Web Appendix to "Linear Models with Outliers: Choosing Between Conditional-Mean and Conditional-Median Methods"

Jeffrey J. Harden* Bruce A. Desmarais†

Contents

1 Implementing MR and CVDM in R and Stata **ii**

1.1 R ii

1.2 Stata iii

2 MR versus OLS Efficiency **iii**

3 Details of the CVDM Test **v**

3.1 Overview and Null Hypothesis v

3.2 Unbiasedness of the CVDM Test vi

3.3 Computing the CVDM Test Statistic vii

4 Additional Replication Examples **viii**

*Ph.D Candidate, Department of Political Science, University of North Carolina at Chapel Hill, 312 Hamilton Hall CB #3265, Chapel Hill, NC 27599, 919.962.0430, jjharden@unc.edu.

†Assistant Professor, Department of Political Science, University of Massachusetts–Amherst, 420 Thompson Hall, 200 Hicks Way, Amherst, MA 01003, 413.545.1992, desmarais@polsci.umass.edu.

1 Implementing MR and CVDM in R and Stata

Here we provide a brief overview for implementing MR and CVDM in R and MR in Stata. For additional resources, see Koenker (2005) or http://www.econ.uiuc.edu/~roger/research/rq/rq.html.

1.1 R

The following code generates data with a Laplace error distribution, estimates MR, and performs the CVDM test.

```
# Load necessary packages
library(quantreg)
library(MASS)
library(normalp)
source("CVDM.R")

set.seed(12345)

# Define data
n <- 1000
x1 <- runif(n, -1, 1)
x2 <- runif(n, -1, 1)
x3 <- runif(n, -1, 1)
b1 <- .1
b2 <- .2
b3 <- .3
y <- b1*x1 + b2*x2 + b3*x3 + rnormp(n, p = 1)

# Model estimation
mr.model <- rq(y ~ x1 + x2 + x3)
summary(mr.model, se = "iid")

# CVDM test
cvdm.test <- CVDM(y ~ x1 + x2 + x3, data = data.frame(y, x1, x2, x3))
cat("CVDM test statistic:", cvdm.test$cvjt, "\n")
```

```
cat("A negative test statistic favors MR; positive favors OLS", "\n")
```

1.2 Stata

The following code generates data with a normal error distribution and estimates MR.

```
set seed 12345


* Define data
set obs 1000
generate x1 = uniform()
generate x2 = uniform()
generate x3 = uniform()
generate y = .1*x1 + .2*x2 + .3*x3 + invnorm(uniform())

* Model estimation
qreg y x1 x2 x3
```

2 MR versus OLS Efficiency

A key element of the result that MR can be more efficient than OLS even when OLS is BLUE is the important distinction between BLUE ("best linear unbiased estimator") and MVUE ("minimum variance unbiased estimator"). We explain this and review the concepts of efficiency, BLUE, and MVUE below.

The "best" in BLUE refers to the fact that, *among linear estimators*, OLS has minimum variance. Generally, the efficiency of an estimator is judged by the inverse of its sampling variation, which is the variance of the estimator over an infinite number of samples of the same size drawn from the population that generated the data. The higher the variance of the estimator, the lower the efficiency. For instance, it is well known that the variance of the sample mean, computed on a sample of size N drawn from a population with variance σ^2 is $\frac{\sigma^2}{n}$, regardless of the particular distribution that characterizes the population (e.g., normal, binomial, or Poisson). Thus, the efficiency of the sample mean is $\frac{n}{\sigma^2}$. Political scientists often cite the BLUE property of OLS as justification for its exclusive use as an estimator of the linear model when the Gauss-Markov assumptions are

plausible. Take, for example, the following excerpt from Beck and Katz (1995, 636):

> Ordinary least squares is optimal (best linear unbiased) for TSCS models if the errors are assumed to be generated in an uncomplicated ('spherical') manner. In particular, for OLS to be optimal it is necessary to assume that all the error processes have the same variance (homoscedasticity) and that all of the error processes are independent of each other.

This perspective on the BLUE property, that it is synonymous with optimality, seems to emphasize the "best" and ignore the "linearity" qualification. It is critical to separate the concepts of the linearity of the estimator and the linear-additive relationship between the independent variables and the response. The linearity in BLUE refers to a property of the algorithm used to estimate the parameters, and is not a reference to the linear relationships in the model. An estimator $\hat{\beta}$ of a parameter β is linear if it can be written as

$$\hat{\beta} = \sum_{i=1}^{n} a_i h(y_i, \boldsymbol{x}_i), \tag{1}$$

where $h(\cdot)$ is any (possibly non-linear) function, $\{y, \boldsymbol{x}_i\}$ is a single data point, and a is a weight that does not depend on $\hat{\beta}$ (Casella and Berger 2002). OLS is a linear estimator because the regression coefficient estimate can be represented as a weighted sum of functions of the data points, not due to the linear relationship between the covariates and the conditional mean of the response.

The fact that the model of interest is composed of linear relationships does not imply that the estimator used is a linear estimator. Indeed, though it estimates linear relationships between the median of the dependent variable and the independent variables, MR is a non-linear estimator since the MR coefficient estimates cannot be represented in the form of Equation 1. Also, the fact that a linear estimator is used does not imply that the relationships between independent variables and the dependent variable are linear. It is well known that logistic regression, which is appropriate when the outcome variable is binary, characterizes a non-linear relationship between the dependent variable and an independent variable. However, linear estimators have been derived for large-scale

or otherwise computationally demanding applications of logistic regression (e.g., Veerkamp 2000; van der Laan, Hubbard, and Jewell 2007).

3 Details of the CVDM Test

Here we provide a more detailed justification for the CVDM test.

3.1 Overview and Null Hypothesis

Recall that a likelihood function can be used to compute the likelihood of having observed the data on hand, given that the data were drawn from the model corresponding to that likelihood function and vector of parameters. Moreover, when the observations are assumed to be independent, the likelihood function can be used to compute the likelihood of observing any particular data point (an individual likelihood). One approach to selecting between two models is to consider the model corresponding to the higher average value of the individual likelihood (or log-likelihood) function to be closer to the data generating process, since it is more likely to have generated the data. Indeed, this is more than heuristically reasonable. The average individual log-likelihood value is a consistent estimator of perhaps the most thoroughly-studied and commonly-used measure of distance between the true model f, and an approximating model g: the Kullback-Leibler Divergence, denoted $D_{\text{KL}}(f||g)$ (Kullback and Leibler 1951).[1] Suppose there are two equally theoretically satisfying approximating models under consideration, g_1 and g_2 (e.g., OLS and MR). To decide whether one model should be favored over the other, it would be ideal to test the null hypothesis that $D_{\text{KL}}(f||g_1) = D_{\text{KL}}(f||g_2)$: both models are equally close to the true data generating process.

Testing this null hypothesis that $D_{\text{KL}}(f||g_1) = D_{\text{KL}}(f||g_2)$ serves as the motivation for the tests derived by Vuong (1989) and Clarke (2003, 2007), both of which are used in political science.[2] Let $\boldsymbol{\delta}$ be the vector of N log-likelihood differences computed as $\delta_i = L_{\text{OLS}}(y_i|\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}) - L_{\text{MR}}(y_i|\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are the OLS and MR estimates of the regression coefficients using all N observa-

[1] For instance, the Akaike Information Criterion (Akaike 1974) is derived with the sole purpose of providing an asymptotically unbiased estimator of the Kullback-Leibler Divergence.

[2] See Lubell, Schneider, Scholz, and Mete (2002); Mondak and Sanders (2005), and Bailey (2007) for recent examples that utilize the Vuong test and Souva (2005) and Boockmann (2006) for those that use the Clarke test.

tions, respectively, and $L_{\text{OLS}}(\cdot)$ and $L_{\text{MR}}(\cdot)$ are the OLS and MR likelihood functions, respectively. The Vuong test constitutes a z-test of the null hypothesis that $E[\boldsymbol{\delta}] = 0$, and the null hypothesis of the Clarke test is that MEDIAN$(\boldsymbol{\delta}) = 0$. As a large-sample justification for his approach, Vuong (1989) notes that $\bar{\boldsymbol{\delta}}$ is a consistent estimator of $D_{\text{KL}}(f||g_1) - D_{\text{KL}}(f||g_2)$. Clarke (2007) criticizes the Vuong test in that the assumption that the distribution of $\boldsymbol{\delta}$ is normal, justifying the z-test, may not be appropriate in finite samples. Clarke (2007) suggests using the sign test (the count of the number of times in N that $\boldsymbol{\delta} > 0$) to test the median hypothesis because, regardless of the distribution of $\boldsymbol{\delta}$, the null distribution of the sign test is known to be binomial with N trials and a probability of success in any one trial equal to 0.5.

3.2 Unbiasedness of the CVDM Test

One weakness of the Vuong and Clarke tests is that the difference in the average individual log-likelihoods is a *biased* (albeit consistent) estimator of $D_{\text{KL}}(f||g_1) - D_{\text{KL}}(f||g_2)$ (Akaike 1974; Smyth 2000; Desmarais and Harden N.d.). This means that, in a finite sample, neither the Vuong nor the Clarke procedures effectively test the null hypothesis that $D_{\text{KL}}(f||g_1) = D_{\text{KL}}(f||g_2)$.[3] On these properties of the Clarke and Vuong tests, Desmarais and Harden (N.d.) show that (1) the finite sample bias in the mean of $\boldsymbol{\delta}$ as an estimator of $D_{\text{KL}}(f||g_1) - D_{\text{KL}}(f||g_2)$ can be so severe that the Vuong test chooses the poorer-fitting model more often than the better-fitting model and (2) the distribution of $\boldsymbol{\delta}$ can be skewed such that the median and mean of $\boldsymbol{\delta}$ are differently signed, meaning the Clarke procedure tests a much different null hypothesis than $D_{\text{KL}}(f||g_1) = D_{\text{KL}}(f||g_2)$. The overall problem with the biases in these tests is that there is no particular reason to test whether the mean or median of $\boldsymbol{\delta}$ is zero in a finite sample. Vuong (1989) is careful to justify his test as a large-N procedure because, *in the limit*, the effective null hypothesis is that $D_{\text{KL}}(f||g_1) = D_{\text{KL}}(f||g_2)$. There is simply no reason to base a model selection rule on a test involving $\boldsymbol{\delta}$, when the mean and median of $\boldsymbol{\delta}$ are both finite-sample-biased estimators of $D_{\text{KL}}(f||g_1) - D_{\text{KL}}(f||g_2)$.

The bias inherent in using $\boldsymbol{\delta}$ to estimate $D_{\text{KL}}(f||g_1) - D_{\text{KL}}(f||g_2)$ arises because the exact same

[3] Moreover, since the Clarke test considers the median of $\boldsymbol{\delta}$ rather than its mean, even in large samples, the effective null hypothesis of the Clarke test may not be that $D_{\text{KL}}(f||g_1) = D_{\text{KL}}(f||g_2)$.

data is used to (1) fit the models' parameters and (2) evaluate the fit of the models (Konishi and Kitagawa 1996). The process of measuring the out-of-sample fit of a model consists of using data that is not used to estimate the model, but is thought to be drawn from the same population as the data used to fit the model to evaluate the performance of the model. Political scientists have recently taken note of the fact that out-of-sample evaluation of model fit serves as an automatic guard against over-fitting a statistical model (e.g., Ward, Greenhill, and Bakke 2010). The CVLL, which is an out-of-sample measure of model fit, has been proven to provide an unbiased estimate of $D_{\text{KL}}(f||g_1) - D_{\text{KL}}(f||g_2)$ (Smyth 2000). Thus, using the cross-validated individual log-likelihoods in a difference of means test results in a finite-sample, unbiased test of the null hypothesis that $D_{\text{KL}}(f||g_1) = D_{\text{KL}}(f||g_2)$.

3.3 Computing the CVDM Test Statistic

Desmarais and Harden (N.d.) use the CVLL to construct the CVDM test, which, as we show here, can be applied to selecting between OLS and MR. Recall from the main text (Equation 1), that $\boldsymbol{\delta}^{(cv)}$ is an N-length vector of differences in individual CVLLs. Then, in the context of comparing OLS and MR, Equation 1 states that

$$\delta_i^{(cv)} = L_{\text{OLS}}(y_i|\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_{-i}) - L_{\text{MR}}(y_i|\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}}_{-i}), \tag{2}$$

where the $-i$ subscripts on $\hat{\boldsymbol{\beta}}_{-i}$ and $\tilde{\boldsymbol{\beta}}_{-i}$ indicate that the i^{th} observation is excluded from the sample used to estimate the regression coefficients. The CVDM test is a test of the null hypothesis that $E\left[\boldsymbol{\delta}^{(cv)}\right] = 0$, which Desmarais and Harden (N.d.) show is, at any sample size, equivalent to testing the null hypothesis that $D_{\text{KL}}(f||g_1) = D_{\text{KL}}(f||g_2)$. The starting point for the CVDM test is a conventional t-test applied to $\boldsymbol{\delta}^{(cv)}$. Because it is possible that there is significant skew in the distribution of $\boldsymbol{\delta}^{(cv)}$, the t-statistic is adjusted for skewness using the procedure suggested by Johnson (1978). Specifically, let $\bar{\delta}^{(cv)}$ be the sample mean of $\boldsymbol{\delta}^{(cv)}$. Then the unbiased estimator of the skewness of $\boldsymbol{\delta}^{(cv)}$ is $\hat{\mu}^3 = n(n-1)^{-1}(n-2)^{-1}\sum_{i=1}^{n}(\delta_i^{(cv)} - \bar{\delta}^{(cv)})^3$. The CVDM test statistic

is

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

where s is the conventional estimator of the standard deviation of $\boldsymbol{\delta}^{(cv)}$. To derive p-values, the CVDM statistic is evaluated with respect to a Student's t distribution with $n-1$ degrees of freedom. Desmarais and Harden (N.d.) show, both analytically and in simulations, that the CVDM test is unbiased, and thus performs better than either the Vuong or Clarke tests.

4 Additional Replication Examples

Though MR is a useful tool for state politics scholars, its applicability extends across political science. We conducted 15 different replications of OLS models from several subfields with MR and our CVDM test. Table A.1 summarizes the results. It shows that the test chooses both OLS and MR across the different examples and that MR can lead to several outcomes with respect to either statistical or substantive significance. Our definition of "less support" is a change from a significant p-value (at the 0.05 level) to a non-significant level on at least one coefficient that is part of a hypothesis. Our definition of "more support" is a change from non-significant to significant on at least one coefficient that is part of a hypothesis. Our definition of "mixed results" is at least one coefficient satisfies the "less support" condition while at least one satisfies the "more support" condition.

[Insert Table A.1 here]

References

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transaction on Automatic Control* 19(6):716–723.

Apollonio, D. E. and Raymond J. La Raja. 2004. "Who Gave Soft Money? The Effect of Interest Group Resources on Political Contributions." *Journal of Politics* 66(4):1134–1154.

Bailey, Michael A. 2007. "Comparable Preference Estimates across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51(3):433–448.

Beck, Nathaniel and Jonathan N. Katz. 1995. "What to Do (And Not to Do) With Time-Series Cross-Section Data." *American Political Science Review* 89(3):634–647.

Boockmann, Bernhard. 2006. "Partisan Politics and Treaty Ratification: The Acceptance of International Labour Organisation Conventions by Industrialised Democracies, 1960–1996." *European Journal of Political Research* 45(1):153–180.

Braithwaite, Alex. 2006. "The Geographic Spread of Militarized Disputes." *Journal of Peace Research* 43(5):507–522.

Brown, Robert D., Robert A. Jackson and Gerald C. Wright. 1999. "Registration, Turnout, and State Party Systems." *Political Research Quarterly* 52(3):463–479.

Casella, George and Roger L. Berger. 2002. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury.

Clarke, Kevin A. 2003. "Nonparametric Model Discrimination in International Relations." *Journal of Conflict Resolution* 47(1):72–93.

Clarke, Kevin A. 2007. "A Simple Distribution-Free Test for Nonnested Hypotheses." *Political Analysis* 15(3):347–363.

Czaika, Mathias and Krisztina Kis-Katos. 2009. "Civil Conflict and Displacement: Village-Level Determinants of Forced Migration in Aceh." *Journal of Peace Research* 46(3):399–418.

Desmarais, Bruce A. and Jeffrey J. Harden. N.d. "An Unbiased Model Comparison Test Using Cross-Validation." Available at SSRN: http://ssrn.com/abstract=1765845.

Gerber, Elisabeth R. and Jeffrey B. Lewis. 2004. "Beyond the Median: Voter Preferences, District Heterogeneity, and Political Representation." *Journal of Political Economy* 112(6):1364–1383.

Golder, Matt. 2006. "Presidential Coattails and Legislative Fragmentation." *American Journal of Political Science* 50(1):34–48.

Griffin, John D. 2006. "Electoral Competition and Democratic Responsiveness: A Defense of the Marginality Hypothesis." *Journal of Politics* 68(4):909–919.

Heger, Lindsay and Idean Salehyan. 2007. "Ruthless Rulers: Coalition Size and the Severity of Civil Conflict." *International Studies Quarterly* 51(2):385–403.

Huber, John D., Charles R. Shipan and Madelaine Pfahler. 2001. "Legislatures and Statutory Control of Bureaucracy." *American Journal of Political Science* 45(2):330–345.

Johnson, Norman J. 1978. "Modified t Tests and Confidence Intervals for Asymmetrical Populations." *Journal of the American Statistical Association* 73(363):536–544.

Johnson, Timothy, Paul J. Wahlbeck and James Spriggs. 2006. "The Influence of Oral Arguments on the U.S. Supreme Court." *American Political Science Review* 100(1):99–113.

Joshi, Madhav and T. David Mason. 2008. "Between Democracy and Revolution: Peasant Support for Insurgency versus Democracy in Nepal." *Journal of Peace Research* 45(6):765–782.

Koenker, Roger. 2005. *Quantile Regression*. New York: Cambridge University Press.

Konishi, Sadanori and Genshiro Kitagawa. 1996. "Generalised Information Criteria in Model

Selection." *Biometrika* 83(4):875–890.

Kullback, Solomon and Richard A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22(1):79–86.

Li, Quan and Rafael Reuveny. 2006. "Democracy and Environmental Degradation." *International Studies Quarterly* 50(4):935–956.

Lubell, Mark, Mark Schneider, John T. Scholz and Mihriye Mete. 2002. "Watershed Partnerships and the Emergence of Collective Action Institutions." *American Journal of Political Science* 46(1):148–163.

Mondak, Jeffery J. and Mitchell S. Sanders. 2005. "The Complexity of Tolerance and Intolerance Judgments: A Response to Gibson." *Political Behavior* 27(4):325–337.

Smyth, Padhraic. 2000. "Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood." *Statistics and Computing* 10(1):63–72.

Souva, Mark. 2005. "Foreign Policy Determinants: Comparing Realist and Domestic-Political Models of Foreign Policy." *Conflict Management and Peace Science* 22(2):149–163.

van der Laan, Mark J., Alan Hubbard and Nicholas P. Jewell. 2007. "Estimation of Treatment Effects in Randomized Trials with Non-Compliance and a Dichotomous Outcome." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 69(3):463–482.

Veerkamp, Wim J. J. 2000. "Taylor Approximations to Logistic IRT Models and Their Use in Adaptive Testing." *Journal of Educational and Behavioral Statistics* 25(3):307–343.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57(2):307–333.

Ward, Michael D., Brian D. Greenhill and Kristin M. Bakke. 2010. "The Perils of Policy by p-value: Predicting Civil Conflicts." *Journal of Peace Research* 47(4):363–375.

Table A.1: Summary of Replication Results

Article	Dependent Variable	N	CVDM Test	Consequences
Huber, Shipan, and Pfahler (2001, *AJPS*)	Words added to Medicaid legislation	38	MR*	Less support
Paul and Brown (2006, *SPPQ*)	Voter support for referendums	41	MR*	Less support
Apollonio and La Raja (2004, *JOP*)	Corporate soft money contributions	47	MR*	More support
Gerber and Lewis (2004, *JPE*)	NOMINATE scores	55	OLS*	Original selected
Joshi and Mason (2008, *JPR*)	Voter turnout %	75	MR^{ns}	Mixed results
Brown, Jackson, and Wright (1999, *PRQ*)	Voter registration %	180	OLS^{ns}	Original selected
Li and Reuveny (2006, *ISQ*)	Deforestation rate	204	MR*	Less support
Heger and Salehyan (2007, *ISQ*)	Battle deaths (logged)	213	OLS*	Original selected
Li and Reuveny (2006, *ISQ*)	Forested area share of land area	255	MR*	Less support
Braithwaite (2006, *JPR*)	Battle land area	296	MR*	Less support
Golder (2006, *AJPS*)	Effective no. of electoral parties	603	MR*	No change
Johnson, Wahlbeck, and Spriggs (2006, *APSR*)	U.S. Supreme Court oral argument grades	1,118	OLS*	Original selected
Hogan (2008, *AJPS*)	Challenger/incumbent spending ratio	1,816	MR*	Less support
Czaika and Kis-Katos (2009, *JPR*)	Population change %	5,211	MR*	Less support
Griffin (2006, *JOP*)	NOMINATE scores	5,685	MR^{ns}	No change

Note: Cell entries report replication summary information, including dependent variable, sample size, results of the CVDM test, and, when applicable, the consequences of MR for inferences on the authors' hypotheses. * $p < 0.05$; ^{ns} *Not significant.*