When BLUE Is Not Best

Non-Normal Errors and the Linear Model*

Daniel K. Baissa†

Carlisle Rainey‡

Abstract

Researchers in political science often estimate linear models of continuous outcomes using least squares. While it is well-known that least-squares estimates are sensitive to single, unusual data points, this knowledge has not led to careful practices when using least-squares estimators. We highlight the importance of using more robust estimators along with variable transformations and discuss several approaches to detect, summarize, and communicate the influence of particular data points. We conclude with a reanalysis of Clark and Golder (2006) and show that the residuals are highly non-normal under their model specification. We show that applying an empirically-chosen transformation and using a robust estimator allows us to improve and extend their analysis in important ways.

*We thank Bill Clark and Matt Golder for making their data available to us. The analyses presented here were conducted with R 3.2.2. All data and computer code necessary for replication are available at github.com/carlislerainey/heavy-tails.

†Daniel K. Baissa is a Ph.D. student in the Department of Government, Harvard University, 1737 Cambridge St., Cambridge , MA, 02138 (dbaissa@g.harvard.edu).

‡Carlisle Rainey is Assistant Professor of Political Science, Texas A&M University, 2010 Allen Building, College Station, TX, 77843 (crainey@tamu.edu).

Introduction

Linear models of the form $y_i = X_i\beta + \epsilon_i$ estimated with least squares remain one of the most common statistical tools in political science research (Krueger and Lewis-Beck 2008). Yet there seems to remain some confusion about the conditions under which least squares serves as a good estimator of the linear model. After assuming that the model $y_i = X_i\beta + \epsilon_i$ is correct and that the matrix X is full rank, we need to make further assumptions about the errors ϵ_i to obtain desirable properties for the least-squares estimator. We might use some of the following assumptions:

A1: Errors have mean equal to zero.

A2: Errors have a constant, finite variance.

A3: Errors are independent.

A4: Errors follow a normal distribution.

By assuming only A1, we obtain an unbiased, consistent estimator (e.g., Wooldridge 2013, p. 810 and pp. 385, 815-816). However, by assuming A1, A2, A3, and A4, we obtain the best unbiased estimator (BUE), which has the smallest possible variance among the class of unbiased estimators (e.g., Wooldridge 2013, pp. 809-815).

However, political scientists tend to focus on a different property of least squares. By the Gauss-Markov theorem, we can remove A4—the assumption of normal errors—and still obtain the best *linear* unbiased estimator (BLUE), which has the smallest possible variance among the class of unbiased, *linear* estimators (e.g., Wooldridge 2013, pp. 809-812). Researchers have primarily justified least squares using the Gauss-Markov theorem because it seems to impart desirable small-sample properties without the overly restrictive assumption of normal errors. For example, Berry and Feldman (1985) write:

> [The assumption of normally distributed errors] is necessary *only* for tests of significance; its violation will have no effect on the estimation of the parameters of the

regression model. It is quite fortunate that normality is not required for estimation, because it is often very difficult to defend this assumption in practice.[1]

However, notice that a tradeoff occurs when relaxing the assumption of normal errors. In order to relax the assumption of normal errors (and keep desirable small sample properties), we must restrict ourselves to linear estimators. This raises a critical, but often overlooked question: Under what conditions can a researcher safely restrict herself to linear estimators? We argue that a restriction to linear estimators makes sense only when the errors follow a normal distribution. If the errors do not follow a normal distribution, then least-squares is still a best *linear* unbiased estimator, but other, non-linear estimators may be more more efficient. Our claim is that the restriction to linear estimators is artificial and can only be justified by assuming normal errors—an assumption that Berry and Feldman (1985) note is very difficult to defend in practice.

The Gauss-Markov theorem has convinced researchers in political science that as long as A1, A2, and A3—the Gauss-Markov assumptions—are met, the distribution of the errors is unimportant. But the distribution of the errors is crucial to a linear regression analysis. Deviations from normality, especially large deviations commonly found in regression models in political science, can devastate the performance of least squares compared to alternative estimators.

In this paper, we emphasize the importance of errors and residuals from a statistical and substantive perspective. We adopt and defend a skeptical perspective toward least squares in favor of more robust estimators. We proceed as follows: we (1) clarify the crucial distinction between a linear *model* and a linear *estimator*, (2) explain that the BLUE estimator is not the best estimator unless the errors are normally distributed, (3) highlight powerful, robust alternatives to least-squares estimators that are unbiased and more efficient for a wide range of substantively

[1] Similarly, Wooldridge (2013, p. 101) writes that the Gauss-Markov theorem "justifies the use of the OLS method rather than using a variety of competing estimators."

plausible error distributions, (4) provide concrete, practical advice to substantive researchers using linear models, and (5) provide a compelling example that illustrates the importance of robust estimators.

Is a BLUE Estimator the Best Estimator?

The linear model can be written as $y = X\beta + \epsilon$.[2] Researchers in political science commonly estimate this model with least squares by minimizing the sum of the squared residuals, such that $\hat{\beta}^{ls} = \arg\min_b S(b)$, where $S(b) = \sum_{i=1}^{n}(y_i - X_i b)^2$. If we assume that the errors ϵ follow independent and identical normal distributions with mean zero and unknown variance, which we refer to as a "normal-linear model," then the least squares estimator is the best (i.e., minimum variance) unbiased estimator (BUE) (Casella and Berger 2002, pp. 334-342, and Wooldridge 2013, 807-815). This is a powerful result. Under the assumption of normally distributed errors, least squares is the most efficient unbiased estimator.

If we relax the assumption of normality and simply require that the errors have mean zero and constant variance, then the Gauss-Markov Theorem guarantees that the least-squares estimator is the best *linear* unbiased estimator. While this result is often emphasized, it should provide little comfort to researchers because there is little statistical or substantive reason to restrict themselves to *linear* estimators.

At first glance, one might take the linearity restriction in the Gauss-Markov theorem to refer to the structure of the model (i.e., "linear in the parameters"). Indeed, this is the sense in which we use "linear" in the phrase "linear model." However, the "linear" restriction in the Gauss-Markov Theorem refers something else—a linear *estimator*, which is a technical condition that has little connection to the substance of the problem. Linearity of the *estimator*

[2] As usual, y is an outcome variable of interest (usually roughly continuous), X is a $n \times (k+1)$ matrix containing a single column of ones and k columns holding k explanatory variables, β is a $(k+1) \times 1$ matrix of model coefficients, and ϵ is an $n \times 1$ matrix of errors. As usual, the statistical properties of these estimators depend on this model being correct and a full rank X.

simply requires that the estimates be a linear function of the outcome variable, so that $\hat{\beta}_j = \lambda_1 y_1 + \lambda_2 y_2 + ... + \lambda_n y_n$, where the weights λ_i are allowed to depend on X, but not on y. In other words, the Gauss-Markov theorem assumes a linear *model* of the form $E(y|X) = X\beta$, but it also restricts researchers to linear *estimators* of the form $\hat{\beta}_j = \lambda_1 y_1 + \lambda_2 y_2 + ... + \lambda_n y_n$.

We can see that the least-squares criterion produces a linear estimator with some simple algebra. First, recall that we wish to minimize $S(b) = \sum_{i=1}^{n}(y_i - X_i b)^2$ with respect to b. To do this, we can simply set $\frac{\partial S(\hat{\beta}^{ls})}{\partial \hat{\beta}^{ls}} = 0$ and solve for the vector $\hat{\beta}^{ls}$. $\frac{\partial S(\hat{\beta}^{ls})}{\partial \hat{\beta}^{ls}} = \sum_{i=1}^{n} 2(y_i - X_i\hat{\beta}^{ls})(-X_i) = 0$ implies that $\sum_{i=1}^{n} X_i(y_i - X_i\hat{\beta}^{ls}) = 0$. This is simply a system of $k+1$ *linear* equations $\sum_{i=1}^{n} X_{ij}(y_i - X_i\hat{\beta}^{ls})$ for $j = \{0, 1, 2, ..., k\}$. Of course, the matrix form $X'(y - X\hat{\beta}^{ls}) = 0 \Rightarrow (X'X)\hat{\beta}^{ls} = X'y \Rightarrow \hat{\beta}^{ls} = (X'X)^{-1}X'y$ is much more common. In matrix form, linearity of the estimator requires that $\hat{\beta} = My$, where M depends on the matrix X. We can clearly see that the least squares estimator $\hat{\beta}^{ls} = (X'X)^{-1}X'y$ has the form My.

However, restricting ourselves to linear estimators is neither necessary nor productive. Note that we are not arguing against linear models (i.e., linear in the parameters), such as

$$y_i = \beta_0 + \beta_1 \log(x_i) + \epsilon_i,$$
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \text{ or}$$
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i.$$

This collection of linear models illustrates that the linear model can represent a wide range of theoretically-relevant relationships, especially when it includes explanatory variables non-linearly. But there is no statistical reason to restrict ourselves to linear *estimators* for these linear models, except for mathematical convenience and computational ease. The only substantive reason to restrict ourselves to linear estimators is if we are willing to assume that we have normally distributed errors. Under this assumption, linearity is a reasonable restriction—indeed, the linear estimator is BUE when the errors are normal. Restriction to linearity does not get us

away from the assumption of normality because BLUE estimators are BUE estimators only under the assumption of normal errors. If the errors are not normally distributed, then a researcher can easily better the BLUE estimator.

If the researcher has a substantive or empirical reason to assume a non-normal distribution for the errors, such as a slightly heavier-tailed t_{10} distribution, then the linear restriction in the Gauss-Markov theorem prohibits comparisons to the more efficient (but non-linear) MLE estimator implied by the assumed t_{10} error distribution. Similarly, the linear restriction prohibits comparisons to the least absolute deviation estimator, which the is the MLE and is more efficient than least squares when the errors follow a Laplace distribution (Harden and Desmarais 2011).

Many researchers simply assume a normal-linear model for little or no substantive or empirical reason. Even while knowing that the assumed normal-linear model is *in*correct, researchers use this model as an approximation. But if the model is only an approximation, then the desirable statistical properties are no longer guaranteed (e.g., unbiasedness, minimum variance). With this in mind, it makes more sense to use a robust estimator with the following qualitative properties in typical sample sizes:

1. When the normal-linear model is exactly correct, the estimator should be approximately unbiased with efficiency comparable to, but less than, least squares.
2. When the deviation from the normal-linear model is small, the estimator should be approximately unbiased with efficiency comparable to, or perhaps greater than, least squares.
3. When the deviation from the normal-linear model is large, the estimator should exhibit relatively little bias and be much more efficient than least squares.

The "best" model for a social scientist might not be the optimal estimator for an assumed model, but an estimator that works reasonably well for the assumed model and many substantively plausible deviations.

To see the importance of this in practice, we simulated 10,000 data sets of 50 observations

of variables x and y, where the relationship between x and y is given by $y = x + \epsilon$, where ϵ follows a t distribution with three degrees of freedom. Note that the t_3 distribution is symmetric, bell-shaped, and resembles the normal distribution, except it has heavier tails. For each of these 10,000 data sets, we used least-squares to estimate the slope of the relationship between x and y, where the true value equals one. Because we simulated these data, we know that the Gauss-Markov assumptions hold. This means that least squares is the *best* linear unbiased estimator. The left panel of Figure 1 shows the distribution of the estimated slopes using least squares.

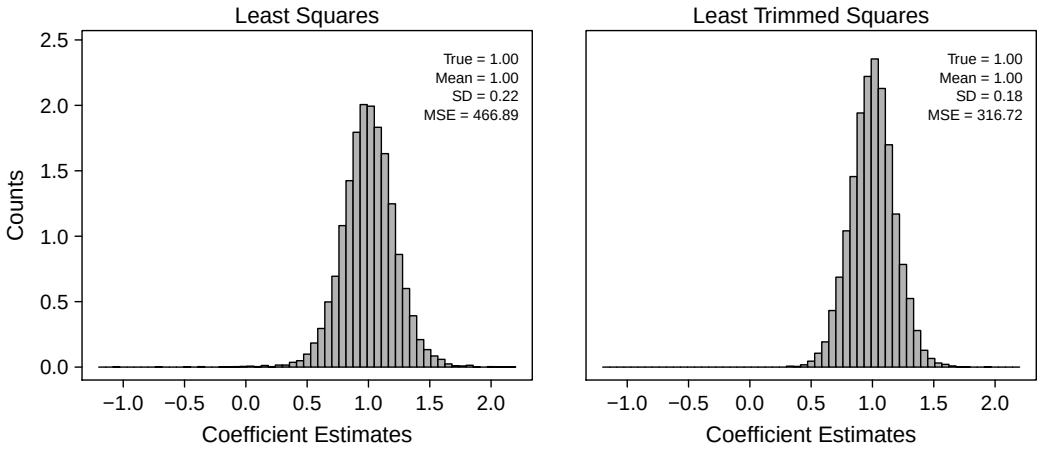

Figure 1: These two histograms show the sampling distribution for the least squares estimator and the least trimmed squares estimator under the true model $y = x + \epsilon$, where $\epsilon \sim t_3$. Notice that despite least squares being the best linear unbiased estimator for the problem, the least trimmed squares estimator is a better estimator.

But we also consider a least trimmed squares (LTS) estimator in which we minimize the smallest 90% of the residuals. This method literally throws away data. Though it lacks the elegant theory of the least squares estimator, the right panel of Figure 1 shows that it is unbiased and more efficient that the least-squares estimator. The standard deviation of the estimates from the LTS estimator is about about 18% smaller than the BLUE estimator, and the mean squared error is about 32% smaller. By any reasonable standard, the LTS estimator is a better estimator than the least squares estimator in this example, yet the least squares estimator is

BLUE. We obtain this improvement by expanding our focus to non-linear estimators, such as the LTS estimator. In this case, the LTS estimator is non-linear because it places zero weight on the largest 10% of the residuals and weights of one on the smallest 90% of the residuals.

The Relative Emphasis on Standard Errors

There has been a great deal of attention in the methodological literature to the sensitivity of standard errors to violations from the assumed model—and substantive scholars have paid attention. White's (1980) seminal paper developing heteroskedasticity-consistent standard errors has received over 20,000 citations, making it one of the most cited papers in economics. Beck and Katz's (1995) development panel corrected standard errors has received over 4,300 citations, making it one of the most cited papers in political science.

On the other hand, there has been scant attention paid by substantive political scientists to the sensitivity of the *estimates* to similar violations. This is particularly problematic, since it makes little sense to find a good standard error for a poor estimate (Freedman 2006 and King and Roberts 2014). Two papers in political science have addressed the issue of robust estimation. Western (1995) introduces political scientists to robust estimators, but this work has been essentially ignored. Although it is more broadly applicable than Beck and Katz (1995) and was published in the same year, it has received only 99 citations, or about 2% of the citations that Beck and Katz have received. Similarly, Harden and Desmarais (2011) have received only one citation, and it comes from the authors themselves. Anderson's (2008) broad and accessible introduction to robust estimation methods has received only about 150 citations, most from outside political science.

The relative focus on obtaining reasonable standard errors at the expense of reasonable estimates can be seen in Gujarati (2004). Though the text deals with robust standard errors in some detail, Gujarati (2004, p. 339) writes in a footnote:

> In passing, note that the effects of departure from normality and related topics are often discussed under the topic of robust estimation in the literature, a topic *beyond the scope of this book* [italics ours].

Angrist and Pischke (2009) devote an entire chapter to robust standard errors and completely ignore robust estimation of model coefficients. Wooldridge (2013) does devote about two pages to robust estimation, though the tone is skeptical.

Dealing with Skewness: Transforming the Outcome

Despite the lack of attention devoted by substantive scholars to non-normal errors, there are two ways in which the errors can deviate from normality, and both negatively affect inferences when using least squares.

1. The error distribution might be skewed.
2. The error distribution might have heavy tails.

We suggest dealing with these two deviations differently, so we discuss each separately.

Skewed error distributions create two problems for the linear model. First, least squares estimates the quantity $E(y|X)$ and the mean is not a good summary of location for skewed variables. Symmetric error distributions are easier to understand.

Second, and perhaps most importantly, skewed residuals from a least-squares fit indicate model misspecification. While we cannot be certain of the correct model in this situation, we can be confidence that the normal-linear model did not produce such data. In some cases, it is theoretically intuitive that explanatory variables have increasing effects on non-negative outcome variables, such as an individual's annual income. Rather than a college degree increasing one's expected income by \$10,000, perhaps a college degree increases it by 10%. If this intuition is correct and a researcher relies on the statistical model $\text{Income}_i = \beta_0 + \beta_1 \text{College Degree}_i + \epsilon_i$, then the errors will have a strong skew to the right. Simply logging the outcome, or using the

model $\log(\text{Income}_i) = \beta_0 + \beta_1 \text{College Degree}_i + \epsilon_i$, better captures the theoretical intuition.

Even if we remain indifferent toward the theoretical implications of skewed error distributions, we must remain cautious about the statistical implications. Indeed, the log-transformation in the example above improves the efficiency of the least squares estimator by making the assumption of normally-distributed errors more appropriate (not to mention the linearity of the model). The performance of least squares estimators improves as the error distribution approaches a normal distribution.

It is quite common in disciplines such as economics, for example, to log-transform non-negative outcome variables by default. Since non-negative (or strictly positive) outcomes are bounded below by zero, then these variables are likely skewed to the right—they are squeezed from the left by zero. In this case, the model $\log(y) = X\beta + \epsilon$ will likely provide a better approximation to the data.

When handling skewed residuals, it is impossible to know whether (1) the skew is due to a misspecification of the outcome variable (i.e., failing to transform) or (2) the errors simply follow a heteroskedastic, skewed distribution. However, we can be confident that heavily skewed residuals are inconsistent with normal errors—the researcher must do something to address the skew. Transforming the outcome variable is *one* effective method for making the model more consistent with the data.

The Box-Cox Transformation

While we agree with the spirit of the suggestion to log-transform a non-negative outcome variable y, statisticians have created more precise empirical methods for choosing *whether* and *how* to do the transformation. Box and Cox (1964) propose the Box-Cox transformation

$$y^{(\lambda)} = BC(y, \lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log y & \text{for } \lambda = 0 \end{cases},$$

where the transformation parameter λ is estimated with maximum likelihood. In this case, the model becomes $y^{(\lambda)} = X\beta + \epsilon$. This is particularly convenient because $\hat{\lambda} \approx 1$ suggests no transformation is needed and $\hat{\lambda} \approx 0$ suggests that only an intuitive log-transformation is needed.

Researchers can easily assess the skewness in the residuals using a simple histogram of the residuals or a QQ plot of the residuals compared to their normal quantiles. For a formal test of skewness, researchers might use a direct test for symmetry on residuals $\hat{e}$, such as the Mira test (Mira 1999) or simply test whether $\lambda \neq 1$ under the Box-Cox framework. However, we do not want to argue for a particular test, but to highlight that (1) asymmetries worsen the performance of least squares and many robust methods, (2) researchers can easily detect asymmetries by carefully examining the residuals, and (3) researchers can address this problem with simple, easy-to-use transformations.

Mean or Median?

Applying a non-linear transformation to the outcome variable y does raise an interpretational difficulty. The usual, untransformed linear model is given by $y = X\beta + \epsilon$ and the quantity of interest is usually $E(y|X)$ or $\frac{\partial E(y|X)}{\partial x_j}$. For concreteness, consider the log-transformation. Using the same logic, then the model is $log(y) = X\beta + \epsilon$ and we might take the quantity of interest to be $E[\log(y)|X]$ or $\frac{\partial E[\log(y)|X]}{\partial x_j}$. However, the substantive researcher is usually interested in y, not $\log(y)$, making $\frac{\partial E[\log(y)|X]}{\partial x_j}$ more difficult to understand than $\frac{\partial E(y|X)}{\partial x_j}$. To make the results more interpretable, we simply need to "undo" the transformation. But $E[\log(y)|X] \neq \log[E(y|X)]$, which means that the log cannot simply be undone without additional computation.

These interpretational difficulties are not due to the choice to transform the data, but imbedded in the data themselves. In the context of skewed distributions, the mean $E(\cdot)$ is not a good summary of the "center" of the data. While the mean often makes calculations easier, the median offers a better summary of location. The median also has an intuitive interpretation because one-half of the distribution lies above the median and one-half lies below. If a researcher

uses med($y_{new}|X_{new}$) to predict the unknown outcome y_{new} for a known case X_{new}, then she has a 50% chance of being too high and a 50% chance of being too low.

In addition to the intuitive substantive interpretation of med($y|X$), the median has another desirable property. Because the log-transformation is order-preserving, med[log(y)$|X$] = log[med($y|X$)], which means that the log *can* easily be undone because $e^{\text{med}[\log(y)|X]} = e^{\log[\text{med}(y|X)]} = \text{med}(y|X)$. Therefore, by adopting med($y|X$) and $\frac{\partial \text{med}(y|X)}{\partial x_j}$ as the quantities of interest, the researcher eases the interpretation of the results and can easily move between transformed and untransformed outcomes (e.g., med[log(y)] $\rightarrow$ med(y)). This holds for the more general case of $y^{(\lambda)}$ as well.

Simulating Quantities of Interest Under Transformation

To obtain quantities of interest relating to med(y) when the estimated model has the generic form $y^{(\lambda)} = X\beta + \epsilon$, one can simply use the algorithm described in King, Tomz, and Wittenberg (2000).

1. Estimate the Box-Cox transformation parameter $\hat{\lambda}$ using maximum likelihood. (If the values one or zero fall within the confidence interval, then one may wish to use those values to maintain the direct interpretability of the model coefficients.)
2. Estimate the transformed model $y^{(\hat{\lambda})} = X\beta_{trans} + \epsilon$ and obtain the estimated model coefficients $\hat{\beta}_{trans}$ and covariance matrix Σ_{trans}.
3. Choose a hypothetical case or set of cases X_{pred} for which to calculate the quantity of interest. If one is interested in calculating a first difference, it is convenient to use X_{hi} and X_{lo}, where the first-difference $\Delta(y, X_{hi}, X_{lo}) = \text{med}(y|X_{hi}) - \text{med}(y|X_{lo})$.
4. Following King, Tomz, and Wittenberg (2000), for i from one to a large (e.g., 1,000) number of iterations n_{sims}:
    a. Simulate $\tilde{\beta}_{trans} \sim N\left(\hat{\beta}_{trans}, \Sigma_{trans}\right)$.
    b. If interested in the predicted value, then calculate and store $\tilde{Q}_i = \text{med}(y|X_{pred}, \tilde{\beta}_{trans}) =$

$BC^{-1}(X_{pred}\tilde{\beta}_{trans}, \hat{\lambda})$. If interested in the first-difference, the calculate and store $\tilde{Q}_i = \Delta(X_{hi}, X_{lo}, \tilde{\beta}_{trans}) = BC^{-1}(X_{hi}\tilde{\beta}_{trans}, \hat{\lambda}) - BC^{-1}(X_{lo}\tilde{\beta}_{trans}, \hat{\lambda})$.

5. Summarize the n_{sims} simulations. The mean or median of $\tilde{Q}$ serves as an estimate of the quantity of interest and the standard deviation of $\tilde{Q}$ serves as an estimate of the standard error. The 5th and 95th percentiles of $\tilde{Q}$ serve as an estimate of the (likely asymmetric) 90% confidence interval for the quantity of interest.

Dealing with Heavy-Tails: M-Estimation

In spite of the scant attention paid to robust estimators in political science, statisticians have developed and refined many robust methods since the seminal work of Box (1953) and Huber (1964). Huber and Ronchetti (2009) provide a detailed review of these developments and Anderson (2008) provides an accessible introduction. Adjudicating among the many robust alternatives to least squares is beyond the scope of our paper, but, to fix ideas, we do introduce one robust estimator in detail which has several desirable properties—the M-estimator with Tukey's biweight function. However, there are many other options: M-estimators with other objective functions (e.g., Huber 1973), LMS- and LTS-estimators (Rousseeuw 1984), S-estimators (Rousseeuw and Yohai 1984), and MM-estimators (Yohai 1987).

While least squares yields the coefficients that minimize the sum of the squared residuals, so that $\hat{\beta}^{ls} = \arg\min_b \sum_{i=1}^{n}(y_i - X_i b)^2$, M-estimators minimize an arbitrary, (usually) less-rapidly increasing function of the residuals $\hat{\beta}^{\rho} = \arg\min_b \sum_{i=1}^{n} \rho(y_i - X_i b)$. The function $\rho(\cdot)$ is typically chosen to be non-negative, symmetric about zero, and increasing away from zero. For example, Harden and Desmarais (2011) recommend the least absolute deviation (LAD) estimator (Dodge 1987) that such $\rho(\cdot) = \text{abs}(\cdot)$. However, other estimators offer better performance, particularly when the normal-linear model is approximately correct. In particular,

we recommend Tukey's biweight function, so that

$$\rho_{bw}(r_i) = \begin{cases} \dfrac{k^2}{6}\left\{1-\left[1-\left(\dfrac{r_i}{k}\right)^2\right]^3\right\} & \text{for } |r_i| \leq k \\ \dfrac{k^2}{6} & \text{for } |r_i| > k \end{cases},$$

where $r_i = y_i - X_i b$ and k is a tuning parameter usually set to 4.685 to ensure good performance under the normal-linear model. We refer to the M-estimator using the biweight objective function as the "biweight (BW) estimator." The BW estimator is a compelling alternative to the LAD estimator suggested by Harden and Desmarais (2011) for two reasons. First, the biweight objective function is redescending, which means that it has the ability to weight unusual observations all the way down to zero. The absolute value objective function, on the other hand, does downweight unusual observations, but these always received some weight. Secondly, the BW estimator is much more efficient than the LAD estimator when the errors are approximately normal.

Two cautions are in order. First, the optimization problem is not convex, so standard minimization routines can produce a local rather than the global minimum. This concern might lead researchers to choose another objective function, such as the Huber objective function. However, researchers can usually handle this problem in practice by using a good starting value, such as the LTS estimate. Second, because the solution is not scale invariant, the residuals $\hat{e}_i$ are standardized by a robust estimate of scale $\hat{\sigma}_{mad}$, which must of course be estimated jointly, so that $\hat{\beta}^{bw} = \arg\min_b \sum_{i=1}^n \rho_{bw}\left(\dfrac{y_i - X_i b}{\hat{\sigma}_{mad}}\right)$, where $\hat{\sigma}_{mad} = \dfrac{\text{med}\,(|y - Xb|)}{0.6745}$. Dividing by 0.6745 makes $\hat{\sigma}_{mad}$ a consistent estimator of the standard deviation of normal distribution.

While the theory for M-estimators remains less complete than the theory for least squares estimators, M-estimators do have desirable statistical properties. In general, M-estimators are consistent as long as (1) $\rho(\cdot)$ is convex or (2) the errors follow strongly unimodal distribution (i.e., decreasing away from zero). Because the biweight objective function is not convex, we

must assume that the errors follow a strongly unimodal distribution, which ensures that the estimates are consistent and distributed asymptotically normal.

M-estimators in general, and the biweight estimator in particular, have the desirable substantive property that they allow unusual cases to "be unusual." Least squares, on the other hand, sacrifices fit on typical cases to better fit unusual cases. Allowing unusual cases to stand out, though, is extremely important because unusual cases can inform and improve subsequent analyses. Knowing what cases fall outside the explanatory power of the model enables the researcher to ask "Why?" and raise issues relating to concepts, theory, and measurement that might otherwise have been missed.

Estimation

The model parameters $\hat{\beta}^{bw}$ and $\hat{\sigma}_{bw}$ can be quickly estimated jointly using the following iterative algorithm.

1. Start with initial estimate of the coefficients $\hat{\beta}^{(0)}$. The choice of initial estimator is not trivial. In the case of extreme outliers and/or many parameters, starting with least squares might lead the algorithm to a local minimum. We recommend using the least trimmed squares method discussed earlier to obtain starting values.
2. Extract the residuals $r^{(0)} = y - X\hat{\beta}^{(0)}$. Use these residuals to estimate the rescaled MAD so that $\hat{\sigma}^{(0)}_{mad} = \dfrac{\text{med}\left(|r^{(0)}|\right)}{0.6745}$.
3. For i from one until convergence:
    a. Using $\hat{\beta}^{(i-1)}$ and $\hat{\sigma}^{(i-1)}_{mad}$ assign weights w according to the function ρ and denote $\text{diag}(w) = W^{(i)}$.
    b. Calculate $\hat{\beta}^{(i)} = (X'W^{(i)}X)^{-1}X'W^{(i)}y$.
    c. Calculate $\hat{\sigma}^{(i)}_{mad} = \dfrac{\text{med}\left(|y - X\hat{\beta}^{(i)}|\right)}{0.6745}$
    d. The algorithm has converged when $r^{(i-1)} \approx r^{(i)}$.

If we assume that the errors are symmetricaly distributed about zero, then any objective

function ρ that is also symmetric about zero (including, for example, the biweight objective function) produces an unbiased estimate of the parameters. But this estimator is *linear* if and only if $\rho(r_i) = r_i^2$. Other choices of $\rho(\cdot)$ might produce better estimators than the BLUE estimator.

The theory for the variance for this broad class of unbiased M-estimators, though, is asymptotic. The required sample size for the asymptotic approximations to work well depends on the problem, but valid confidence intervals for small data can easily be computed by bootstrapping (Efron 1981 and Mooney and Duval 1993).

Monte Carlo Simulations

To understand and illustrate how the performance of the biweight (BW) estimator compares with the common least squares (LS) estimator and the least absolute deviation (LAD) estimator suggested by Harden and Desmarais (2011), we simulated from the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$, where $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = 1$ and the x_i's were generated from independent standard normal distributions. We used four different distribution for the errors.

- *Laplace distribution.* The Laplace distribution has tails that decrease exponentially, but behaves much differently from the normal distribution near zero. Rather than "shoulders," the Laplace distribution has a sharp peak at zero and can be thought of as combining two exponential distributions, one in the positive direction and the other in the negative direction. The LAD estimator is the maximum likelihood estimator when the errors follow a Laplace distribution.
- t_2 *distribution.* The t distribution with two degrees of freedom has very heavy tails. Because the least squares estimator weights all points equally (conditional on X), the extreme outliers produced by the t_2 distribution makes least squares a very inefficient estimator.

- t_{10} *distribution.* The t distribution with ten degrees of freedom has *slightly* heavier tails than the normal distribution. The t_{10} and normal distributions are so similar that a Shapiro-Wilk test of normality only correctly rejects the null in about 65% of repeated samples if 500 observation are simulated from a t_{10} distribution.[3] It is essentially impossible to spot the differences between the normal and t_{10} density functions without plotting the two directly on top of each other.
- *Normal Distribution.* The normal distribution yields the optimal conditions for the LS estimator. When the errors follow a normal distribution, the LS estimator has the smallest variance of all unbiased estimators.

For each of two different sample sizes, 100 and 1,000, and the four different error distributions, we simulated 10,000 data sets, estimated β_1 using the LS estimator, the LAD estimator, and the BW estimator. For each condition, we calculated the expected value of the estimate and the mean squared error (MSE). Table 1 provides the results for the sample size of 100 and 2 provides the results for the sample size of 1,000. These results show that all three estimators are unbiased regardless of the error distribution and sample size. Efficiency, however, varies considerably across the estimators.

	Mean				**Mean Squared Error**			
	Lapl.	t_2	t_{10}	Norm.	Lapl.	t_2	t_{10}	Norm.
Absolute Performance								
Least Squares	1.000	0.999	1.001	1.000	231.072	1571.227	149.507	87.103
Least Absolute Deviation	1.001	0.999	1.000	1.000	164.875	305.173	196.751	133.454
Tukey's Biweight	1.000	0.998	1.001	1.000	171.136	272.269	145.291	92.514
Relative Performance								
LAD/LS	1.001	1.000	0.999	1.000	0.714	0.194	1.316	1.532
BW/LS	1.000	0.999	1.000	1.000	0.741	0.173	0.972	1.062

Table 1: This table summarizes the Monte Carlo simulations for four different error distributions with a sample size of 100. Notice that the BW has the best or nearly best performance in each condition, while the LAD estimator performs quite poorly for the t_{10} and normal distributions and the LS estimator performs quite poorly for the Laplace and t_2 distributions.

The LAD estimator is the MLE when the errors follow a Laplace distribution, so, as we might expect, the LAD performs well for Laplace errors, with a MSE about 30% lower than the LS

[3] One needs about 750 samples to reach 80% power.

	Mean				Mean Squared Error			
	Lapl.	t_2	t_{10}	Norm.	Lapl.	t_2	t_{10}	Norm.
Absolute Performance								
Least Squares	0.999	0.999	1.000	1.000	20.173	165.585	12.793	9.998
Least Absolute Deviation	1.000	1.000	1.001	1.000	11.352	19.928	17.123	15.867
Tukey's Biweight	0.999	1.000	1.000	1.000	14.705	18.567	12.482	10.523
Relative Performance								
LAD/LS	1.001	1.000	1.000	1.000	0.563	0.120	1.338	1.587
BW/LS	1.000	1.000	1.000	1.000	0.729	0.112	0.976	1.053

Table 2: This table summarizes the Monte Carlo simulations identical to those in Table 1, except with a sample size of 1,000 rather than 100. As with Table 1, notice that the BW has the best or nearly best performance in each condition, while the LAD estimator performs quite poorly for the t_{10} and normal distributions and the LS estimator performs quite poorly for the Laplace and t_2 distributions.

estimate. However, the BW estimator also performs quite well for the Laplace distribution, with a MSE about 25% less than the LS estimator for $N = 100$. For $N = 1,000$, the results are similar. The MSE of the LAD estimator is about 43% than the LS estimator and the BW estimator is about 27% lower.

The t_2 distribution is nearly a worst case for the LS estimator, so both robust alternatives perform considerably better. For $N = 100$, the LAD estimator has an MSE about 79% lower than the LS estimator and the BW estimator has a MSE about 81% lower. For $N = 1,000$, the MSE for the LAD estimator is about 85% lower than the LS estimator and the MSE for the BW estimator is about 86% lower.

The t_{10} distribution is a much more interesting case, because it is very similar to a normal distribution. Indeed, even statistical tests have trouble distinguishing the t_{10} from the normal, even with large samples (e.g., with $N = 500$, the Shapiro-Wilk test of normality has about 65% power). In this case, the LAD estimator has a 34% larger MSE than the LS estimator for $N = 100$ and about a 32% larger MSE for $N = 1,000$. The BW estimator, on the other hand, shows a small *improvement* over the LS estimator, with an MSE about 2% *smaller* than the LS estimator for $N = 100$ and about 4% smaller for $N = 1,000$. This is crucial because it demonstrates that only a small deviation from normality is required before the BLUE estimator is no longer a BUE estimator.

The normal distribution is the optimal scenario for the LS estimator and it outperforms the

the LAD estimator considerably when the errors are normal. For both $N = 100$ and $N = 1,000$, the MSE for the LAD estimator is about 57% larger than the MSE for the LS estimator. However, the BW estimator performs nearly as well as the LS estimator for normal errors. The MSE for the BW estimator is only about 6% larger than the MSE for the LS estimator for both $N = 100$ and $N = 1,000$.

Among the estimators we consider, the BW estimator is not the most efficient estimator for the Laplace and normal distributions, but it is a *close* second. It is the most efficient estimator for the t_2 and t_{10} distributions. It considerably outperforms the LS estimator for Laplace and t_2 errors and the LAD estimator for normal and t_{10} errors. Thus, the BW estimator works quite well across a range of error distributions, whereas the LS and LAD estimators work well only in particular situations. And even in particular situations where the LS and LAD estimators work well, the BW estimator performs comparably.

To better understand how the heaviness of the tails of the error distribution affects the efficiency of these estimators, we repeated these simulates for t distributions for degrees of freedom ranging from two to thirty and sample sizes 25, 100, 500, and 2,000. Figure **??** shows the MSE of the LAD and BW estimators relative to the LS estimator.

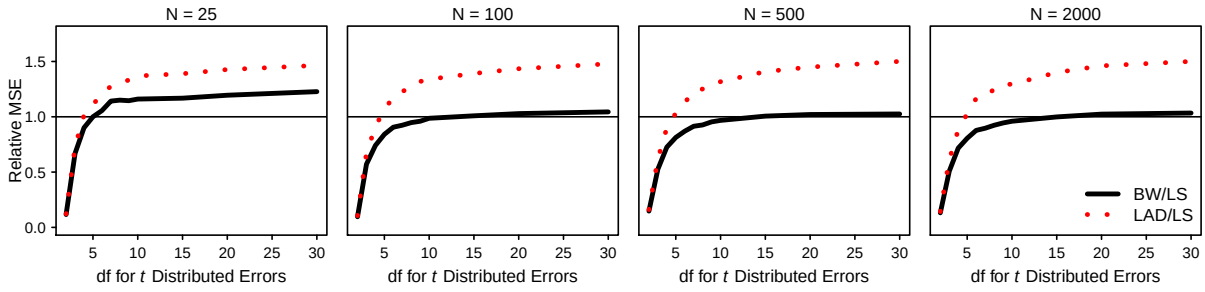

Figure 2: This figure shows the relative MSE for the LAD estimator and the BW estimator compared to the LS estimator for t distributed errors as the degrees of freedom varies. Notice that for very heavy-tailed distributions (e.g., two to five degrees of freedom), both the LAD and the BW estimators significantly outperform the LS estimator. And while the performance of the LAD estimator significantly worsens as the distribution becomes more normal, the BW estimator remains comparable to the LS estimator.

Notice that the LAD and BW estimators perform quite well for very heavy-tailed distributions

(i.e., degrees of freedom from two to four), but as the tails grow lighter, the LS estimator quickly begins to outperform the LAD estimator. Except for all but very heavy-tailed distributions, the LS estimator is considerably more efficient than the LAD estimator.

The BW estimator, other the other hand, is a much stronger competitor for the LS estimator. While the LS estimator is more efficient for lighter-tailed distributions (i.e., more than ten degrees of freedom), the difference is tiny except in very small samples. Indeed, for sample sizes of 100 or larger, the LS estimator is only about 5% more efficient than the BW estimator *at best*. This second simulation also suggests that the BW estimator works almost as well as the LS estimator under ideal conditions for LS estimator and considerably better across a wide range of other, substantively plausible scenarios.

Recommendation for Applied Researchers

When using the linear model, we suggest that researchers take steps to ensure that the assumption of normal errors makes theoretical and empirical sense.

1. Initially fit the model using least squares.
2. As a "robustness check," re-fit the model using a robust alternative, such as the biweight estimator.
3. If the inferences change (and even if not), carefully examine the residuals using histograms and QQ plots. Be careful to check for skewness.
4. If the residuals are not symmetrically distributed, then consider a transformation. This transformation might be critical because it allows the model to represent non-linear relationships implied by the skewness and allows the statistical model to more closely approximate the data. The log-transformation has a nice substantive interpretation, so it makes sense as a first cut, especially for variables naturally bounded below by zero or one. If the log-transformation over- or under-corrects the skewness, then the Box-Cox

transformation should ensure that the residuals are roughly symmetric.

5. Once the residuals are roughly symmetric, it makes sense to re-fit the model using least squares and a robust alternative. Especially if the residuals appear to have heavy tails, then the robust estimator might serve as a more efficient estimator. However, the robust estimator also allows for greater substantive interpretation as well, because it allows unusual cases to stand out.
6. Always pay close attention to the residuals from each model, especially differences, as these can be especially substantively informative.
7. To the extent that some cases seem unusual, especially in the context of the robust regression, give these cases careful review. Is it possible that these unusual outcomes are simply data entry errors? In light of these cases, can the measurement be improved? Might a subset of the cases be operating under a substantially different causal process that could be built into the statistical model?

Replication of Clark and Golder (2006)

Clark and Golder (2006) attempt to "rehabilitate" Duverger's (1963) classic explanation for the number of political parties in a system. They write:

> According to Duverger, the mechanical effect of electoral institutions favoring large parties creates incentives for strategic entry and strategic voting. Parties that have no chance of winning are encouraged to withdraw. If these parties fail to withdraw, then voters will have an incentive to vote strategically in favor of better placed parties. Thus disproportional systems with low district magnitudes are likely to reduce the demand for political parties created by social heterogeneity (p. 694).

For our replication, we focus specifically on their hypothesis:

HYPOTHESIS: Social heterogeneity increases the number of electoral parties only when the district magnitude is sufficiently large.

This suggests that the marginal effect of social heterogeneity should be positive and statistically significant under permissive electoral rules (i.e., large district magnitude) and about zero and statistically insignificant (though see Rainey 2014) under restrictive electoral rules (i.e., district magnitude near one).

To measure their key outcome variable, the number of political parties, Clark and Golder use the *effective* number of political parties (ENEP), where ENEP $= \sum_{i=1}^{n} \frac{1}{v_i^2}$, for v_i represents the vote share of party i in the election. Similarly, for their measure of social heterogeneity, Clark and Golder use the effective number of ethnic groups (ENEG), taken from Fearon's (2003) measure of ethnic fragmentation. To measure the restrictiveness of the electoral rules, Clark and Golder simply use the average district magnitude in the election.

Figure 3 plots the key explanatory variables, district magnitude and the effective number of ethnic groups, along the horizontal and vertical axes, respectively. The size of the points indicates the effective number of political parties for each election. The hypothesis predicts that the large points should lie in the upper-right section of the plot and small points should lie near the horizontal and vertical axes.

To test this hypothesis, Clark and Golder fit the following linear model using least squares:

$$\begin{aligned}\text{ENEP}_i = \beta_0 &+ \beta_1 \text{ENEG}_i + \beta_2 \log(\text{Magntiude}_i) + \beta_3 \text{Upper-Tier Seats}_i \\ &+ \beta_4 \text{Presidential Candidates}_i + \beta_5 \text{Proximity}_i \\ &+ \beta_6 \text{Ethnic}_i \times \log(\text{Magnitude}_i) + \beta_7 \text{Ethnic}_i \times \text{Upper-Tier Seats}_i \\ &+ \beta_8 \text{Presidential Candidates}_i \times \text{Proximity}_i + \epsilon_i \ ,\end{aligned}$$

The first key coefficient in this analysis is β_1, which captures the effect of social heterogeneity when district magnitude is one (i.e., the log of district magnitude is zero) and there are no upper-

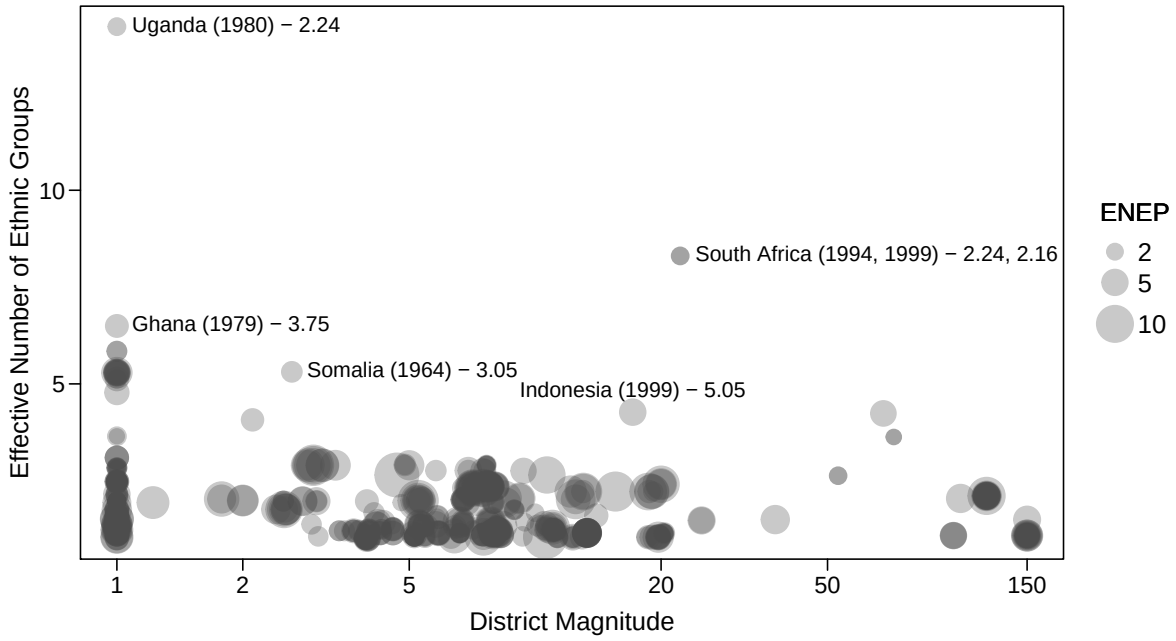

Figure 3: This figure shows the distribution of district magnitude (on the log scale) and ethnic heterogeneity, where the point sizes indicate the number of political parties. According to Clark and Golder's hypothesis, large points should lie in the upper-right portion of the plot and small points should lie along the axes.

tier seats. According to the hypothesis, β_1 should be about zero. The second key coefficient is β_6, which captures how the effect of social heterogeneity changes with the electoral rules. According to the hypothesis, β_6 should be positive, so that the effect of social heterogeneity becomes (perhaps more) positive as the district magnitude increases.

Clark and Golder use least squares to obtain their estimates of the model coefficients, but worry about their estimates of the standard errors. They write that "[t]he crucial thing to remember is that although OLS is consistent with longitudinal data, the standard errors may be incorrect" (p. 690). They discuss several options and ultimately settle on robust standard errors clustered by country, but demonstrate that their conclusions are robust to alternative approaches to estimating standard errors. However, they do not address the possibility of

a non-normal error distribution or its potential impact on the coefficient estimates. This is especially concerning given that the effective number of electoral parties is bounded below by zero, perhaps creating an error distribution with a skew to the right.

To get an initial sense of how the results might change using an alternative (perhaps more efficient) estimator, we replicated the estimates from four models in their Table 2 using least squares and the biweight estimator. For these initial estimates, we make no attempt to account for the clustered nature of the data in calculating the standard errors, but do supply the usual 90% confidence intervals to serve as a lower-bound on the uncertainty. Figure 4 presents these estimates and confidence intervals.

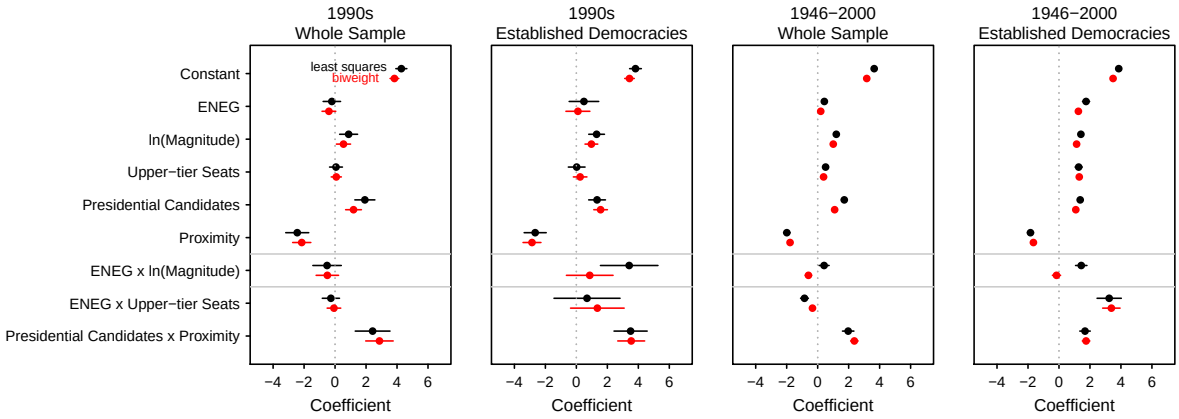

Figure 4: This figure shows the coefficient estimates of Clark and Golder's (2006) linear model using a least-squares and biweight estimator with explanatory variables standardized to have mean zero and standard deviation one-half. The black lines and points show the least squares estimates and 90% confidence intervals and the red lines and points show the biweight estimates and confidence intervals. Notice that the coefficient for the product of the effective number of ethnic groups and the district magnitude changes drastically with the choice of estimator.

The crucial estimate $\hat{\beta}_6$ changes substantially depending on the choice of estimator. This key estimate, which the theory suggests should be positive, remains negative in the 1990s sample including new democracies, shrinks substantially toward zero in the 1990s sample that includes only established democracies, and *becomes negative* in the large sample of countries from 1946-2000 that includes new democracies *and* the large sample that only includes established

democracies. When results depend on the choice of estimator, it is especially important to carefully examine the residuals.

For the remainder of our analysis, we focus on the estimates from large sample of countries from 1946-2000 that includes only established democracies. Figure 5 presents the histogram of the residuals from the least-squares estimates and Figure 6 presents the QQ plot for these residuals. Both figures indicate a substantial skew to the right. While this does not necessarily lead to biased estimates, it does, in our view, suggest that the linear model would be more appropriate for a transformed outcome variable. If the transformation makes the errors more closely approximate a normal distribution, then the least squares estimators will be more efficient. The model for the transformed outcome can capture potentially interesting substantive effects as well.

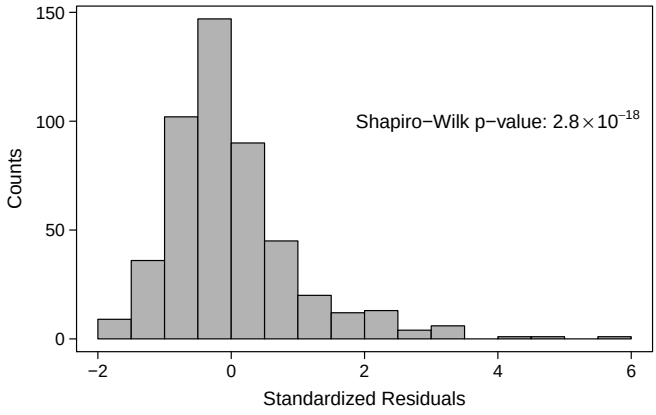

Figure 5: This histogram shows the distribution of the residuals from Clark and Golder's (2006) main model. Notice that these residuals do not seem approximately normal. They have a strong skew to the right. For example, one would rarely expect to observe residuals more that three standard deviations from zero if the assumption of normality holds. In these data, we have several residuals more that three standard deviations away and one nearly six standard deviations away. This suggests that some transformation of the outcome variable might be useful.

The maximum likelihood estimate of the Box-Cox transformation parameter λ is about $-\frac{1}{3}$ and the confidence interval does not include zero, which suggests that a log-transformation does not quite eliminate the skew. We re-estimated the model using both a log-transformation and

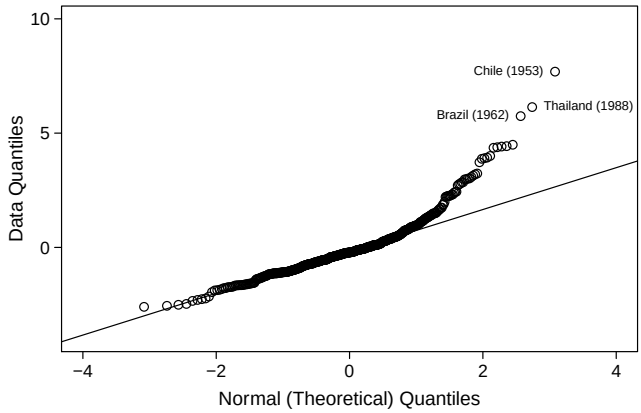

Figure 6: This QQ plot shows the deviation of the residuals from normality in Clark and Golder's (2006) main model. If the residuals were approximately normal, then the points in the QQ plot would approximately follow the line. However, notice that the positive residuals deviate sharply from the theoretical expectations. This also suggests that some transformation of the outcome variable might be useful.

Box-Cox transformation with $\lambda = -\frac{1}{3}$. Figure 7 presents the histograms of the residuals from these two regression models. Notice that log-transforming the effective number of electoral parties does not quite eliminate the skew in the residuals. However, the Box-Cox transformation with $\lambda = -\frac{1}{3}$ produces highly symmetric residuals.

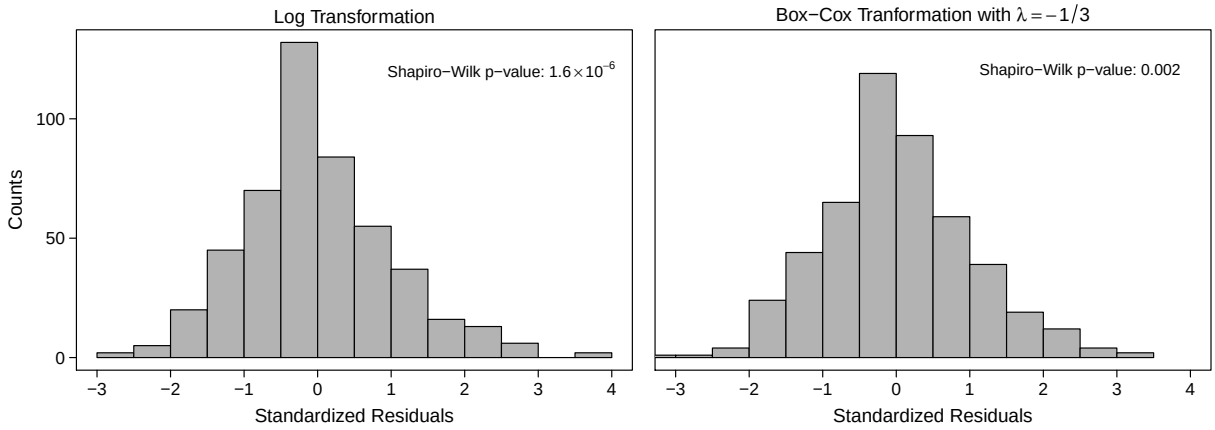

Figure 7: These histograms shows the distribution the residuals after transforming the outcome variable. The left panel shows that the log transformation does not quite remove all of the skew, but the right panel shows that the Box-Cox transformation with $\lambda = -\frac{1}{3}$ creates an approximately symmetric error distribution.

Figure 8 shows the QQ plot for the residuals. The left panel confirms that a right-skew remains after the log-transformation, as suggested by the histogram in the left panel of Figure 7. The right panel of Figure 8 confirms that the Box-Cox transformation removes much or all of this skew, as suggested by the right panel of Figure 7.

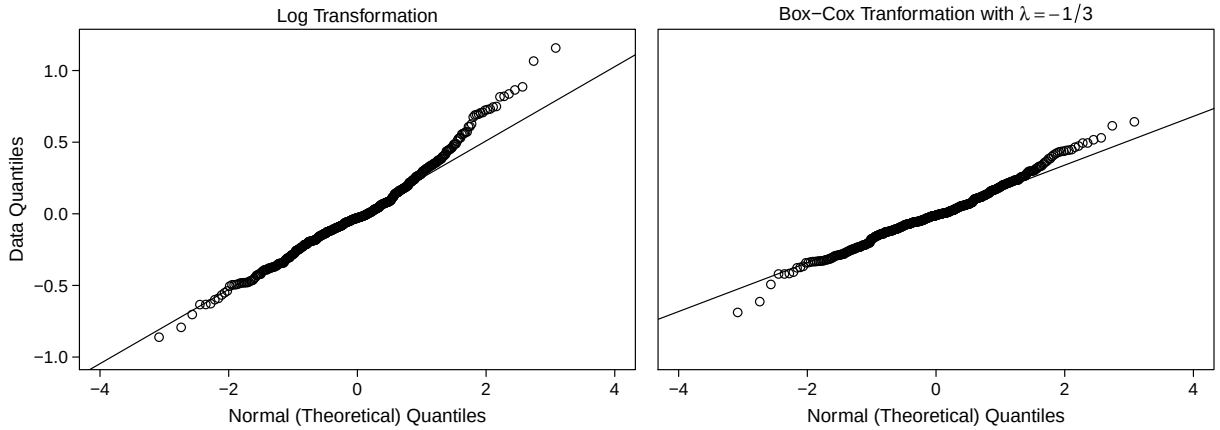

Figure 8: These QQ plots show residuals after transforming the outcome variable. The left panel shows that the log transformation does not quite remove all of the skew, but the right panel shows that the Box-Cox transformation with $\lambda = -\frac{1}{3}$ creates an approximately symmetric error distribution, though it has heavier tails than a normal distribution.

Although the Box-Cox transformation removes much or all of the skew, the residuals retain tails that are slightly heavier than the tails of the normal distribution. The residuals have larger positive and negative quantiles than one would expect under a normal distribution. This suggests the residuals follow a heavy-tailed distribution, perhaps resembling a t distribution with degrees of freedom in the range of six to twelve. Even after the Box-Cox transformation, the Shapiro-Wilk test rejects the null hypothesis of normality with $p = 0.002$. Assuming that these residuals follow a t-distribution, we can estimate the degrees of freedom using maximum likelihood, which points toward a t distribution with about ten degrees of freedom. Recall that the biweight estimator is a (slightly) more efficient estimator for t_{10} distributed errors. However, the least squares estimates *assumes* normally distributed errors and thus the error distribution might have *even heavier* tails.

Most importantly, after the Box-Cox transformation, the residuals are well-behaved. They are highly symmetric and only slightly heavy-tailed. In spite of a nearly ideal situation, the simulations in Tables 1 and 2 and Figure 2 suggest that the biweight estimate is slightly more efficient in this application. But these differences are slight so it can be valuable to carefully examine a least squares estimator and a more robust estimator.

Quantities of Interest

To estimate the quantity of interest—the effect of social heterogeneity as the permissiveness of the electoral rules varies—we re-estimate Clark and Golder's model with and without the Box-Cox transformation using both the least squares and biweight estimators. To obtain standard errors, we use the cluster bootstrap suggested by Harden (2012) to calculate the confidence intervals for each model.

Figure 9 shows the effect of increasing social heterogeneity from ENEG $= 1.06$ (10th percentile) to ENEG $= 2.48$ (90th percentile) as the district magnitude varies. The upper-left figure replicates Clark and Golder's approach (except for the cluster-bootstrap confidence intervals) and replicates their finding, which they summarize:

> [These results] clearly illustrate that in established democracies, ethnic heterogeneity significantly increases the number of parties once the electoral system is sufficiently permissive. This is exactly what Duverger's theory predicts. To be more specific, Figure 1a [our upper-left panel of Figure 9], based on the pooled model with established democracies, indicates that ethnic heterogeneity will increase the number of electoral parties once we move beyond nonpermissive electoral systems with single-member districts when [Magnitude $= 1$].

But this evidence breaks down once we adjust for the non-normality of the errors. The lower-left panel of Figure 9 shows that simply using the biweight estimator as a robust alternative

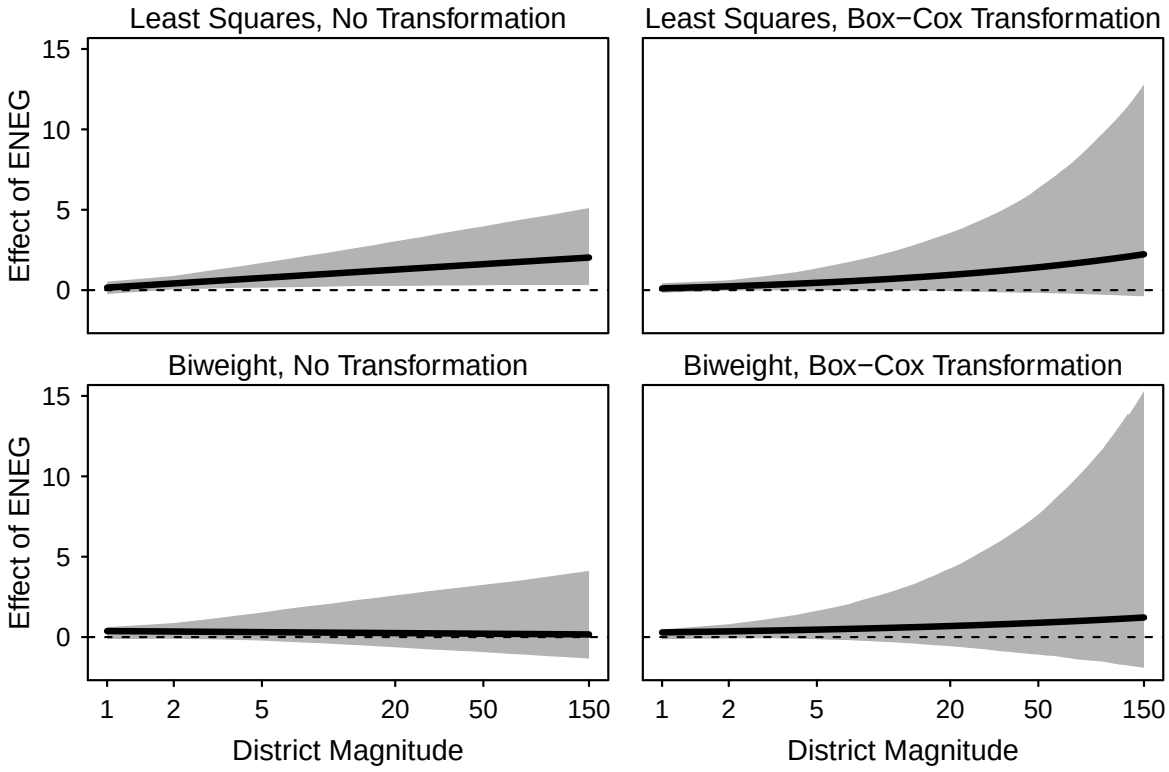

Figure 9: This figures shows the estimated effects and 90% confidence intervals of increasing social heterogeneity from ENEP = 1.06 (10th percentile) to ENEP = 2.48 (90th percentile) on the effective number of electoral parties as the district magnitude varies. Notice that only the model that does not use robust estimators or transform the outcome variable provides evidence in support of the hypothesis.

to least squares produces estimated interaction in the opposite direction as the hypothesis predicts, though small and not statistically significant. The upper-right panel of Figure 9 shows that simply transforming the outcome variable to make the data more consistent with the assumed normal-linear model substantially increases the uncertainty across the range of district magnitude, so that small negative effects are now plausible, as well as much larger positive effects.

However, we argue that the best approach for these data is to transform the outcome variable to obtain a roughly symmetric error distribution *and* use the biweight estimator to handle the remaining heavy tails shown in the left panel of Figure 8. This approach reduces

effect of social heterogeneity somewhat across the range of district magnitude.

Because the scale of the estimates and uncertain differ drastically across the choice to transform the outcome variable or not, as well as the range of district magnitude, we present the relevant quantities of interest in Table 3. In this case, we are interested in the effect of substantially increasing social heterogeneity from ENEG = 1.06 (10th percentile) to ENEG = 2.48 (90th percentile) on the number of political parties when district magnitude is one (10th percentile) and also when district magnitude is 14 (90th percentile). Since the hypothesis suggests that this effect should be larger when the district magnitude is larger, we are also interested in the difference between these two effects. For simplicity, we simply focus our discussion on the differences between the typical approach, least squares with no transformation, and the approach we recommend, transformation *and* robust estimation.

The first row of Table 3 suggests that in countries with single-member districts, a substantial increase in ENEG (10th to 90th percentile) increases the ENEP by about 0.16 [-0.24; 0.53] parties. On the other hand, in large, multimember districts (magnitude of 14), the same increase in social heterogeneity increases the ENEP by about 1.14 [0.26; 2.67] parties. This is just as the hypothesis predicts. Further, this increase of 0.98 [0.06; 2.65] is large and statistically significant.[4]

	First-Difference When $M = 1$		**First-Difference When $M = 14$**		**Second-Difference**	
	Est.	90% CI	Est.	90% CI	Est.	90% CI
No Transformation						
Least Squares	0.16	[-0.24; 0.53]	1.14	[0.26; 2.67]	0.98	[0.06; 2.65]
Biweight	0.37	[-0.12; 0.62]	0.26	[-0.51; 2.33]	-0.11	[-0.96; 2.05]
Box-Cox Transformation						
Least Squares	0.11	[-0.16; 0.43]	0.80	[-0.03; 2.84]	0.69	[-0.19; 2.73]
Biweight	0.29	[-0.13; 0.48]	0.62	[-0.41; 3.34]	0.33	[-0.78; 3.14]

Table 3: This table shows the quantities of interest from least squares and biweight estimates, with and without the Box-Cox transformation of the outcome variable. Notice that the least squares estimates without transforming the outcome variable are consistent with Clark and Golder's hypothesis. However, transforming the outcome variable, using the robust biweight estimator, or both substantially reduces the amount of evidence that these data offer in favor of the hypothesis.

[4]Using cluster-robust standard errors, Clark and Golder find that the product term is *not quite* significant. We use cluster-bootstrap standard errors and find the the coefficient is *barely* significant.

However, once we make an effort to account for the non-normality of the residuals by transforming the outcome variable *and* using the robust biweight estimator, this evidence for interaction weakens slightly. This model suggests that, in single-member districts, a substantial increase in social heterogeneity increases the ENEP by about 0.29 [-0.13; 0.48] parties. This is about twice Clark and Golder's initial estimate. In large, multimember districts (magnitude of 14), the estimate shrinks to 0.62 [-0.41; 3.34], which is about half of Clark and Golder's estimate. This leads to an estimated increase of 0.33 [-0.78; 3.14] in the effect of social heterogeneity as we move from single-member districts to large, multimember districts—about one-third of Clark and Golder's initial estimate.

Differences Between the Estimates

One major advantage of robust estimators is that these estimators allow unusual cases to stand out. Figure 10 compares the residuals from the least squares fit and the biweight fit after the Box-Cox transformation. Notice that the residuals tend to agree, with the exception of the 1980 election in Uganda, which is the largest residual in the biweight fit, but does not stand out among the least squares residuals.

Similarly, Figure 11 presents the 35 smallest weights from the biweight fit. Recall that as cases become increasingly inconsistent with the majority of the data, the biweight estimator increasingly down weights these cases, potentially to zero. As we might expect, given the residuals in Figure 10, the 1980 election in Uganda receives zero weight. This election is consistent with the hypothesis because it features single member districts, an extremely large number of ethnic groups, and only two major political parties. Unfortunately for the hypothesis, though, this case is inconsistent the the majority of the data.

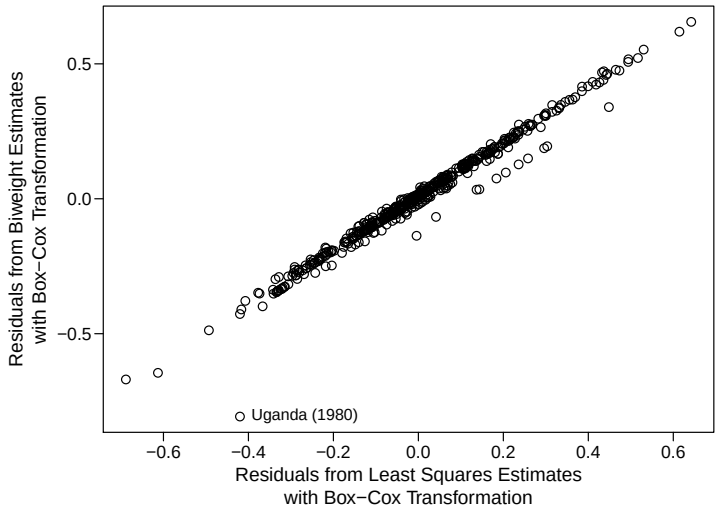

Figure 10: This figure shows the relationship between the least squares and biweight estimates after the Box-Cox transformation. Notice that the biweight estimate allows the unusual case of Uganda to stand out from the others.

Some Implications the Results

We offer a different approach than most applied research in political science adopts in practice. Rather than relying on the artificial Gauss-Markov theorem and BLUE estimators, we suggest that a careful consideration of the residuals is important. We show that there can be tremendous gains in efficiency from using robust estimators in the face of heavy-tailed error distributions, and we have shown that our approach leads to somewhat different conclusions than Clark and Golder (2006). But the approach we advocate is not powerful because it gives different results—it is powerful because it is substantively informative. Below we discuss several *substantive* points that one can take from our brief analysis of Clark and Golder's data.

Questions about the Validity of the Theory

At first glance, these results might make it seem that Duverger's logic lacks the empirical support suggested by the literature. Indeed, the evidence offered by the normal-linear model

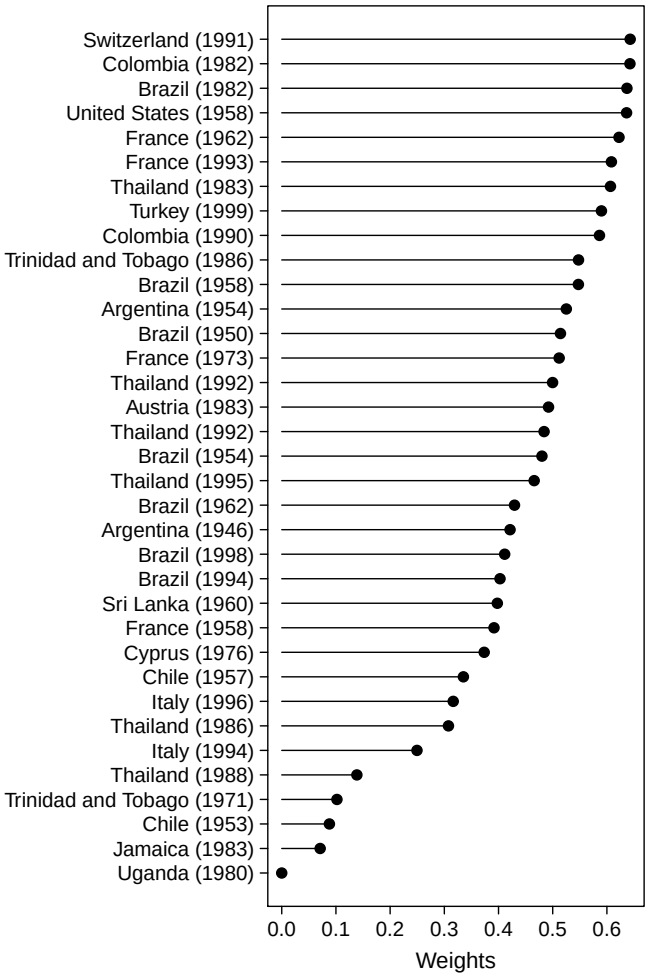

Figure 11: This figure shows the final weights implied by the biweight estimator. Because the 1980 election in Uganda is quite different from the remainder of the data, the biweight estimator downweights this case all the way to zero.

seems to hinge on the improper assumption of normal errors. Once we relax this assumption using a transformation and robust estimators, the evidence shrinks. However, rejection of Duverger's theory or Clark and Golder's (2006) analysis is premature for four reasons. First, the theoretical logic for Duverger's hypotheses is clear and compelling (e.g., Duverger 1963, Riker 1982, Cox 1997, Amorim Neto and Cox 1997, and Cox 1999). Second, many empirical studies beyond Clark and Golder (2006) find substantial empirical support for the hypotheses, including observational (e.g., Chhibber and Kollman 1998 and Singer and Stephenson 2009,

quasi-experimental (e.g., Blais et al. 2011 and Fujiwara 2011), and incentivized experimental studies (for a review, see Rietz 2008). Thirdly, in addition to the confidence intervals including effects that are inconsistent with the Clark and Golder's hypothesis, the confidence intervals now include *even larger* effects, which suggests the effect might be even larger than Clark and Golder suggest. Finally, the biweight M-estimator suggests several potential shortcomings in terms of concepts, theory, measurement that might currently undermine the evidence for these theories.

Questions about Measuring "Established Democracies"

These results suggest we have room for improvement in our measurement of "established democracies." Figure 11 suggests that the 1980 election in Uganda is quite different from the remaining cases, so we might look more closely at the context of the election. Uganda was under British rule until 1962, when it was granted independence from Britain. The 1962 elections led to the election of Milton Obote as prime minister. Four years later, facing scandal, Obote suspended the constitution, assumed all government powers, and declared himself president. In 1967, the parliament adopted a new constitution that solidified the expansive powers in the hands of the president. During this time, Obote banned all political parties and organizations accept his own (Kasozi 1993, p. 99). In 1971, Obote was ousted by Major General Idi Amin. During Amin's rule, between 50,000 and 300,000 civilians were killed (Kasozi 1993, p. 104). With the support of the Tanzanian military, which was working in conjunction with two Ugandan factions, one lead by Obote and the other Museveni, ended Amin's bloody rule in 1979 (Kasfir 1998, p. 53, and Kasozi 1993, pp. 124-127). In the following year, one leader was forcibly removed from office and the following government experienced yet another military coup. Elections were scheduled for 1980. By May of 1980, a group known as The Military Commission had taken power in order to give it to Obote. Obote wanted the appearance of being a democratic leader, therefore the Military Commission drafted electoral rules and held an election later

that year (Kasozi 1993, pp. 135-137). In an election that was fraught with corruption and irregularities, Obote's party won a majority of the seats—this 1980 Ugandan election is the unusual case in the data set. Amin's bloody rule continued into Obote's second tenure with a death toll between 100,000 and 500,000 over the next five years—a second conflict with civilian casualties on the scale of the recent tragedies in Darfur. Although elections were again held in 1996, 2001, 2006, and 2011, these elections featured irregularities, harassment, and did not lead to a change in power.

The context of the election raises question about whether Uganda was an "established democracy" in 1980, or even a democracy at all. One common measure offered by Cheibub, Gandhi, and Vreeland (2010), does indeed code Uganda as a democracy in 1980.[5] They code a country as a democracy if and only if it meets all of the following four conditions:

1. The chief executive must be chosen by popular election or by a body that was itself popularly elected.
2. The legislature must be popularly elected.
3. There must be more than one party competing in the elections.
4. An alternation in power under electoral rules identical to the ones that brought the incumbent to office must have taken place.

Although the election showed signs of irregularities and harassment of voters and candidates, it is plausible that Uganda in 1980 meets the first three criteria. But what about the forth? Because these authors code Uganda in 1980 as a democracy, they consider this condition met. However, this was a single election, nested between two military dictatorships, and during the winner's administration, he seized powers and dragged the country into civil war. It remains unclear to us how this meets the condition of democracy more broadly or the alternation condition more specifically.

[5]Another common measure of democracy, Polity IV, codes Uganda as a four in 1980, which falls outside range of six to ten that scholars typically consider democratic.

Indeed it is becoming common for authoritarian leaders to hold meaningless elections to present the illusion of democracy. Levitsky and Way (2002) argue that:

> It is essential, however, to distinguish regimes in which democratic institutions offer an important channel through which the opposition may seek power from those regimes in which democratic rules simply serve as to legitimate an existing autocratic leadership.

Secondly, even if Uganda was a democracy in 1980, was it "established"? Surely not, as this was the first election in 18 years following 10 years of brutal dictatorship. However, Clark and Golder define "established democracies" as countries that transitioned to democracy before 1989 (see note b in their Tables 1 and 2). Yet why is it important to focus on *established* democracies? Clark and Golder (2006, p. 706) summarize Duverger's argument:

> Another intriguing finding is that Duverger's theory receives much weaker support when we include elections from countries that transitioned to democracy after 1989. This finding is perhaps understandable if we think that party systems in newly democratic countries take a while to reach their equilibrium. It is interesting that Duverger himself took this view in regard to the fledgling democracies of Central Europe, Latin America, and Africa earlier in the 20th century. By warning about the danger of confusing multipartism with the absence of (fully institutionalized) parties, Duverger was indicating that he did not expect his theory to work particularly well in new democracies.

This suggests some rethinking of the concept of "*established* democracy" might be in order, at least in the context of Duverger's theory. For example, it surely must not be the case that countries transitioning to democracy before 1989 (e.g., Uganda in 1980) are "established" immediately after the transition. As Clark and Golder explain, the hypothesis assumes that the party system has reached an equilibrium. If systems that have not yet reached an equilibrium

are included in the analysis, then we can expect the data to offer little support for the hypothesis. Perhaps with some rethinking of how this equilibrium is reached, we can find a better measure that indicates when systems have reached an equilibrium.

Questions about Dynamics Prior to Equilibrium

In order to judge when systems are likely to be in an equilibrium, we need a stronger theory about the dynamics leading to an equilibrium number of parties. Duverger's logic offers a compelling explanation for the number of political parties *in equilibrium*, but this tells us little about the number of political parties prior to reaching an equilibrium or the dynamics that lead to an equilibrium. A theory about the dynamics is crucial to measuring whether a party system has stabilized, which, in turn, is crucial to testing Duverger's theory. Perhaps new democracies begin with many parties and the number shrinks until equilibrium. Perhaps they begin with few parties and the number grows until equilibrium. Perhaps it depends on the context. Some work has been done in this area (e.g., Moser 1999; Crisp, Olivella, and Potter 2012; and Ferrara 2011), but expanding the theoretical and empirical research in this area is crucial for studying the number of parties in equilibrium.

The fact that the 1980 election in Uganda featured essentially two political parties is quite interesting. Indeed, given the social heterogeneity, we would expect many more political parties— more than eight according to the biweight estimate and more than four according to the least squares estimate. Indeed, this case is so different from the rest of the data that it receives zero weight in the biweight estimator. But why? One potential explanation might rely on the fact that Milton Obote, elected 18 years before, had returned. The people were quite familiar with Obote. Perhaps this enabled groups to coordinate in support of or in opposition to Obote. It is difficult to know without a thorough analysis, but our analysis hints that the dynamics toward an equilibrium number of parties might depend on the nature of the prior regime. When parties representing the prior authoritarian regime participate in the election,

perhaps this helps solve the coordination problem. On the other hand, when the prior regime does not participate in the elections, perhaps this makes coordination difficult. Our point is not to answer this question but to highlight that a careful consideration of the residuals along with an estimation approach that allows unusual cases to be unusual encourages the researcher to raise these types of questions.

Conclusion

In this paper, we adopt and defend a skeptical perspective toward least squares and explain the importance of carefully scrutinizing residuals. We first note that the restriction to linear estimators, as required by the Gauss-Markov theorem, is unnecessary, artificial, and counterproductive. When the errors do not follow a normal distribution, restricting ourselves to linear estimators limits our ability to efficiently estimate the model coefficients. And as Berry and Feldman (1985) note, the assumption of normal errors is usually difficult to defend using either theory or data.

Secondly, we use Monte Carlo simulations to show that our preferred estimator, an M-estimator with a biweight objective function, performs almost as well as least squares for normal errors, slightly better than least squares for small deviations from normality, and much better than least squares for large deviations from normality. In our view, political scientists should prefer this robust behavior to the more fragile least squares estimator.

Thirdly, we use a practical example to show that a more robust estimator can *constructively* alter substantive conclusions. Robust estimators not only more efficiently estimate the model coefficients under deviations from normality, but more importantly, allow unusual cases to stand out as unusual. This, in turn, allows researchers to identify and carefully investigate these cases and improve the theoretical model, the statistical model specification, and/or the measurement of the concepts.

References

Amorim Neto, Octavio, and Gary W. Cox. 1997. "Electoral Institutions, Cleavage Structures, and the Number of Parties." *American Journal of Political Science* 41(1):149–171.

Anderson, Robert. 2008. *Modern Methods for Robust Regression*. Thousand Oaks, CA: Sage.

Angrist, Joshua D., and Jörn-Seffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

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

Berry, William D., and Stanley Feldman. 1985. *Multiple Regression in Practice*. Quantitative Applications in the Social Sciences Thousand Oaks, CA: Sage.

Blais, André, Romain Lachat, Airo Hino, and Pascal Doray-Demers. 2011. "The Mechanical and Psychological Effects of Electoral Systems a Quasi-experimental Study." *Comparative Political Studies* 44(12):1599–1622.

Box, George E. P. 1953. "Non-Normality and Tests on Variances." *Biometrika* 40(3/4):318–335.

Box, George E. P., and David R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society, Series B* 26(2):211–252.

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

Cheibub, José Antonio, Jennifer Gandhi, and James Raymond Vreeland. 2010. "Democracy and Dictatorship Revisited." *Public Choice* 143(1-2):67–101.

Chhibber, Pradeep, and Ken Kollman. 1998. "Party Aggregation and the Number of Parties in India and the United States." *American Political Science Review* 92(2):329–342.

Clark, William Roberts, and Matt Golder. 2006. "Rehabilitating Duverger's Theory: Testing the Mechanical and Strategic Modifying Effects of Electoral Laws." *Comparative Political Studies* 39(6):679–708.

Cox, Gary. 1997. *Making Votes Count: Strategic Coordination in World's Electoral Systems*. Cambridge: Cambridge University Press.

Cox, Gary. 1999. "Electoral Rules and Electoral Coordination." *Annual Review of Political Science* 2(1):145–161.

Crisp, Brian F., Santiago Olivella, and Joshua D. Potter. 2012. "Electoral Contexts that Impede Voter Coordination." *Electoral Studies* 31(1):143–158.

Dodge, Yadolah, ed. 1987. *Statistical Data Analysis Based on the L1-Norm and Related Methods*. Amsterdam: North-Holland.

Duverger, Maurice. 1963. *Political Parties: Their Organization and Activity in the Modern State*. New York: Wiley.

Efron, Bradley. 1981. "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Methods." *Biometrika* 68(3):589–599.

Ferrara, Federico. 2011. "Cleavages, Institutions and the Number of Parties: A Study of Third Wave Democracies." *Journal of Elections, Public Opinion, and Parties* 21(1):1–27.

Freedman, David A. 2006. "On the So-Called 'Huber Sandwich Estimator' and 'Robust Standard Error'." *The American Statistician* 60(4):299–302.

Fujiwara, Thomas. 2011. "A Regression Discontinuity Test of Strategic Voting and Duverger's Law." *Quarterly Journal of Political Science* 6(3-4):197–233.

Gujarati, Damodar N. 2004. *Basic Econometrics*. 4th ed. Boston, MA: McGraw Hill.

Harden, Jeffrey J. 2012. "Improving Statistical Inference with Clustered Data." *Statistics, Politics, and Policy* 3(1):1–27.

Harden, Jeffrey J., and Bruce A. Desmarais. 2011. "Linear Models with Outliers: Choosing between Conditional-Mean and Conditional-Median Methods." *State Politics and Policy Quarterly* 11(4):371–389.

Huber, Peter J. 1964. "Robust Estimation of a Location Parameter." *The Annals of Mathematical Statistics* 35(1):73–101.

Huber, Peter J. 1973. "Robust Regression: Asymptotics, Conjectures, and Monte Carlo." *The Annals of Statistics* 1(5):799–821.

Huber, Peter J., and Elvezio M. Ronchetti. 2009. *Robust Statistics*. Vol. 2nd Hoboken, NJ: Wiley.

Kasfir, Nelson. 1998. "'No-Party Democracy' in Uganda." *Journal of Democracy* 9(2):49–63.

Kasozi, A. B. K. 1993. *Social Origins of Violence in Uganda, 1964-1985*. Canada: McGill-Queen's University Press.

King, Gary, and Margaret E. Roberts. 2014. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It." Forthcoming in *Political Analysis*.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):341–355.

Krueger, James S., and Michael S. Lewis-Beck. 2008. "Is OLS Dead?" *The Political Methodologist* 15(2):2–4.

Levitsky, Steven, and Lucan A Way. 2002. "The Rise Of Competitive Authoritarianism." *Journal of Democracy* 13(2).

Mira, Antonietta. 1999. "Distribution-Free TEst for Symmetry Based on Bonferroni's Measure." *Journal of Applied Statistics* 26(8):959–972.

Mooney, Christopher Z., and Robert D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Quantitative Applications in the Social Sciences Newbery Park, CA: Sage.

Moser, Robert G. 1999. "Electoral Systems and the Number of Parties in Post-Communist States." *World Politics* 51(3):359–384.

Rietz, Thomas. 2008. "Three-Way Experimental Election Results: Strategic Voting, Coordinated Outcomes and Duverger's Law." In *Handbook of Experimental Economics Results*, ed. Charles R. Ploot, and Vernon L. Smith. Vol. 1 North-Holland.

Riker, William H. 1982. "The Two-Party System and Duverger's Law: An Essay on the History of Political Science." *The American Political Science Review* 76(4):753–766.

Rousseeuw, Peter J. 1984. "Least Median of Squares Regression." *The Journal of the American Statistical Association* 79(388):871–880.

Rousseeuw, Peter J., and Victor Yohai. 1984. "Robust Regression by Means of S-Estimators." In *Robust and Nonlinear Time Series Analysis*, ed. Jürgen Franke, Wolfgang Härdle, and Douglas Martin. Vol. 26 of *Lecture Notes in Statistics* Springer US pp. 256–272.

Singer, Matthew M., and Laura B. Stephenson. 2009. "The Political Context and Duverger's Theory: Evidence at the District Level." *Electoral Studies* 28(3):480–491.

Western, Bruce. 1995. "Concepts and Suggestions for Robust Regression Analysis." *American Journal of Political Science* 39(3):786–817.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4):817–838.

Wooldridge, Jeffrey M. 2013. *Introductory Econometrics: A Modern Approach*. 5th ed. Mason, Ohio: South-Western Cengage Learning.

Yohai, Victor. 1987. "High Breakdown-Point and High Efficiency Robust Estimates for Regression." *The Annals of Statistics* 15(2):642–656.