential) distributed error term (median regression, MR).[6] The covariate is included in both models.

The results from the simulation are presented in Fig. 1. Note that at every value of df over the 10,000 iterations, the average value of $\tilde{\mu}(l_i^{(d)})$ is positive, and the average value of $E[l_i^{(d)}]$ is negative, meaning that $\tilde{\mu}(l_i^{(d)})$ points to OLS as the best-fitting

[4] We discovered this example by starting with two misspecified models and varying parameters in the data generating process until we arrived upon a simulation-based proof that it is possible for the signs of $\tilde{\mu}(l_i^{(d)})$ and $E[l_i^{(d)}]$ to be different.

[5] It may seem odd to see partial degrees of freedom because the t-distribution is often used with reference to the number of observations less the number of parameters estimated (i.e., an integer). However, the t-distribution is a valid probability distribution for any $df > 0$. This interval for df is chosen to produce the divergence in the sign of $\tilde{\mu}(l_i^{(d)})$ and $E[l_i^{(d)}]$.

[6] The Laplace distribution is a symmetric, unbounded continuous distribution that has significantly heavier tails than the normal distribution (Clarke 2007). The MLE of the regression parameters with a Laplace distributed error term is equivalent to the estimate of the coefficients in median regression (Koenker 2005).



Note: The graph shows the difference between $\tilde{\mu}(l_i^{(d)})$ and $E[l_i^{(d)}]$ when normal and Laplace models are estimated with a Student's t error. The mean and medians plotted are the average values of the mean and median difference in log-likelihoods over the 10,000 iterations in the simulation (normal log-likelihood − Laplace log-likelihood). This illustrates that it is possible under the simple condition of a regression model with no skewed inputs to result in a skewed distribution of $l_i^{(d)}$ that renders $\tilde{\mu}(l_i^{(d)})$ and $E[l_i^{(d)}]$ in disagreement regarding the better-fitting model.

Fig. 1 The difference between $\tilde{\mu}(l_i^{(d)})$ and $E[l_i^{(d)}]$ when normal and Laplace models are estimated with a Student's t error

model and $E[l_i^{(d)}]$ points to MR as fitting the data better.[7] Also, note on the right-axis of Fig. 1 that there is negative skewness in the distribution of $l_i^{(d)}$ at each value of df. Since it is an analytically intractable problem, it is not possible to say which estimator actually has the lower KLD. This simulation exercise simply illustrates that it is possible under the simple condition of a regression model with no skewed inputs to result in a skewed distribution of $l_i^{(d)}$ that renders $\tilde{\mu}(l_i^{(d)})$ and $E[l_i^{(d)}]$ in disagreement regarding the better-fitting model. This emphasizes the fact that the Clarke test may, on average, reach a different conclusion than that of the Vuong test, and not the same conclusion with a different power.

Thus, both the Vuong and Clarke tests, though essential tools in selecting between theoretically plausible models, have problems, especially in finite samples. In the next section we derive the CVDM test as a solution to these problems. Specifically, the CVDM test compares the mean CVLL of the two models under study, which, as we show, is a *direct* test of the null hypothesis that the two KLDs are equal *at any sample size*.

3 Cross-validated log-likelihood and the CVDM test

There are two major obstacles to the derivation of a correction to the observed log-likelihood as an estimator of the expected log-likelihood that removes the exact, finite-sample bias. These are (1) the true distribution generating the data is unknown, and (2) even if it were known,

[7] We attempted to depict 95 % confidence intervals around the mean estimates of $\tilde{\mu}(l_i^{(d)})$ and $E[l_i^{(d)}]$ over the 10,000 iterations, but it was impossible to distinguish them on the graph.

due to analytic intractability, it may not be possible to arrive at a closed-form expression for the bias. This is why the most commonly used information-theoretic model selection criteria employ either an asymptotic approximation (e.g., AIC and BIC) or an exact asymptotic estimate (TIC) of the bias. However, the leave-one-out CVLL offers an elegant solution to this problem of estimating the expected log-likelihood. The benefit of using the CVLL is that it provides an unbiased estimate of the expected log-likelihood of each model in samples of size $n - 1$ from f.

Let g be the probability density function, which constitutes the model for $\boldsymbol{y}$. The data are an i.i.d. sample of size n from f, and $\hat{\boldsymbol{\theta}}_{-i}$ is a parameter estimated by ML with the ith observation excluded from the sample. Then the ith contribution to the CVLL is

$$cl_i = \ln\left[g\left(y_i|\hat{\boldsymbol{\theta}}_{-i}\right)\right]. \tag{5}$$

For each observation $i \in \{1, 2, \ldots, n\}$, the data used to estimate the parameters are independent of the data used to estimate the log-likelihood, and the parameter estimate is therefore independent of the draw from f used to compute the CVLL. This independence between y_i and $\hat{\boldsymbol{\theta}}_{-i}$ renders the average value of cl_i an unbiased estimator of the expected log-likelihood (Smyth 2000).

Moving back to the example from above, let the difference in the ith observation's contributions to the CVLLs of models g_1 and g_2 be $cl_i^{(d)} = \left[g_1(y_i|\hat{\boldsymbol{\theta}}_{-i}^{g1})\right] - \left[g_2(y_i|\hat{\boldsymbol{\theta}}_{-i}^{g2})\right]$. Note that $\hat{\boldsymbol{\theta}}_{-i}^{g1}$ and $\hat{\boldsymbol{\theta}}_{-i}^{g2}$ are the MLEs (i.e., the regression coefficients) of g_1 and g_2, respectively, when the ith observation is excluded from the sample. The null hypothesis of the CVDM test is that $E[cl_i^{(d)}] = 0$ (i.e., the mean of $cl_i^{(d)}$ over all i is zero), which is equivalent to the null hypothesis that $E_f\{\ln[g_1(\boldsymbol{y})]\} = E_f\{\ln[g_2(\boldsymbol{y})]\}$, and the null that $D_{KL}(f_t||g_1) = D_{KL}(f_t||g_2)$. Thus, at any sample size, the CVDM constitutes a direct test of the null hypothesis that $D_{KL}(f_t||g_1) = D_{KL}(f_t||g_2)$. In other words, the CVDM can be used to test whether the two models—g_1 and g_2—are equally close to the true model that generated the data.

3.1 The CVDM test statistic

Clarke's (2007) central critique of the Vuong test is that in finite samples, the null distribution of the test statistic is unknown. This is, in part, the motivation for performing a test where the null hypothesis is defined with respect to the sample median. The paired-sign test statistic, which is used for the Clarke test, has a binomial null distribution with n trials and probability of success 0.5 regardless of the parametric form of the distribution under testing (Clarke 2007). There is no test statistic such that its sampling distribution is known under the null hypothesis that the mean is zero, without any information regarding the form of the distribution.

Thus, we cannot advise that analysts test whether the median of $cl_i^{(d)} = 0$ simply because the null distribution of the test statistic is known. The various nonparametric alternatives to the Student's t-test are typically motivated with the objective of testing for a difference in location where the particular measure of location is changed so as to relax assumptions about the located distribution. This is sometimes done using heuristics about the preferable measure of location given the shape of the distribution under study. In this particular case, since the sole objective of testing the location of $cl_i^{(d)}$ is to determine whether its mean is different from zero and thus $D_{KL}(f_t||g_1) \neq D_{KL}(f_t||g_2)$, we cannot compromise on the use of the mean as the measure of location, since no other measure constitutes an unbiased estimator of $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2)$. We may indeed find that the median of $cl_i^{(d)} \neq 0$, but there is no way to determine how such a finding should inform the selection between g_1 and g_2.

Thus, the CVDM test statistic is a skewness-corrected t-test, using the skew correction introduced by Johnson (1978). Another motivation for using the sign test cited by Clarke (2007) is that it achieves greater power than the t-test when the distribution of $l^{(d)}$ has heavy tails (i.e., high kurtosis). The t-test is conservative and has a lower type-I error rate in small samples when the underlying distribution has higher kurtosis than a normal distribution (i.e., leptokurtic, see Chaffin and Rhiel 1993). Of course, by the central limit theorem, as n increases the t-statistic converges to a standard-normal null distribution regardless of the underling distribution of $cl^{(d)}$. We consider accepting the conservativeness of the t-test when the distribution of $cl_i^{(d)}$ is leptokurtic as a reasonable price to pay for the fact that it constitutes a direct test of the null hypothesis that $D_{KL}(f_t||g_1) = D_{KL}(f_t||g_2)$.

However, as is noted above, kurtosis is not the only departure from normality that needs to be considered when designing a test. Just as in the distribution of $l_i^{(d)}$, the distribution of $cl^{(d)}$ can be skewed. Skewness can cause the t-test to have higher type-I error than the nominal level (Chaffin and Rhiel 1993). We apply the correction to the t-statistic derived by Johnson (1978) to remove the bias introduced by skewness. Let $\bar{cl}^{(d)}$ be the sample mean of $cl^{(d)}$. Then the unbiased estimator of the skewness of $cl^{(d)}$ is $\hat{\mu}^3 = n(n-1)^{-1}(n-2)^{-1}\sum_{i=1}^{n}(cl_i^{(d)} - \bar{cl}^{(d)})^3$. The test statistic to be used for testing the null hypothesis that $D_{KL}(f_t||g_1) = D_{KL}(f_t||g_2)$ is

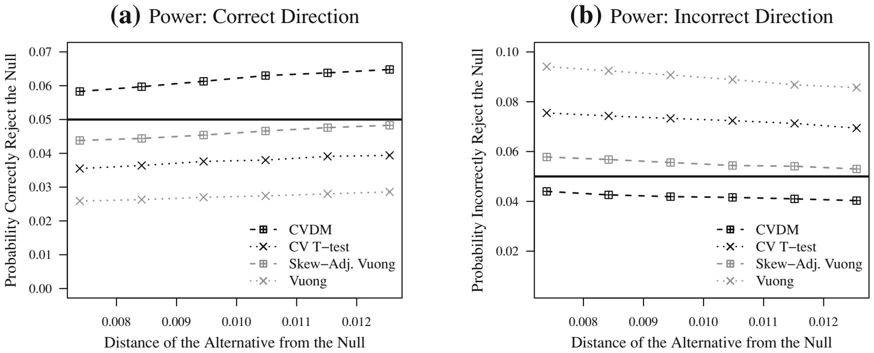
$$\text{CVDM} = \left[\bar{cl}^{(d)} + \frac{\hat{\mu}^3}{6s^2 n} + \frac{\hat{\mu}^3}{3s^4}\left(\bar{cl}^{(d)}\right)^2\right]\frac{s}{\sqrt{n}}, \tag{6}$$

where s is the conventional sample estimate of the standard deviation of $cl^{(d)}$. The CVDM test statistic can be evaluated with respect to the Student's t distribution with $n-1$ degrees of freedom. The CVDM is a direct test of the null hypothesis that $D_{KL}(f_t||g_1) = D_{KL}(f_t||g_2)$. It is conservative when the only departure from normality is leptokurtosis, and due to the application of Johnson's (1978) correction, exhibits better behavior than a conventional t-test when the distribution of $cl^{(d)}$ is highly skewed. It is equivalent to a t-test when there is no skewness.

3.2 Evaluating the CVDM test

Equation 3 gives conditions for the fixed variances of two normal models (σ_1^2 and σ_2^2) and the actual variance of the underlying distribution τ^2 such that the expected values of the observed and expected log-likelihoods are differently signed. Though we show that the Vuong test is biased in this instance, we have yet to demonstrate that the CVDM test performs better. We turn to that task next by evaluating the CVDM test through simulation.

Evidence of improved performance would be demonstrated by the CVDM correction of the bias in the observed log-likelihoods producing a switch in the model that is favored by the test on average. Let $g_1 = \mathcal{N}(\bar{y}, \sigma_1^2)$, $g_2 = \mathcal{N}(\bar{y}, \sigma_2^2)$, and $f_t = \mathcal{N}(0, \tau^2)$. If $\sigma_1 = 1$ and $\sigma_2 = 2$, $\tau \in (1.36, 1.365)$, and $n = 100$, then g_1 has a higher expected observed log-likelihood and g_2 has a lower KLD from f_t (i.e., g_1 is favored by the log-likelihood and g_2 is actually the better-fitting model). We perform 10,000 iterations of this model comparison scenario with τ set at six equally-spaced increments from 1.36 to 1.365, and examine the performance of the CVDM test, Vuong test, the Vuong test with Johnson's (1978) skew correction, and the CVDM without the skew correction, using the 0.05 level of significance (one-tailed).



Note: The graphs show the performance of the model comparison tests when the expected value of the observed log-likelihood favors the model with the higher KLD from the true model. The distance of the alternative from the null is $D_{KL}(f_t||g_1) - D_{KL}(f_t||g_2)$. Since this quantity is always positive, the probability of correctly rejecting the null is the proportion of times in the 10,000 iterations that g_2 is selected at the 0.05 level (one-tailed), and the probability of incorrectly rejecting the null is the proportion of times g_1 is selected at the 0.05 level. The results illustrate the improvements in test performance achieved through the cross-validation of the log-likelihood and Johnson's (1978) skew correction. The CVDM test is the only unbiased test among the four considered in that the CVDM test's power (probability of correctly rejecting the null) is larger than its size (probability of incorrectly rejecting the null).

Fig. 2 Test performance when the expected value of the observed log-likelihood favors the model with the higher KLD

This simulation experiment illustrates the improvements in test performance achieved through the cross-validation of the log-likelihood and Johnson's (1978) skew correction. The results of the simulation are presented in Fig. 2. The CVDM test is the only unbiased test among the four considered in that the CVDM's power (probability of correctly rejecting the null) is larger than its size (probability of incorrectly rejecting the null) (Greene 2008, p. 1036). Moreover, the Vuong test performs the worst of the four; it is the least likely to correctly reject the null and the most likely to incorrectly reject the null. In contrast, the cross-validation of the log-likelihood improves the test performance, increasing the power and lowering the probability of a type-I error with or without the skewness correction. The skewness correction increases the power and lowers the probability of a type-I error regardless of the use of cross-validation of the log-likelihood. Most importantly, the CVDM performs well, favoring the model with the lower KLD, in spite of the fact that that model has, on average, a lower log-likelihood.

4 Applications of the CVDM test

Having shown the problems of the Vuong and Clarke tests and offered the CVDM test as a potential solution, our final objective is to apply the CVDM test to two empirical examples from political science in which the main interest is in selecting between two likelihood-based models.[8] This is a common use for the Vuong and Clarke tests.[9]

[8] Though this may seem somewhat restrictive, note that the general method of cross-validation can be used to conduct model comparison outside of ML estimators (see Diebold and Mariano 2002).

[9] For examples using the Vuong test, see Mebane and Sekhon (2002), Abbe et al. (2003), Mondak and Sanders (2005), Bailey (2007), Shellman and Stewart (2007), and Konisky and Woods (2009). For those employing the Clarke test see Souva (2005), Boockmann (2006), and Travis (2010).

Here we focus on choices for estimation of the linear model with heavy-tailed data. OLS regression is by far the most common method for estimating the linear model in social science. Analysts are often justified in this choice because OLS holds many desirable statistical properties; for instance, it is the minimum variance unbiased estimator (MVUE) when the error term is distributed normally. However, heavy-tailed error distributions can reduce OLS efficiency by masking the true systematic behavior with extreme and atypical random variation, or outliers. This is problematic because inefficiency can produce misleading estimates in a single sample of data, even when the estimator is unbiased. Thus, researchers using OLS with a heavy-tailed error term are more likely to draw inferences that are not warranted by the data. One potential solution is a robust regression (RR) estimator.[10] Outliers do not influence RR as heavily as OLS, and thus RR is more likely to provide estimates closer to the true parameters in heavy-tailed data (Western 1995).

4.1 Robust regression

Before presenting our empirical examples, it is first necessary to define RR and the implementation of it that we use here. First, recall that outlying observations are those that occur with a frequency that departs markedly from expectations. In the context of a continuous and unbounded response variable conditioned on a vector of covariates, the conventional expectation, either implicit or explicit, is that the residual will conform to a normal distribution. RR methods are those that are less influenced by departures from the normality assumption than is the OLS estimator. Whether one has the perspective that outliers exist due to some data contamination such as measurement error, or that the process generating the data has heavier tails than a normal distribution, these robust methods will be more capable of estimating the relationships between independent variables and the center of a response variable than will OLS.

In the examples below, we compare OLS to RR with a t likelihood function, as given in Western (1995, pp. 796–797). Specifically, the ith observation has density

$$p(y_i|\boldsymbol{\beta}_i, \psi, \nu) = \frac{\Gamma(\nu+1)/2}{\psi^{1/2}\Gamma(1/2)\Gamma(\nu/2)\nu^{1/2}} \times \left(1 + \frac{\left(y_i - \boldsymbol{\beta}'\boldsymbol{x}_i\right)^2}{\psi\nu}\right)^{-(\nu+1)/2} \tag{7}$$

where ψ is a dispersion parameter and ν is the degrees of freedom.[11] The t has heavier tails than the normal, and thus by assuming the dependent variable is a conditional t distribution, outliers are effectively downweighted. We use this particular implementation of RR because it is robust to outliers while remaining within the ML framework, thus allowing for comparisons between OLS and RR with our CVDM test.

4.2 Replication examples

We report replication results of two recent articles that use OLS to test linear hypotheses. We report the re-analysis of each model graphically. First, we plot a density estimate of the dependent variable (solid lines) along with a normal density (dotted lines) and a t density

[10] Another possibility is MR, which appears in one of our simulation examples above. For simplicity we only focus on the choice between OLS and RR here, though results do not change if we compare OLS and MR.

[11] Following Lange et al. (1989) and Western (1995), we set $\nu = 4$. However, this parameter could also be estimated from the data (Western 1995). In fact, analysts could use our CVDM test to compare a model in which ν is estimated to a model setting ν a priori.

(dashed lines) with the same parameterizations as that dependent variable.[12] Note that in each case the density estimate of the dependent variable looks more like the density of the estimator selected by the CVDM test. This is done to illustrate the intuition behind the CVDM test. It only constitutes *preliminary* and *informal* evidence and is *not* a substitute for the CVDM itself. We also report the Vuong and Clarke test results to compare to the CVDM results. We then plot the coefficient estimates and 95 % confidence intervals for the original OLS models (in black) and the RR models (in gray).[13]

4.2.1 Voting law reform after the 2000 presidential election

Palazzolo and Moscardelli (2006) examine election reform in the American states in the wake of the 2000 presidential election. They contend that the election served as a "focusing event" that may have galvanized some states to adopt election reforms to prevent future problems like those in Florida in 2000. Building from policy innovation literature, the authors specify an "internal determinants" model, which emphasizes the role of social, political, and economic factors in driving innovation. Their main goal is to test the role of political leaders. They posit that the 2000 election's role as a focusing event gave more power to public officials, and thus increases in action on the part of officials led to more reform in the states.

More specifically, the authors created *Leadership Index*, a measure of the amount of action taken in favor of election reform by a state's chief elections officer, governor, or legislative leaders in 2001 and 2002. This measure comes from content analysis of newspaper articles mentioning either of those officials in connection to election reform. The dependent variable is the number of election reforms enacted by the state during this time, weighted and standardized to correct for differences in the likelihood of passage and amount of reform enacted prior to 2000 (see Palazzolo and Moscardelli 2006, p. 304). Positive values indicate more election reform. The authors expect *Leadership Index* to exert a positive influence—as political leaders take more action regarding election law, more reform occurs. They also include several other variables relating to the political environment in the state, including *Moralistic Culture*, *State Ideology*, and *Term Limits*.

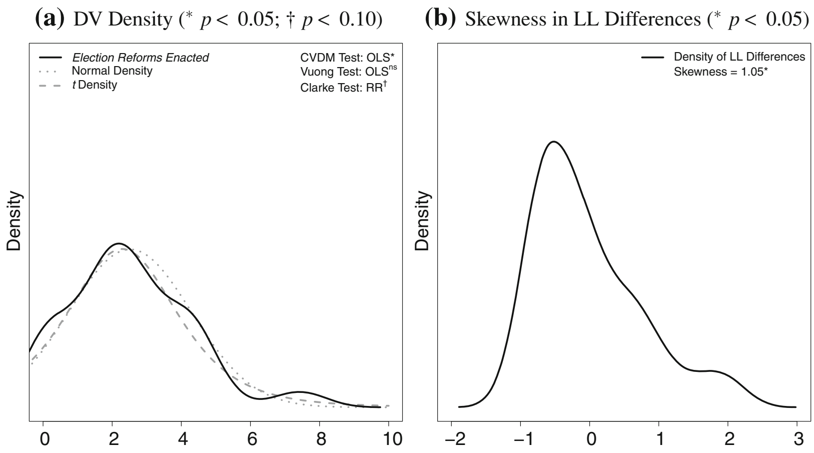
Figure 3a plots the density of the dependent variable along with parameter-matched normal and t densities. This is done to provide an intuitive look at the severity of outliers in these data. Notice that the three lines are very similar to one-another, indicating that an estimator that is less influenced by outliers may not be needed. This is also shown by a small kurtosis value for the dependent variable (3.6). This lack of outliers in the distribution of the unconditioned dependent variable is suggestive of OLS as the better choice. The CVDM test supports this assertion, selecting OLS as the better-fitting estimator at a statistically significant level ($|t| = 3.77$, $p < 0.05$). However, the Vuong and Clarke tests do not show the same result. While the Vuong test points weakly in the same direction as the CVDM test, it is not statistically significant ($|z| = 0.51$, $p = 0.62$). Further, the Clarke test makes a selection in the *opposite* direction (RR) and is statistically significant at the 0.10 level.

Figure 3b provides a possible explanation for this difference in selection between the CVDM and Clarke tests. That graph plots the density of the differences in individual log-likelihoods used to compute the Clarke test. Note that the graph shows these values to be skewed, with a great deal of density in the right tail.[14] Thus, drawing from the discussion

[12] The distributions are fit to the data by ML. The estimated parameters are the mean and variance of the normal distribution and the median and dispersion parameter for the t.

[13] We replicated each model exactly. All coefficients are standardized to allow ease of presentation.

[14] More formally, the skewness of these values is a statistically significant 1.05. The individual *cross-validated* log-likelihoods also exhibit this skewness.



Note: Panel (a) plots the density of the dependent variable along with parameter-matched normal and *t* densities. Notice that the three lines are very similar to one-another, indicating that an estimator that is less influenced by outliers may not be needed. Panel (b) plots the density of the differences in individual log-likelihoods used to compute the Clarke test. Note that the graph shows these values to be skewed, with a great deal of density in the right tail. Thus, from the discussion above, this is likely a case where $\tilde{\mu}(l_i^{(d)}) \neq E[l_i^{(d)}]$, in which the Clarke test leads to the choice of the model with the larger KLD. In contrast, the CVDM test is unaffected by the skewness. In sum, there is good evidence to trust the CVDM test selection of OLS more than the Clarke test selection of RR.

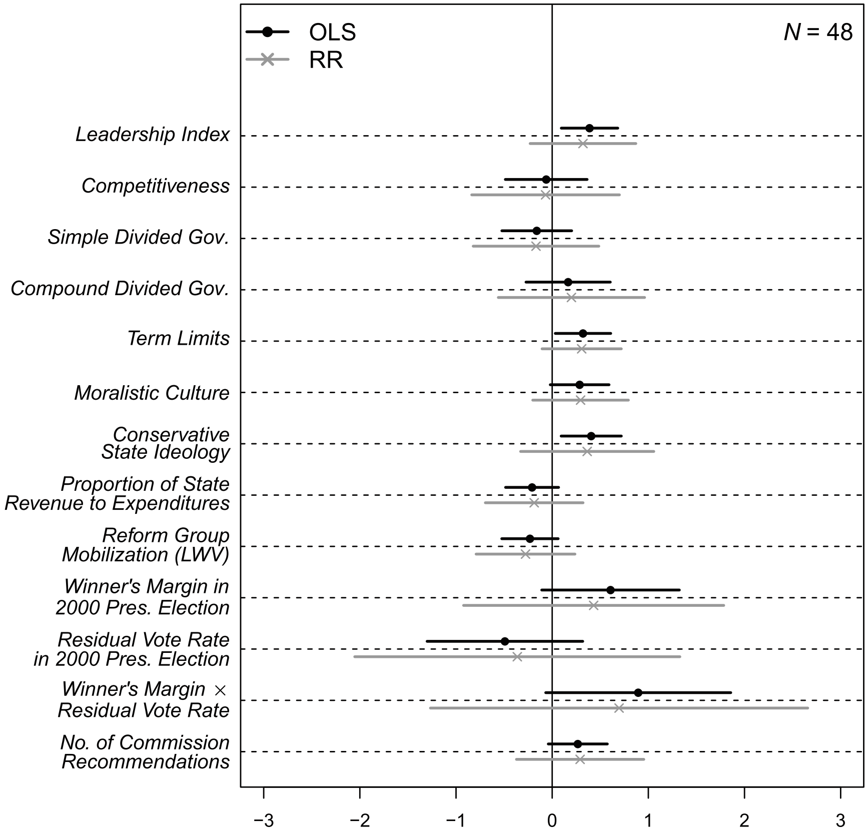
Fig. 3 Dependent variable density and skewness in log-likelihood differences for the Palazzolo and Moscardelli (2006) model

above, this is likely a case where $\tilde{\mu}(l_i^{(d)}) \neq E[l_i^{(d)}]$, in which the Clarke test leads to the choice of the model with the larger KLD. In other words, we have evidence in this instance that the Clarke test is choosing the worse-fitting model. In contrast, the CVDM test is unaffected by the skewness. Thus, there is good evidence to trust the CVDM test selection of OLS more than the Clarke test selection of RR.

Figure 4 plots the model results. Note that the choice between OLS and RR is critical in this case for inference on the authors' key variable of interest, *Leadership Index*. The OLS results show support, producing a positive coefficient on *Leadership Index* that is statistically significant. Thus, the authors conclude that "the activities of elected or appointed policymakers…had a decisive effect on the extent of election reforms adopted by the states" (Palazzolo and Moscardelli 2006, p. 300). In contrast, the RR coefficient on *Leadership Index* is slightly smaller in magnitude and not statistically significant. In short, this is an example in which results are dependent on model selection. Using our CVDM test to make that selection gives strong support for OLS as the better-fitting estimator, and thus more support for the authors' hypotheses. In contrast, the Vuong and Clarke tests produce no selection and the opposite selection, respectively, which leads to less support.

4.2.2 Voter turnout among peasants in Nepal

In our next example, we re-analyze a study on district-level election turnout among peasants in Nepal. Joshi and Mason (2008) provide an explanation of how the large population of subsistence farmers in Nepal, who were active in the Maoist wing of the Nepal Communist Party's 1996 overthrow of the democratic government, consistently fail to provide support for the party at the polls. They posit that patron–client relationships—the extent to which



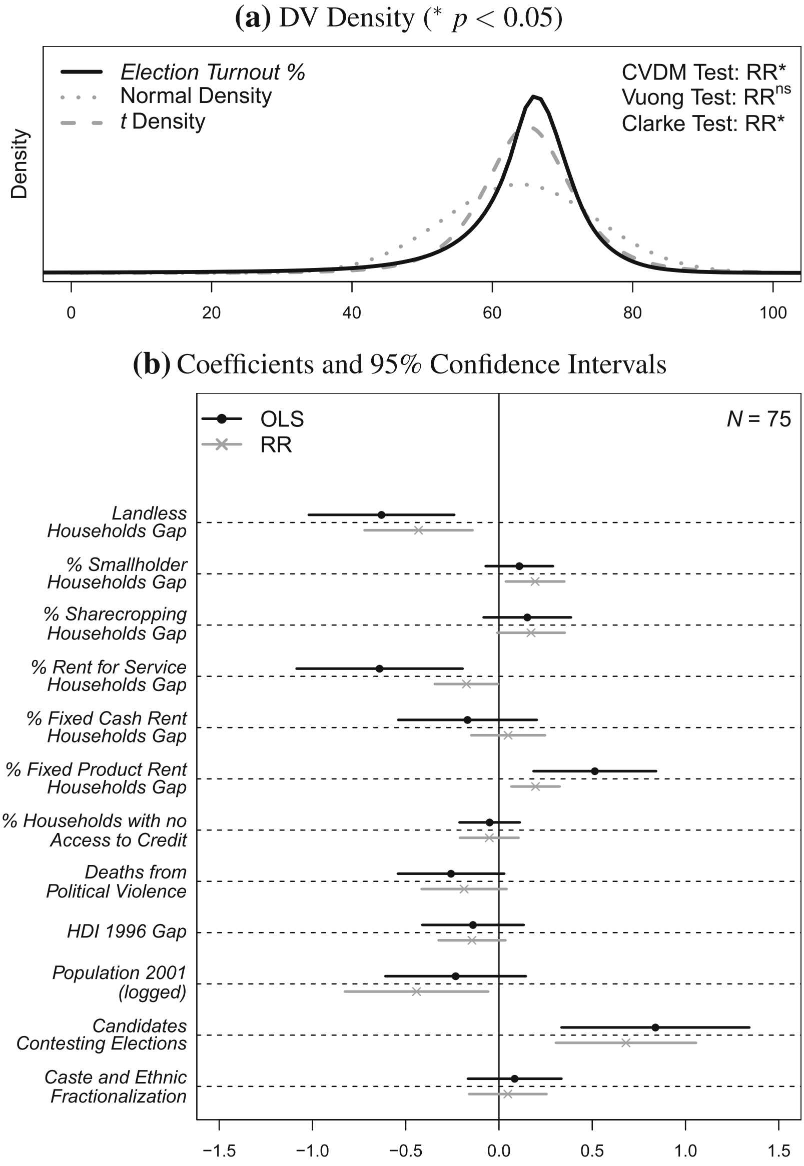
Note: The graph plots coefficient estimates and 95% confidence intervals for the original OLS model (in black) and the RR model (in gray).

Fig. 4 Re-analysis of impacts on election reform in the states, 2001–2002 (Palazzolo and Moscardelli 2006, Table 1)

peasant farmers depend on the local landed elite for subsistence—drive this process. More specifically, they expect that turnout increases in areas where more peasants are dependent on landed elite for survival. While those peasants may be supportive of the parties advocating land and wealth redistribution, it is more important for their immediate interests to go to the polls to support the parties that their patrons support.

The authors test several hypotheses stemming from this theoretical basis, three of which we focus on here. First, they expect that turnout is positively related to the number of tenants farming under a fixed-cash or fixed-product rent system. These tenants pay a portion of their income to the owner of the land on which they farm in exchange for several goods and services. Thus, they have incentive to carry out the wishes of their landlord. Second, the authors posit that in areas with more landless peasants (i.e., those with no patron), turnout decreases, because those not bound to a landowner have no incentives to vote. Finally, the authors expect that areas with more smallholders—those that own a small piece of land—will be less tied to the elite, and thus will be more likely to vote.

Figure 5a plots the distribution of turnout among registered voters in 1999 along with best-fitting normal and t densities. In this case, the dependent variable density estimate corresponds



Note: The graph in panel (a) plots the density of the dependent variable along with parameter-matched normal and *t* densities. Notice that the dependent variable density corresponds closely to the *t* density and is less similar to the normal. The graph in panel (b) plots coefficient estimates and 95% confidence intervals for the original OLS model (in black) and the RR model (in gray).

Fig. 5 Re-analysis of election turnout as a function of land-tenure patterns (Joshi and Mason 2008, Table 4)

closely to the t density and is less similar to the normal. Additionally, the dependent variable kurtosis is 14.7, which provides further evidence that it has heavier tails than a normal. This is consistent with the CVDM test results, which selects RR ($|t| = 2.11$, $p < 0.05$). Thus, RR is likely the better estimator for this model. However, the Vuong and Clarke results are somewhat mixed. While both select RR, the Clarke selection is statistically significant ($p < 0.05$), while the Vuong test statistic is not ($|z| = 1.53$, $p = 0.13$).

Figure 5b shows that each of the hypotheses discussed above has support with the OLS results. The coefficient on *% Fixed Product Rent Households Gap* is positive, the coefficient on *Landless Households Gap* is negative, and both are statistically significant at the 0.05 level. The coefficient on *% Smallholder Households Gap* is positive, as expected, though not statistically significant. However, the better-fitting RR model shows mixed results. In regard to the first two hypotheses mentioned above, coefficient magnitude drops considerably: the RR coefficient on *% Fixed Product Rent Households Gap* is approximately 38 % of the size of the OLS estimate while the RR coefficient on *Landless Households Gap* is about 69 % of its OLS counterpart. However, despite this drop in magnitude, both of these variables remain correctly-signed and statistically significant with RR. In addition, the positive coefficient on *% Smallholder Households Gap* increases in magnitude compared to OLS and is statistically significant at the 0.05 level with RR. Thus, RR provides slightly less support for two hypotheses and more support for another in this model. As in the previous study, this example exhibits model dependence, the CVDM test is helpful in making a choice, and the Vuong test does not make a statistically significant selection.[15]

5 Conclusions

Testing competing models of the same process is a crucial element of quantitative social science. Models are often formulated to represent competing theoretical accounts of the process under study, and model comparison tests can be used to test the hypothesis that one theory better explains the data than another. The methods proposed by Vuong (1989) and Clarke (2001, 2003, 2007) offer innovations in that they test the hypothesis that two competing models are equally close to the true data generating process. However, both of these tests use finite-sample biased measures of the quantity they are designed to test—the difference in the accuracies of two models as approximations of the true model. This bias stems from the fact that both the Vuong and Clarke tests use observation-wise measures of fit that are produced from a model using those same data.

In contrast, our CVDM test, which is based on leave-one-out cross-validation, has the same limiting distribution as the Vuong test, but is unbiased in finite samples because it calculates each observation's log-likelihood value based on a model estimated without that observation. In addition, by using a skew-correction for the t-statistic, the CVDM test addresses Clarke's (2007) critique that the Vuong test is susceptible to skewed differences in log-likelihood contributions. However, the CVDM test also retains the difference in means framework that is a consistent estimator of the difference in KLDs, unlike the Clarke test's paired-sign difference in medians test, which is not shown to be consistent. In short, the CVDM test eliminates the finite-sample bias in the Vuong and Clarke tests, and thus provides a direct and unbiased test of the hypothesis that two competing models are equally close to the true data generating process.

Finally, we demonstrate the utility of the CVDM test in applying it to the choice of linear regression estimators in the face of heavy-tailed data. First, we show that differences between OLS and RR results can produce different substantive inferences from model results, and that the CVDM test can be used to select between the two estimators, thus making a more statistically-sound determination of which conclusions

[15] Although the Clarke test makes the same selection as the CVDM test in this case.

to draw. Second, we show that there can be differences between the CVDM, Vuong, and Clarke tests, including one example (Palazzolo and Moscardelli 2006) where the CVDM and Clarke tests select different estimators entirely at statistically significant levels. In short, the conclusions made from empirical results can depend on both the estimator used and test used to select an estimator. Given the desirable finite-sample properties of the CVDM test over the Vuong and Clarke tests, we recommend that researchers use the CVDM test when comparing two models of the same process.

Appendix: Proof of Vuong test finite sample bias

Here we derive the inequality given in Eq. 3 of the main text. Suppose $\boldsymbol{y}$ is a sample of n independent observations from a normal distribution with zero mean and variance τ^2. Also, let g be a normal probability density function with the mean estimated as the sample mean of $\boldsymbol{y}$, and the variance fixed at σ^2. Then the expected value of the observed log-likelihood is

$$\begin{aligned} E[ll_o] &= E_{\boldsymbol{y}}\left[\frac{1}{n}\ln\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-1}{2\sigma^2}\left(y_i-\frac{1}{n}\sum_{j=1}^{n}y_j\right)^2\right]\right)\right] \\ &= -\ln\left(\sqrt{2\pi\sigma^2}\right)-\frac{1}{2\sigma^2 n}E_{\boldsymbol{y}}\left[\sum_{i=1}^{n}\left(y_i-\frac{1}{n}\sum_{j=1}^{n}y_j\right)^2\right] \\ &= -\ln\left(\sqrt{2\pi\sigma^2}\right)-\frac{\tau^2(n-1)}{2\sigma^2 n}. \end{aligned} \tag{8}$$

The expected value of the expected log-likelihood is[16]

$$\begin{aligned} E[ll_e] &= E_{\bar{y}}\left[E_{\boldsymbol{y}}\left(\frac{1}{n}\sum_{i=1}^{n}\ln\frac{1}{\sqrt{2\pi\sigma^2}}-\frac{(y_i-\bar{y})^2}{2\sigma^2}\right)\right] \\ &= -\ln\left(\sqrt{2\pi\sigma^2}\right)-\frac{1}{2\sigma^2}E_{\bar{y}}\left[E_{\boldsymbol{y}}\left(\frac{1}{n}\sum_{i=1}^{n}y_i^2-2y_i\bar{y}+\bar{y}^2\right)\right] \\ &= -\ln\left(\sqrt{2\pi\sigma^2}\right)-\frac{1}{2\sigma^2}E_{\bar{y}}\left[\tau^2+\bar{y}^2\right] \\ &= -\ln\left(\sqrt{2\pi\sigma^2}\right)-\frac{\tau^2+\frac{\tau^2}{n}}{2\sigma^2} \end{aligned} \tag{9}$$

Now, considering two different values of σ^2, σ_1^2 and σ_2^2 with $\sigma_1^2<\sigma_2^2$, $E_1[ll_o]>E_2[ll_o]$ iff

$$\begin{aligned} -\ln\left(\sqrt{2\pi\sigma_1^2}\right)-\frac{\tau^2(n-1)}{2\sigma_1^2 n} &> -\ln\left(\sqrt{2\pi\sigma_2^2}\right)-\frac{\tau^2(n-1)}{2\sigma_2^2 n} \\ \tau^2 &< \frac{\ln\left(\sigma_2^2\right)-\ln\left(\sigma_1^2\right)}{\frac{n-1}{n}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)}, \end{aligned} \tag{10}$$

[16] It may sound odd to state the “expectation of the expected likelihood”, but this conveys the fact that the expected log-likelihood varies with the sample mean, resulting in the need for an outer expectation taken over the sampling distribution of the mean.

and $E_1[ll_e] < E_2[ll_e]$ iff

$$-\ln\left(\sqrt{2\pi\sigma_1^2}\right) - \frac{\tau^2 + \frac{\tau^2}{n}}{2\sigma_1^2} < -\ln\left(\sqrt{2\pi\sigma_2^2}\right) - \frac{\tau^2 + \frac{\tau^2}{n}}{2\sigma_2^2}$$

$$\frac{\ln\left(\sigma_2^2\right) - \ln\left(\sigma_1^2\right)}{\left[1 + \frac{1}{n}\right]\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)} < \tau^2. \qquad (11)$$

Combining these two conditions gives the interval from Eq. 3.

References

Abbe, O.G., Goodliffe, J., Herrnson, P.S., Patterson, K.D.: Agenda setting in Congressional elections: the impact of issues and campaigns on voting behavior. Political Res. Q. **56**(4), 419–430 (2003)

Achen, C.H.: Let's put garbage-can regressions and garbage-can probits where they belong. Confl. Manag. Peace Sci. **22**(4), 327–339 (2005)

Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)

Amaral, M.A., Dunsmore, I.R.: Optimal estimates of predictive distributions. Biometrika **67**(3), 685–689 (1980)

Bailey, M.A.: Comparable preference estimates across time and institutions for the court, Congress, and presidency. Am. J. Political Sci. **51**(3), 433–448 (2007)

Boockmann, B.: Partisan politics and treaty ratification: the acceptance of international labour organisation conventions by industrialised democracies, 1960–1996. Eur. J. Political Res. **45**(1), 153–180 (2006)

Chaffin, W.W., Rhiel, S.G.: The effect of skewness and kurtosis on the one-sample *t* test and the impact of knowledge of the population standard deviation. J. Stat. Comput. Simul. **46**(1), 79–90 (1993)

Clarke, K.A.: Testing nonnested models of international relations: reevaluating realism. Am. J. Political Sci. **45**(3), 724–744 (2001)

Clarke, K.A.: Nonparametric model discrimination in international relations. J. Confl. Resolut. **47**(1), 72–93 (2003)

Clarke, K.A.: A simple distribution-free test for nonnested hypotheses. Political Anal. **15**(3), 347–363 (2007)

Diebold, F.X., Mariano, R.S.: Comparing predictive accuracy. J. Bus. Econ. Stat. **20**(1), 134–144 (2002)

Gilula, Z., Haberman, S.J.: Density approximation by summary statistics: an information-theoretic approach. Scand. J. Stat. **27**(3), 521–534 (2000)

Greene, W.H.: Econometric Analysis, 6th edn. Prentice Hall, Upper Saddle River (2008)

Hall, P.: On Kullback–Leibler loss and density estimation. Ann. Stat. **15**(4), 1491–1519 (1987)

Johnson, N.J.: Modified *t* tests and confidence intervals for asymmetrical populations. J. Am. Stat. Assoc. **73**(363), 536–544 (1978)

Joshi, M., Mason, T.D.: Between democracy and revolution: peasant support for insurgency versus democracy in Nepal. J. Peace Res. **45**(6), 765–782 (2008)

Koenker, R.: Quantile Regression. Cambridge University Press, New York (2005)

Konishi, S., Kitagawa, G.: Generalised information criteria in model selection. Biometrika **83**(4), 875–890 (1996)

Konisky, D.M., Woods, N.D.: Exporting air pollution? Regulatory enforcement and environmental free riding in the United States. Political Res. Q. **63**(4), 771–782 (2010)

Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)

Lange, K.L., Little, R.J.A., Taylor, J.M.G.: Robust statistical modeling using the *t* distribution. J. Am. Stat. Assoc. **84**(408), 881–896 (1989)

Mebane, W.R., Sekhon, J.S.: Coordination and policy moderation at midterm. Am. Political Sci. Rev. **96**(1), 141–157 (2002)

Mondak, J.J., Sanders, M.S.: The complexity of tolerance and intolerance judgments: a response to Gibson. Political Behav. **27**(4), 325–337 (2005)

Palazzolo, D.J., Moscardelli, V.G.: Policy crisis and political leadership: election law reform in the states after the 2000 presidential election. State Politics Policy Q. **6**(3), 300–321 (2006)

Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(379–423), 623–656 (1948)

Shellman, S.M., Stewart, B.M.: Political persecution or economic deprivation? A time-series analysis of Haitian exodus, 1990–2004. Confl. Manag. Peace Sci. **24**(2), 121–137 (2007)

Smyth, P.: Model selection for probabilistic clustering using cross-validated likelihood. Stat. Comput. **10**(1), 63–72 (2000)

Souva, M.: Foreign policy determinants: comparing realist and domestic-political models of foreign policy. Confl. Manag. Peace Sci. **22**(2), 149–163 (2005)

Travis, R.: Problems, politics, and policy streams: a reconsideration us foreign aid behavior toward Africa. Int. Stud. Q. **54**(3), 797–821 (2010)

Vuong, Q.H.: Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica **57**(2), 307–333 (1989)

Ward, M.D., Greenhill, B.D., Bakke, K.M.: The perils of policy by *p*-value: predicting civil conflicts. J. Peace Res. **47**(4), 363–375 (2010)

Western, B.: Concepts and suggestions for robust regression analysis. Am. J. Political Sci. **39**(3), 786–817 (1995)