When BLUE Is Not Best

Non-Normal Errors and the Linear Model*

Dan Baissa†

Carlisle Rainey‡

Abstract

Researchers studying the consequences of comparative electoral institutions, as well as other areas of political and social science, often estimate linear models of continuous outcomes of interest using least squares. While it is well known that least-squares estimates are often sensitive to single, influential data points, this knowledge has not led to appropriate practices when using least-squares estimators. We highlight the importance of using more robust estimators in conjunction with least squares 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) an show that the residuals are highly non-normal under their model specification. Applying an empirically-chosen transformation and/or using a robust estimator weakens their key conclusions about the conditional relationship between social heterogeneity and electoral rules in influencing the number of political parties.

*We thank Bill Clark and Matt Golder for making their data available to us. The analyses presented here were conducted with R 3.1.0. All data and computer code necessary for replication are available at github.com/carlislerainey/heavy-tails .

†Dan Baissa is an M.A. student in the Department of Political Science, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (dkbaissa@buffalo.edu).

‡Carlisle Rainey is Assistant Professor of Political Science, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (rcrainey@buffalo.edu).

Introduction

In this manuscript, our goals are to (1) highlight powerful, robust alternatives to least-squares estimators, (2) provide concrete, practical advice to substantive researchers using linear models, and (3) provide a compelling example that illustrates the importance of robust estimators.

[Add more introduction later.]

Is a BLUE Estimator the "Best" Estimator?

The linear model can be written as $y = X\beta + \epsilon$.[1] 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 uniform minimum variance unbiased estimator (UMVUE) (Casella and Berger 2002, pp. 334-342, and Wooldridge 2013, pp. 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 (i.e., uniform minimum variance) *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

[1] 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.

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. It 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 is restricted 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 $\hat{\beta}^{LS}$. Noting that $\frac{\partial S(\hat{\beta}^{LS})}{\partial \hat{\beta}^{LS}} = \sum_{i=1}^{n} 2(y_i - X_i\hat{\beta}^{LS})(-X_i) = 0$ 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 $k = \{0,1,2,...,k\}$. 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 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

$$y_i = \beta_0 + \beta_1 \log(x_i) + \epsilon_i,$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \text{ or}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i.$$

These are all linear models and illustrate 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.

Further, there is a substantive argument *against* linear estimators (i.e., weighting all observations equally). We suggest two potentially desirable properties of estimators. First, we might prefer estimators that provide an excellent fit to *most* of the data rather rather than a mediocre fit to *all* the data. Secondly, we might prefer an estimate that treats unusual data equally–as inconsistent the the model. Neither of these behaviors is possible with linear estimators.

For example, consider the estimates shown in the left panel of Figure 1. Which of the two estimates, A or B, best summarizes the relationship between the explanatory and outcome variable? Estimate A fits all the data reasonably well, but estimate B provides an excellent summary for most of the data. Which is preferred? At least in some cases, we might prefer estimate B because we wish to discount the three unusual cases as inconsistent with the model (perhaps these cases are elections marred with scandals, etc.).

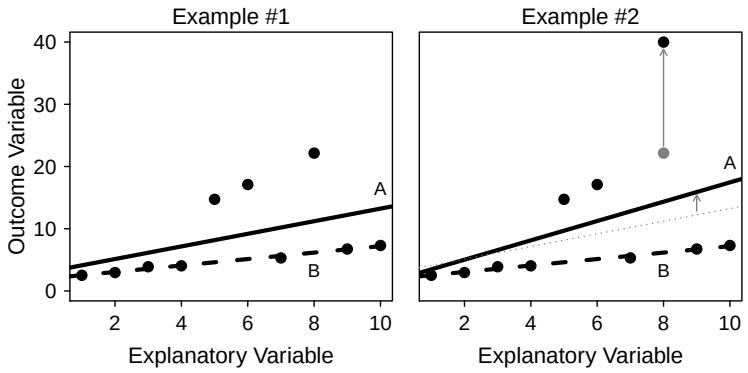

Figure 1: Two scatterplots of hypothetical data. [Make sure that captions have a similarly styled lead.] The left panel illustrates that one estimator (A) might provide an excellent fit to most of the data while another estimator (B) might provide a mediocre fit to all the data. The right panel illustrates that one estimator (A) might ignore changes among the unusual cases while anther estimator (B) might be quite responsive.

Secondly, we might prefer an estimator that gives zero weight to cases that fall outside the explanatory power of the model. Is is better to talk about an estimator or estimators? These cases seem somehow different and we might prefer that these unusual data to not degrade

the excellent fit for the majority of the data. That is, if one unusual case becomes *even more* unusual, the estimates should not change. Stated differently, we might prefer an estimator that *ignores* changes in the unusual cases, such as estimator B in the right panel of Figure 1, over an estimator that *resonds to* changes in the unusal cases, such as estimator A in the right panel of Figure 1.

Even if one does not to use an estimator with either of these behaviors, then following example illustrates the artificiality of the linearity restriction. 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.

It is not often appreciated, at least in practice, that if the errors do not follow independent and identical normal distributions, then the least squares is no longer the UMVUE–other well-understood and easily computed estimators might outperform least squares. Indeed, for some error distributions, least squares might be substantially inferior.

Many researchers simply assume a statistical model for which a UMVUE is easily available 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 not 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 unbiased (or perhaps nearly so) with efficiency comparable to, but less than, least squares.
2. When the deviation from the normal-linear model is small, the estimator should be

unbiased (or perhaps nearly so) with efficiency comparable to, but 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 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 2 shows the distribution of the estimated slopes using least squares.

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 2 shows that it is essentially unbiased and, compared to the least squares estimator, more efficient. 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 make this gain 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

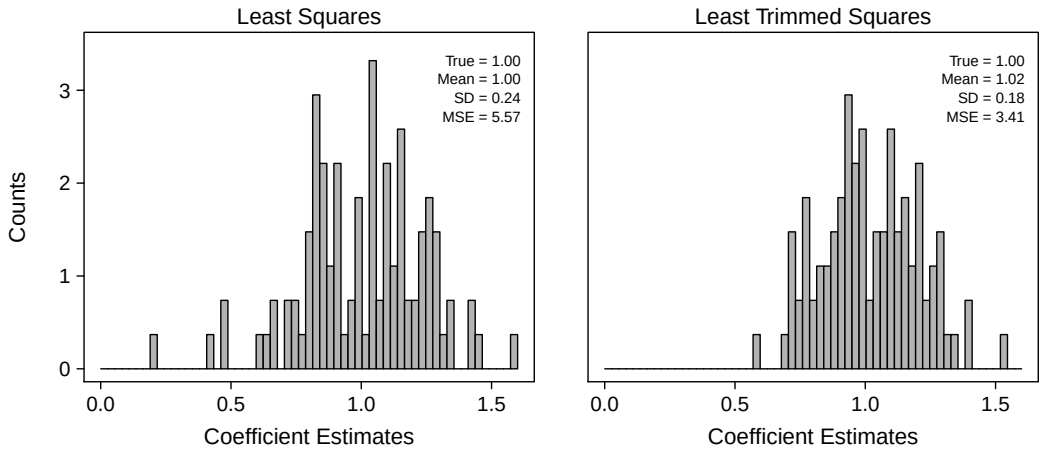

Figure 2: These two histograms show the sampling distribution for the least squares estimator and the least trimmed squares estimator under the true model $y = x + \epsilon$, where $\epsilon \sim t_3$. Notice that despite least squares being the best linear unbiased estimator for the problem, the least trimmed squares estimator is a better estimator.

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 2012 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, but it has received only 99 citations, or about 2% of the citations that Beck and Katz have received. Similarly, Desmarais and Harden (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.

This focus on obtaining reasonable standard errors at the expense of reasonable estimates can be seen in Gujarati (????, p. ??,), a popular textbook for political science classes focusing on linear models. Though the text deals with robust standard errors in some detail, Gujarati 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 (2008) devote an entire chapter to robust standard errors and completely ignore robust estimation of model coefficients. Wooldridge (????) 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 error distributions indicate a lack of model fit. It is sometimes theoretically intuitive to believe that the explanatory variables should have increasing effects 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.

The Box-Cox Transformation

While we agree with the spirit of the suggestion to log-transform 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

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

, 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 if $\hat{\lambda} \approx 1$ suggests no transformation is needed and $\hat{\lambda} \approx zero$ 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 $\text{med}(y_{new}|X_{new})$ 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 $\text{med}(y|X)$, the median has another desirable property. Because the log-transformation is order-preserving, $\text{med}[\log(y)|X] = \log[\text{med}(y|X)]$, which means that the log *can* easily be undone because $e^{med[\log(y)|X]} = e^{\log[\text{med}(y|X)]} = \text{med}(y|X)$. Therefore, by adopting $\text{med}(y|X)$ 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., $\text{med}[\log(y)] \rightarrow \text{med}(y)$). 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 $\text{med}(y)$ 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 $y^{(\lambda)} = X\beta + \epsilon$ 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. Calculate and store $\tilde{Q}_i = \text{med}(y|X_{red}, \tilde{\beta}_{trans}) = X_{red}\tilde{\beta}_{trans}$ or, if interested in the first-difference, $\tilde{Q}_i = \tilde{\Delta}(y, X_{hi}, X_{lo}, \tilde{\beta}_{trans}) = X_{hi}\tilde{\beta}_{trans} - X_{lo}\tilde{\beta}_{trans}$.
5. Summarize the n_{sims} simulations. The mean or median of $\tilde{Q}_i$ serves as an estimate of $\text{med}(y|X_{pred})$, the standard deviation of $\tilde{Q}_i$ serves as an estimate of the standard error of $\text{med}(y|X_{pred})$, and the 5th and 95th percentiles of $\tilde{Q}_i$ serve as an estimate of the (likely asymmetric) 90% confidence interval for $\text{med}(y|X_{pred})$.

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 that are beyond the scope of this paper, such as M-estimators with other objective functions (e.g., Huber 1973), LMS- and LTS-estimators (Rousseeuw 1984), S-estimators (Rousseeux 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 minimizes an arbitrary, 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 decreasing away from zero. For example, Harden and Desmarais (2011) recommend the least absolute deviation (LAD) estimator (Dodge 1987) that such $\rho(\cdot) = abs(\cdot)$. 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

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

. where $r_i = y_i - X_i b$ and k is a tuning parameter usually set to ???? 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 routines can produce a local rather than the global minimum. 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 $\hat{\beta}^{bw} = \arg\min_b \sum_{i=1}^{n} \rho_{bw}\left(\frac{y_i - X_i b}{\hat{\sigma}_{mad}}\right)$, where $\hat{\sigma}_{mad} = \frac{\text{med}\,(|y - X\underline{|})}{0.6745}$. 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 unusal 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 $r^{(0)} = X\hat{\beta}^{(0)}$. Use these residuals to estimate the rescaled MAD so that $\hat{\sigma}^{(0)}_{mad} = \dfrac{\text{med}\left(|y - X\hat{\beta}^{(0)}|\right)}{0.6745}$.
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 $\text{diag}(w) = W$.

b. Calculate $\hat{\beta}^{(i)} = (X'WX)^{-1}X'Wy$.

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 ????, Duval and Mooney ????). maybe more here on required sample size.

Monte Carlo Simulations

To understand and illustrate how the performance of the biweight (BW) estimator compares with the least squares (LS) and least absolute deviation (LAD)8 estimators, 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.[2] 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.

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 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 $N = 1,000$,

[2] 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	1.019	0.996	1.012	0.998	2.176	6.076	1.145	0.698
Least Absolute Deviation	1.001	0.983	1.017	0.996	1.411	1.783	1.592	1.164
Tukey's Biweight	1.000	0.995	1.016	1.000	1.656	2.010	1.125	0.838
Relative Performance								
LAD/LS	0.983	0.987	1.005	0.998	0.648	0.293	1.390	1.666
BW/LS	0.982	0.998	1.003	1.001	0.761	0.331	0.982	1.200

Table 1: Summarizes of 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.

	Mean				Mean Squared Error			
	Lapl.	t_2	t_{10}	Norm.	Lapl.	t_2	t_{10}	Norm.
Absolute Performance								
Least Squares	1.000	0.998	1.000	1.000	17.287	368.347	12.772	10.144
Least Absolute Deviation	1.000	0.999	1.000	1.000	9.642	21.167	16.966	15.787
Tukey's Biweight	1.000	1.000	1.000	1.000	12.523	19.748	12.274	10.683
Relative Performance								
LAD/LS	1.000	1.002	0.999	1.000	0.558	0.057	1.328	1.556
BW/LS	1.000	1.002	1.000	1.000	0.724	0.054	0.961	1.053

Table 2: Summaries of 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.

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 a MSR 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., $N = 500 \rightarrow \approx 40\%$ power for a Shapiro-Wilk test of normality). In this case, the LAD estimator has a MSE about 34% more than the LS estimator for $N = 100$ and about 32% more for $N = 1{,}000$. The BW estimator, on the other hand, shows a small *improvement* over the LS estimator, with an MSE about 2% *less* than the LS estimator for $N = 100$ and about 4% less 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 the UMVUE estimator.

The normal distribution is the optimal scenario for the LS estimator and it outperforms the the LAD estimator considerably with 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. 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 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.

Notice that the LAD and BW estimators perform quite well for very heavy tailed distribution (i.e., degrees of freedom from two to four), but as the tails grow lighter, the LS

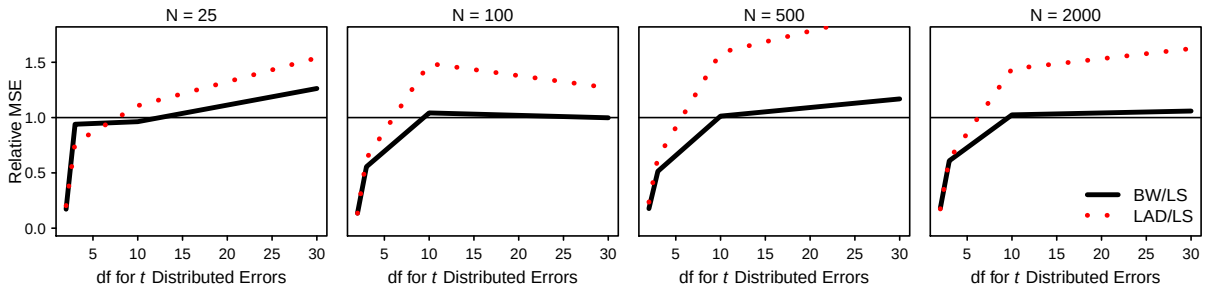

Figure 3: The relative MSE for the LAD estimator and the BW estimator compared to the LS estimator for t distributed errors as the degrees of freedom varies. Notice that for very heavy-tailed distributions (e.g., two to five degrees of freedom), both the LAD and the BW estimators significantly outperform 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.

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 very small 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. These steps ensure that inferences are consistent with the majority of the data and not driven exclusively by a small, unusual minority.

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 (????) 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 4 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.

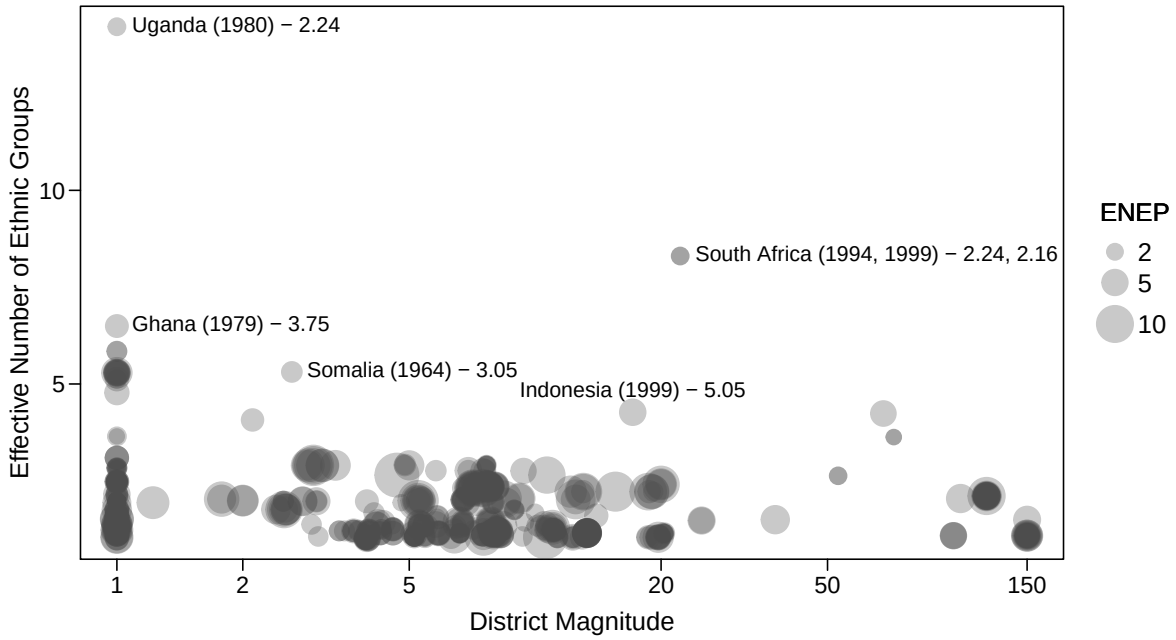

Figure 4: This figure shows the distribution of district magnitude (on the log scale) and ethnic heterogeneity, where the point sizes indicate the number of political parties. According to Clark and Golder's hypothesis, large points should lie in the upper-right portion of the plot and small points should lie along the axes.

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

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

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-tier seats. According to the hypothesis, β_1 should be about zero. The second key coefficient is β_6, 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 strong 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 5 presents these estimates and confidence intervals.

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

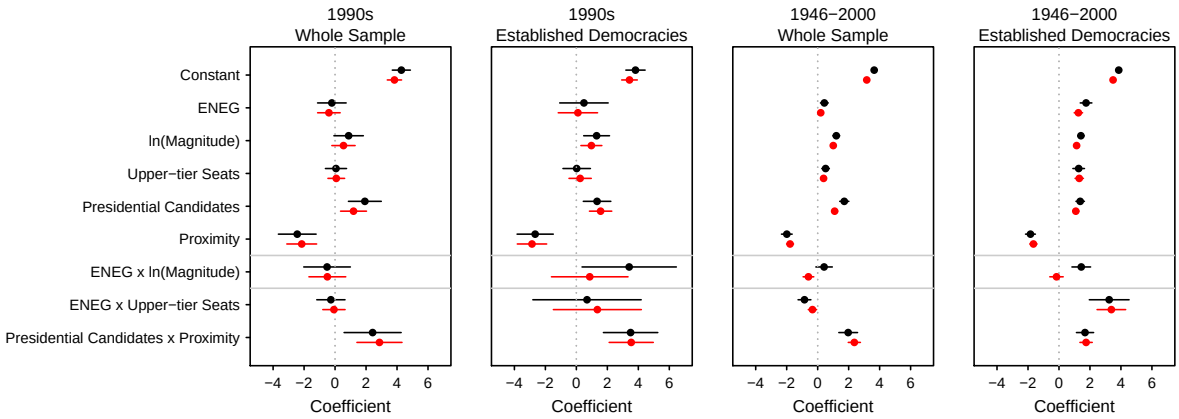

Figure 5: think a bit about this caption Replication of Clark and Golder (2006) using the biweight estimator with explanatory variables standardized to have mean zero and standard deviation one-half. The black lines and points show the least squares estimates and 90% confidence intervals and the red 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.

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 6 presents the histogram of the residuals from the least-squares estimates and Figure 7 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.

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

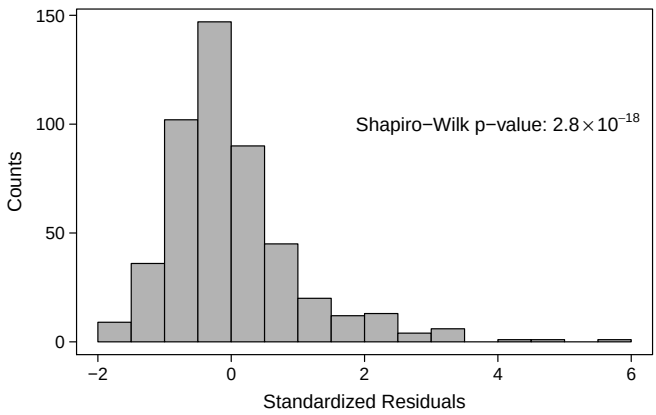

Figure 6: A histogram showing the distribution of the residuals from Clark and Golder's (2006) main model. Notice that these residuals do not seem approximately normal. They have a strong skew to the right. For example, one would rarely expect to observe residuals more that three standard deviations 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.

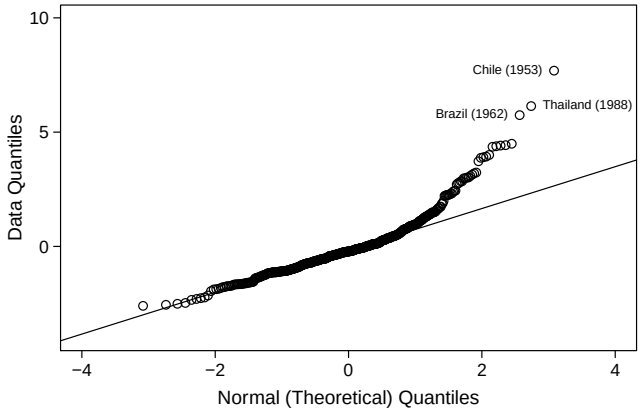

Figure 7: A QQ plot showing the deviation of the residuals from normality. If the residuals were approximately normal, then the points in the QQ plot would approximately follow the line. However, notice that the positive residuals deviate sharply from the theoretical expectations. This also suggests that some transformation of the outcome variable might be useful.

and Box-Cox transformation with $\lambda = -\frac{1}{3}$. Figure 8 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.

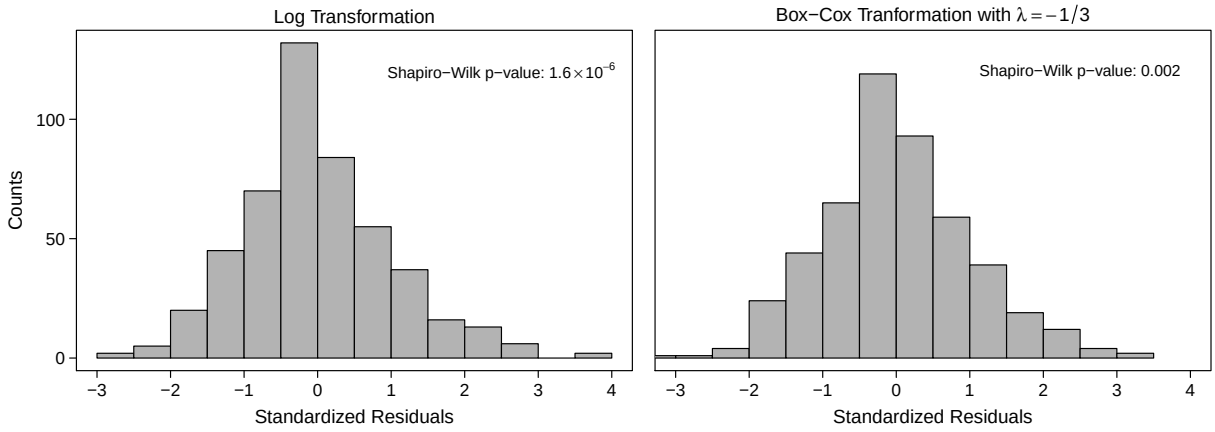

Figure 8: Histograms of the residuals after transforming the outcome variable. The left panel shows that the log transformation does not quite remove all of the skew, but the right panel shows that the Box-Cox transformation with $\lambda = -\frac{1}{3}$ creates an approximately symmetric error distribution.

Figure 9 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 8. The right panel of Figure 9 confirms that the Box-Cox transformation removes much or all of this skew, as suggested by the right panel of Figure 8.

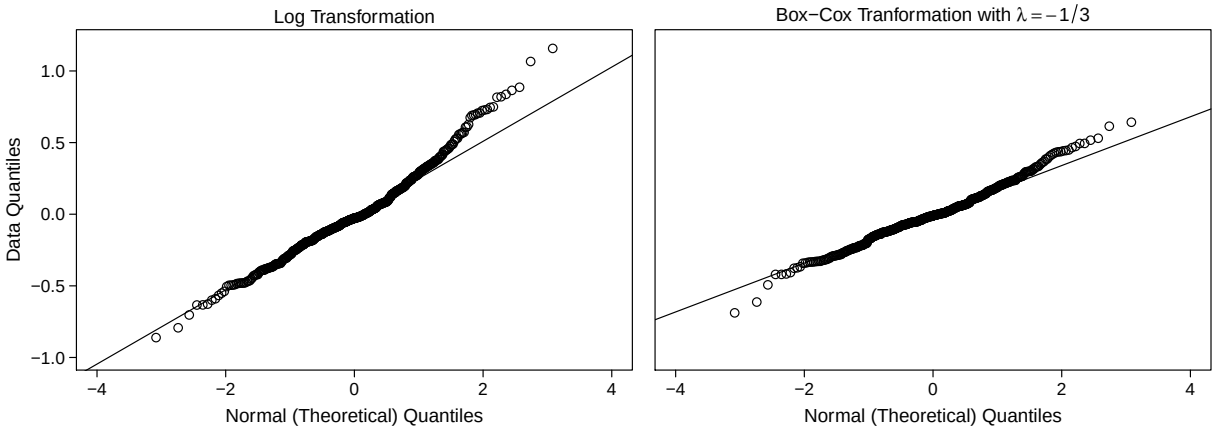

Figure 9: QQ plots of the residuals after transforming the outcome variable. The left panel shows that the log transformation does not quite remove all of the skew, but the right panel shows that the Box-Cox transformation with $\lambda = -\frac{1}{3}$ creates an approximately symmetric error distribution, 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 estimate, 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 nearly ideal situation, the simulations in Tables 1 and 2 and Figure 3 suggest that the biweight estimate is a slightly better estimator. But these differences are slight so it can be valuable to carefully examine examine both models–one that tries to fit all the data well and another that tries to fit most of the data well.

Quantities of Interest

To estimate the quantity of interest–the effect of social heterogeneity as the permissiveness of the electoral rules varies–we re-estimated 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 using the cluster bootstrap suggested by Harden (2012) (see also Esarey and Menger 2015) to calculate the confidence intervals for each model.

Figure 10 plots the effect of increasing social heterogeneity from ENEP = 1.06 (10th percentile) to ENEP = 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 10], 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 10 shows that simply using the biweight estimator as a robust alternative 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 10 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 9. 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 ENEP = 1.06 (10th percentile) to

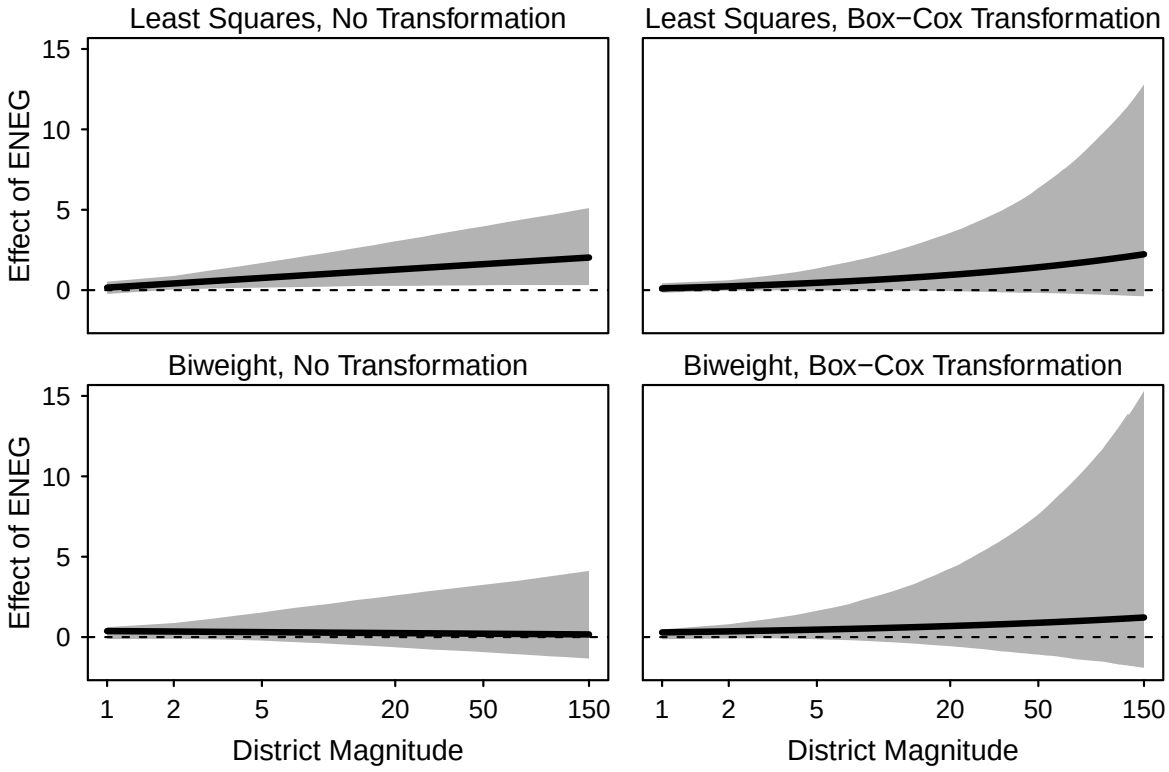

Figure 10: This figures plots the estimated effects and 90% confidence intervals of increasing social heterogeneity from ENEP = 1.06 (10th percentile) to ENEP = 2.48 (90th percentile) on the effective number of electoral parties as the district magnitude varies. Notice that only the model that does not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

ENEP = 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.[3]

	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: Quantities of interest from least squares and biweight estmates, 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.

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 redescending estimators, such as the biweight estimator, is that these estimators allow unusual cases to stand out. Figure 11 compares the residuals from the least

[3]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. See Appendix A for the details.

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.

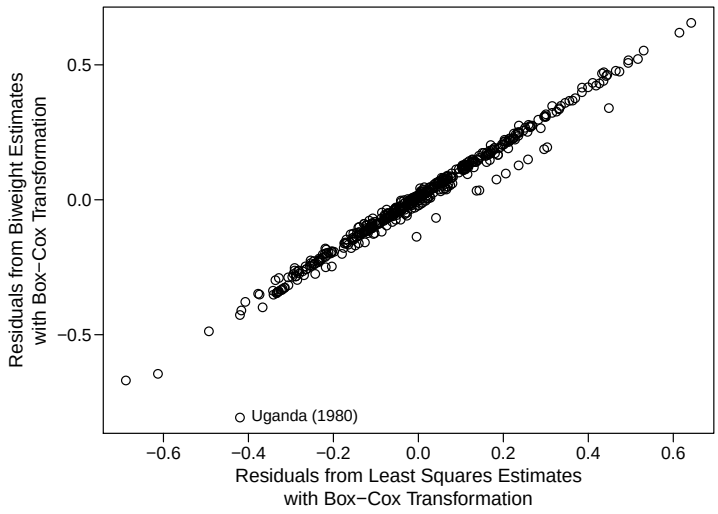

Figure 11: This figure shows the relationship between the least squares and biweight estimates after the Box-Cox transformation. Notice that the biweight estimate allows the unusual case of Uganda to stand out from the others.

Similarly, Figure 12 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 11, 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.

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

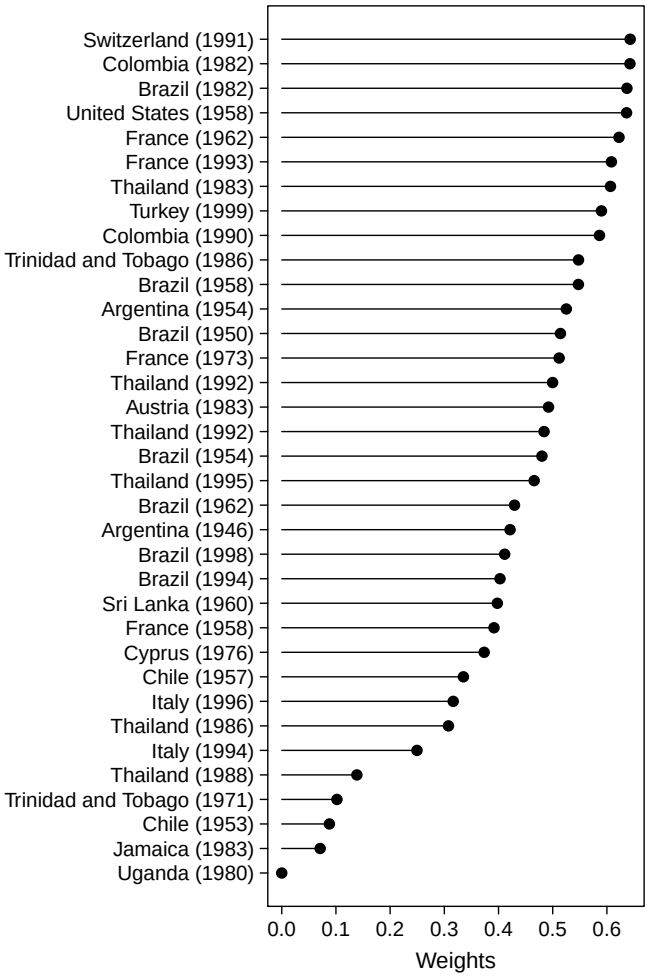

Figure 12: This figure shows the final weights implied by the biweight estimator. Because the 1980 election in Uganda is quite different from the remainder of the data, the biweight estimator downweights this case all the way to zero.

that a careful consideration of the residuals is important. We have shown that there can be tremendous gains in efficiency by 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.

Question about the Validity of the Theory

At first glance, these results might make it seem Duverger's logic lacks the empirical support suggested by the literature. Indeed, the evidence offered by the normal-linear model 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 1954, Riker 1982, Cox 1997, Anorim Neto and Cox 1997, Cox 1999). Second, many empirical studies beyond Clark and Golder (2006) find substantial empirical support for the hypotheses, including observational (e.g., Chibber and Coleman 1998 and Singer and Stephenson 2009), quasi-experimental(e.g., Blais et al. 2009 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. 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 12 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. In 1971, Obote was ousted by Major General Idi Amin.

During Amin's rule, between 80,000 and 300,000 civilians were killed. With the support of the Tanzanian military, Amin's bloody rule ended in 1979. In the following year, one leader was forcibly removed from office and the following government experience yet another military coup and elections were scheduled in 1980. In anticipation of these elections, Milton Obote returned to Uganda. In an elections that were with corruption and irregularity, Obote's party won a majority of the seats. 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 election 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. One common measure offered by Cherub, Gandhi, and Vreeland (2010), does indeed code Uganda as a democracy in 1980.[4] They code countries as a democracy if and only if it meets all 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 author indeed code Uganda in 1980 as a democracy, they consider this condition met. However, this was a single election, nested between two military dictatorships,

[4]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.

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.

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, according to Clark and Golder's define "established democracies" as countries that transitioned to democracy after 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 be the case that countries transitioning to democracy before 1989 (e.g., Uganda in 1980) are "established" immediately after transitioning to democracy. 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 indicated when systems (are likely to) reach an equilibrium.

Questions about Dynamics Prior to Equilibrium

We need a stronger theory about the dynamics with which systems reach 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 to reaching an equilibrium is crucial to measuring whether a party system has reach equilibrium, 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 many and the number grows until equilibrium. Perhaps it depends on the context of the election. Some work has been done in this area (e.g., Moser 1999, Crisp, Olivella, and Potter 2012, and Ferrara 2011), but expanded the theoretical and empirical research in this area is both crucial in its on right but also for studying the number of parties in equilibrium.

The fact that the 1980 election in Uganda featured essentially two political parties is quite interested. 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 majority 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 prior before he seized power, had returned. The people were quite familiar with Obote. Perhaps this enables groups to coordinate in support of or in oppositions to Obote. It is difficult to know without a thorough analysis, but this analysis is suggestive 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. One the other hand, when the prior regime does not participate in the elections, perhaps this makes coordination difficult. The 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 allows the analysis to raise these types of questions.

Conclusion

I might add a little more to the introduction here.

The convention among substantive scholar in political science is to take least squares estimates at face value without thinking carefully about the distribution of the errors. Indeed, the Gauss-Markov theorem encourages a lack of attention to residuals by claiming that least squares is the *best* linear unbiased estimator. However, the restriction to linear estimators is unnecessary and counter-productive. By taking advantage of transformations to make the errors more consistent with the normal-linear model (approximately normal or at least symmetric) and using robust estimators if necessary to deal with remaining heavy tails, we show that researchers can reach different substantive conclusions. Most importantly, though, paying close attention to residuals can raise important questions about concepts, theories, and measurements that might otherwise be overlooked.

Appendix

When BLUE Is Not Best

A Coefficient Estimates

Table A1 presents the coefficient estimate and confidence intervals, creating using the cluster-robust bootstrap recommend by Harden and Desmarais (????). The coefficients for the transformed and untransformed outcomes are not comparable, since the outcome is rescaled. The first column of Table A1 uses Clark and Golder's approach (except for the bootstrapped confidence intervals). Notice that the interaction term ENEG $\times$ log(Magnitude) is large (though not quite statistically significant) when using the least squares estimator, but much smaller when using the biweight estimator.

	Least Squares w/ No Transformation	Biweight w/ No Transformation	Least Squares w/ Box-Cox Transformation	Biweight w/Box-Cox Transformation
Constant	1.916*	1.460*	0.866*	0.810*
	[1.309; 2.580]	[0.989; 2.246]	[0.739; 0.986]	[0.699; 0.987]
ENEG	0.112	0.264	0.018	0.049
	[−0.169; 0.373]	[−0.085; 0.435]	[−0.028; 0.073]	[−0.022; 0.085]
log(Magnitude)	0.078	0.494	0.046	0.075
	[−0.505; 0.435]	[−0.235; 0.840]	[−0.037; 0.113]	[−0.031; 0.162]
Upper-Tier Seats	−0.057	−0.060	−0.004	−0.004
	[−0.108; 0.054]	[−0.103; 0.034]	[−0.012; 0.009]	[−0.013; 0.008]
Presidential Candidates	0.264	0.167	0.025	0.028
	[−0.016; 0.514]	[−0.296; 0.365]	[−0.022; 0.061]	[−0.062; 0.062]
Proximity	−3.098*	−2.884*	−0.523*	−0.536*
	[−3.932; −2.255]	[−3.758; −2.051]	[−0.665; −0.379]	[−0.698; −0.375]
ENEG × log(Magnitude)	0.264*	−0.029	0.025	0.007
	[0.015; 0.708]	[−0.257; 0.549]	[−0.013; 0.088]	[−0.043; 0.089]
ENEG × Upper-Tier Seats	0.059	0.062	0.006	0.006
	[−0.029; 0.092]	[−0.010; 0.088]	[−0.003; 0.011]	[−0.003; 0.012]
Presidential Candidates × Proximity	0.683*	0.711*	0.131*	0.128*
	[0.329; 1.105]	[0.388; 1.315]	[0.069; 0.207]	[0.068; 0.246]
Num. obs.	487	487	487	487

* 0 outside the confidence interval

Table A1: Coefficients estimated using least squares and biweight estimators, with and without the Box-Cox transformation. The brackets contain the 90% confidence intervals estimated with the cluster-bootstrap recommended by Harden and Desmarias (2012). The first column replicates Clark and Golder's (2006) Table 2, column 6, except using cluster-bootstrap confidence intervals. The remaining models make an attempt to address the skewednd/or heavy-tailed residuals. Simply transforming the outcome to be more consistent with the assumed model and/or using the more robust biweight estimator substantial reduces the evidence for their hypothesis.