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 regression models on continuous outcomes of interest using least squares. These outcomes include measures of the number of political parties, proportionality of the electoral system, vote share, and others. While it is well known that least-squares estimates are often sensitive to single, influential data point, this knowledge has not led to appropriate practices when using least-squares estimators. We highlight the importance of using more robust estimators (at least as a robustness check) 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 substantially weakens their key conclusions about the conditional relationship between social heterogeneity and electoral rules in influencing the number of political parties.

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 shows the importance of robust estimators.

[Add more introduction later.]

Is a BLUE Estimator the "Best" Estimator?

The linear regression model can be written as $E(y|X) = X\beta + \epsilon$, where 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. Researchers in political science commonly estimate this model with ordinary least squares (OLS) by minimizing the squared residuals, $\hat{\beta}^{OLS} = \arg\min S(b)$, where $S(b) = \sum_{i=1}^{n}(y_i - X_i b)^2$. That is, OLS estimators choose the estimate $\hat{\beta}$ that minimizes the sum of the squared residuals. Under the assumption that the errors ϵ_i follow independent and identical normal distributions with mean zero and unknown variance, the OLS estimator is the uniform minimum variance unbiased estimator (MVUE) (CITE). This is a powerful result. It shows that of all unbiased estimators, least squares is the most efficient when the errors follow a normal distribution.

Even if the errors do not follow independent and identical normal distributions, the Gauss-Markov Theorem guarantees the least-squares estimator is the best (i.e., minimum variance) *linear* unbiased estimator if the errors have mean zero and constant (and finite) variance. 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 under Gauss-Markov to refer to the structure of the model (i.e., "linear in the parameters"), such that $E(y|X) = X\beta$ falls

into the class of "linear" regression models, but $E(y|X) = e^{X\beta}$ does not. Indeed, this is the sense in which we use "linear" in the terms "linear model" and "linear regression." However, the "linear" restriction in the Gauss-Markov Theorem refers to a highly technical statistical criterion that requires that the estimates be a linear function of the outcome variables, 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.[1] In other words, Gauss-Markov does not require a linear *model* of the form $E(y|X) = X\beta$, but it does require a linear estimator 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 $S(b) = \sum_{i=1}^{n} (y_i - X_i b)^2$ w.r.t. b. To do this, we can simply set $\dfrac{\partial S(b)}{\partial b} \equiv 0$ and solve for the vector b. Noting that $\dfrac{\partial S(b)}{\partial b} = \sum_{i=1}^{n} 2(y_i - X_i b)(-X_i) \equiv 0$ implies that $\sum_{i=1}^{n} X_i(y_i - X_i b) \equiv 0$. This is simply a system of $k+1$ *linear* equations $\sum_{i=1}^{n} X_{ij}(y_i - X_i b)$ for $k = \{0, 1, 2, ..., k\}$. Of course, the matrix form $X'(y - Xb) = 0 \Rightarrow (X'X)b = X'y \Rightarrow b = (X'X)^{-1}X'y$ is much more common. 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 unnecessary and unproductive. Note that we are not arguing against linear models (i.e., models that are linear in the parameters), such as

$$y_i = \beta_0 + \beta_1 \sqrt{(x_i)} + \epsilon_i,$$

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

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

Indeed, each of these is a linear model and illustrate that the linear model can represent a wide range of theoretically-relevant relationships, especially when it includes explanatory variables non-linearly. However, there is no statistical reason to restrict ourselves to linear *estimators*

---

[1] Formally, linearity requires that $\hat{\beta} = My$, where M depends on the matrix X. For the case of least-squares, $M = (X'X)^{-1}X'$.

for these models, except for mathematical convenience and computational ease, and there are substantive reasons to reject this restriction.

Indeed, 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 like estimates that provide an excellent fit to *most* of the data rather and a poor fit to some of the data, 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. Either of these principles exclude linear estimators.

For example, consider the estimates shown in the left panel of Figure **??**. Which of the two estimates, A or B, best summarizes the relationship between the explanatory and outcome variable? Estimate A fits all the data decently, 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.).
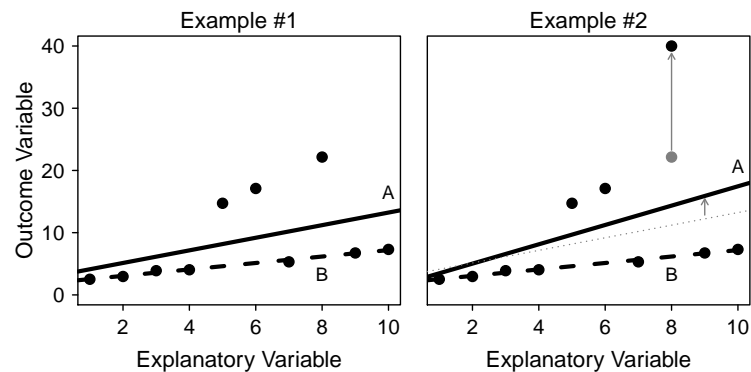

FIGURE 1: Two scatterplots of hypothetical data. The left panel suggests that we might prefer an estimate that provides an excellent fit to most of the data rather than a mediocre fit to all the data. The right panel highlights that we might prefer an estimate that does not respond to movement in points that seem to fall outside the explanatory power of the model. Both the these characteristics violate the assumption of normality.

llabllfig:best-fit-illustration

Secondly, once cases fall outside the explanatory power of the model (i.e., outliers), then we might like our estimators to give them zero weight. After all, these cases seem somehow different and we would not like these unusual data to affect our excellent fit for the majority of the data. That is, if one unusual cases was *even more* unusual, the estimates should not change. Rather than respond to this unusual deviation as does estimate A, we might prefer an estimate that simply ignores this change. After all, the case seems to fall outside the explanatory power of the model in both cases.

Even if one would not like to use an estimator with either of these properties, then consider this example. If the researcher has a substantive 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 effecient (but nonlinear) MLE estimator implied by the assumed t_{10} error 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 MVUE–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 MVUE is easily available for little or no substantive or empirical reason. Knowing that the assumed model (i.e., normality) is *in*orrect, these researchers are using this model as an approximation. But if the model is an approximation, then the desirable statistical properties are not longer guaranteed (e.g., MVUE). With this in mind, it makes more sense to use a robust estimator with the following qualitative properties:

1. Approximately unbiased in typical sample sizes when the errors follow the normal distribution or the deviation from normality is only slight (e.g., $t_1 0$ errors).

2. Good efficiency when the errors follow a normal distribution, though perhaps not as efficient as least squares.

3. Good efficiency when the deviation from normality is slight (e.g., $t_1 0$ errors), perhaps better than least squares.

4. Reasonable efficiency and bias in typical sample sizes with large deviations from normality (e.g., t_2 errors)–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 estimable 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 more efficient (standard deviation about 18% smaller), and has a much smaller mean squared error (about 32% smaller) . By any reasonable standard, it is a better estimator than the least squares estimator. We make this gain by expanding our focus to non-linear estimators. In this case, the LTS estimator is not linear because it places zero weight on the largest 10% of the residuals and weights of one on the smallest 90% of the residuals.
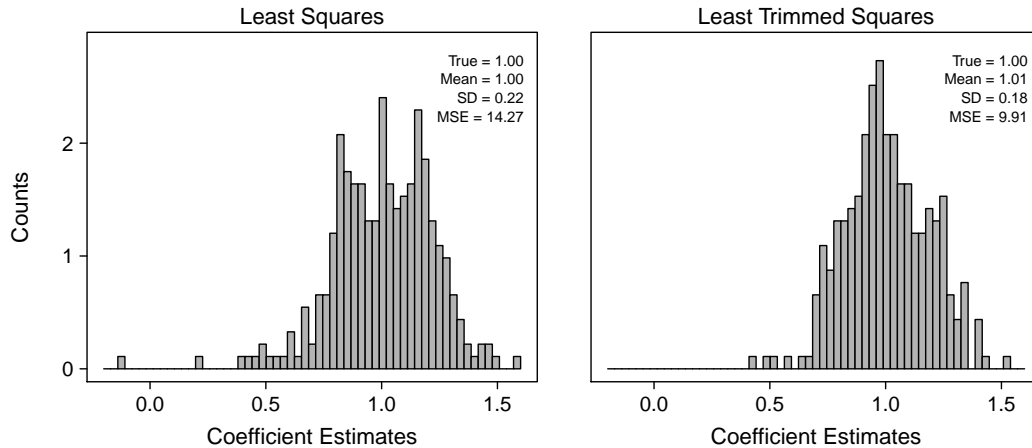
FIGURE 2: caption here

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. For example, 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 political science papers.

On the other hand, there has been scant attention paid by substantive scholars 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's (1995) introduces political scientists to robust estimators, but this work has been essentially ignored. The work is more broadly applicable than Beck and Katz (1995) and was published in the same year, but has received only 99 citations, or about 2% of the citations that Beck and Katz have received. Similarly, Desmarais and Harden (2011) has 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), though, 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

There are two ways in which the errors can deviate from normality. First, the error distribution might be skewed. Second, the error distribution might have heavy-tails. We suggest dealing with these deviations differently so we discuss each separately.

Skewed error distributions create two problems for the linear model. First, least squares models 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, we interpret skewed error distributions as 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. For example, rather than a college degree increasing one expected income by $10,000, perhaps it increases one's

expected income 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 estimated residuals will likely 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. 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. 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 this suggestion, statisticians have created more precise empirical methods for choosing *whether* and *how* to transform the outcome variable y. 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 $\hat{\lambda}$ is near one, this suggests no transformation is needed and if $\hat{\lambda}$ is near zero, then only an intuitive log-transformation is needed.

It is quite easy to 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, one 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, the important point is not to argue for any particular test, but to point out (1) that asymmetries worsen the performance of least squares and alternative methods and (2) this potential problem can be addressed using simple transformations that are easy to implement.

Mean or Median?

Applying 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, these quantities are not as easy to understand substantively. After all, 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 understandable, we simply need to "undo" the transformation. However, it is important to note that $E[\log(y)|X] \neq \log[E(y|X)]$, which means that the log cannot simply be undone (without additional computation) because $e^{E[\log(y)|X]} \neq E(y|X)$.

These interpretational difficulties not due to the choice to transform the data, but imbedded in the data themselves. In the context of skewed distributions, the mean $E(\cdot)$ might be a misleading summary of the "center" of the data. While the mean is often more mathematically convenient, the median offers a better summary of location than the mean. Further, the median has an intuitive interpretation because one-half of the distribution lies above the median and one-half lies below. If we use $\text{med}(y_{new}|X_{new})$ to predict the unknown outcome y_{new} for a known case X_{new}, then there is a 50% chance of being to high and a 50% chance of being to

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 $\frac{\partial\,\text{med}(y|X)}{\partial x_j}$, one gains a more intuitive quantity of interest and can easily move between transformed and untransformed outcomes (e.g., $\text{med}[\log(y)] \to \text{med}(y)$). This holds for the more general case of $y^{(\lambda)}$ as well.

Simulating Quantities of Interest Under Transformation

To obtain quantities of interest for $\text{med}(y)$ 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 model $y = X\beta + \epsilon$ using least squares to obtain the estimated model coefficients $\hat{\beta}^{ls}$ and residuals $\hat{e}^{ls}$. Evaluate whether the assumption of the normal errors matches the estimated residuals using histograms and QQ plots.

2. If needed, 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.)

3. Estimate the transformed model $y^{(\lambda)} = X\beta + \epsilon$ using least squares to obtain the estimated model coefficients $\hat{\beta}^{ls}_{trans}$, residuals $\hat{e}^{ls}_{trans}$, and covariance matrix Σ^{ls}_{trans}.

4. 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 $\Delta(y, X_{hi}, X_{lo}) = \text{med}(y|X_{hi}) - \text{med}(y|X_{lo})$.

5. Following King, Tomz, and Wittenberg (2000), for i from one to a large (e.g., 1,000) number of iterations n_{sims}:

   a. Simulate $\tilde{\beta}^{ls}_{trans} \sim N\left(\hat{\beta}^{ls}_{trans}, \Sigma^{ls}_{trans}\right)$.

b. Calculate and store $\tilde{Q}_i = \text{med}(y|X_{red}, \tilde{\beta}^{ls}_{trans}) = X_{red}\tilde{\beta}^{ls}_{trans}$ or, if interested in the first-difference, $\tilde{Q}_i = \tilde{\Delta}(y, X_{hi}, X_{lo}, \tilde{\beta}^{ls}_{trans}) = X_{hi}\tilde{\beta}^{ls}_{trans} - X_{lo}\tilde{\beta}^{ls}_{trans}$.

6. 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. [Add some more examples of robust estimators here.]

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-estimation 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 that such $\rho(\cdot) = abs(\cdot)$. However, other estimators offer similarly robust alternatives. In particular, we recommend Tukey's biweight function, so that

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

12

.

where $r_i = y_i - X_i b$. 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. Second, the optimization problem is not convex, so standard optimization routines can produce a local rather than a global minimum. However, this problem can usually be handled 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 $\hat{\sigma}_{mad} = \dfrac{\text{med} \left(|y - X \underline{|} \right)}{0.6745}$. Dividing by 0.6745 makes $\hat{\sigma}_{mad}$ a consistent estimator of the standard deviation of normal errors.

While the theory for M-estimators is not as elegant as for least squares estimators, M-estimators do have desirable properties. In general, M-estimators are consistent as long as (1) $\rho(\cdot)$ is convex and (2) the distribution of the errors is strongly unimodal (decreasing away from zero). Because the biweight objective function is non-convex, we must assume that the distribution of the errors is strongly unimodal. Under the assumption of strong unimodality, estimates are distributed asymptotically normal and consistent.

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. Unusual cases can be extremely informative. 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 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 $\hat{\sigma}_{mad}^{(0)} = \dfrac{\text{med}\left(|y - X\hat{\beta}^{(0)}|\right)}{0.6745}$.

3. ssign weights w according to the function ρ and denote $\text{diag}(w) = W$.

4. For i from one until convergence:

   a. Using $\hat{\beta}^{(i-1)}$ and $\hat{\sigma}_{mad}^{(i-1)}$ 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 $\hat{\sigma}_{mad}^{(i)} = \dfrac{\text{med}\left(|y - X\hat{\beta}^{(0)}|\right)}{0.6745}$

   d. The algorithm has converged when $r^{(i-1)} \approx r^{(i)}$.

If we assume the distribution of the errors to be symmetric 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. Only if $\rho(r_i) = r_i^2$ is this estimator a *linear* estimator though. Other choices of $\rho(\cdot)$ might produce estimators that are better than the BLUE estimator. These M-estimators are consistent if (1) $\rho(\cdot)$ is convex *or* (2) the distribution of the errors is strongly unimodal (that is, the density is decreasing away from

14

zero). Because

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

Monte Carlo Simulations

To understand and illustrate how the performance of the BW estimator compares with the LS and LAD estimators, we simulated from the linear model $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i$, where $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = 1$ and the x_i's follow an approximately normal distribution. 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 least absolute deviation 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 produces 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 $t_1 0$ distribution has *slightly* heavier tails than the normal distribution. To get a sense for just how similar a t_{10} distribution is to a normal distribution, note that a Shapiro-Wilk test of normality only successful rejects the null in

	Mean				Standard Deviation				Mean Squared Error			
	Lapl.	t_2	t_{10}	Norm.	Lapl.	t_2	t_{10}	Norm.	Lapl.	t_2	t_{10}	Norm.
Absolute Performance												
Least Squares	0.999	1.000	1.001	0.999	0.151	0.319	0.114	0.099	227.419	1016.182	130.867	98.850
Least Absolute Deviation	0.999	0.999	1.003	1.000	0.126	0.146	0.131	0.124	159.499	212.360	172.911	154.714
Tukey's Biweight	1.000	0.999	1.001	0.999	0.131	0.138	0.113	0.102	172.043	189.519	127.731	104.759
Relative Performance												
LAD/LS	1.001	0.999	1.002	1.001	0.837	0.457	1.149	1.251	0.701	0.209	1.321	1.565
BW/LS	1.001	1.000	1.000	1.000	0.870	0.432	0.988	1.029	0.757	0.187	0.976	1.060

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				Standard Deviation				Mean Squared Error			
	Lapl.	t_2	t_{10}	Norm.	Lapl.	t_2	t_{10}	Norm.	Lapl.	t_2	t_{10}	Norm.
Absolute Performance												
Least Squares	1.000	1.001	1.000	1.000	0.044	0.116	0.035	0.032	19.057	135.301	12.520	10.488
Least Absolute Deviation	1.000	0.999	1.000	1.000	0.033	0.044	0.041	0.041	10.806	19.757	16.472	16.412
Tukey's Biweight	1.000	0.999	1.000	1.000	0.037	0.043	0.035	0.033	13.836	18.321	12.025	11.148
Relative Performance												
LAD/LS	1.000	0.998	1.000	1.000	0.753	0.382	1.147	1.251	0.567	0.146	1.316	1.565
BW/LS	1.000	0.999	1.000	1.000	0.852	0.368	0.980	1.031	0.726	0.135	0.960	1.063

TABLE 2: Summary of 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.

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, in which case the t_{10} has *slightly* heavier tails.

- *Normal Distribution.* The normal distribution yields the optimal conditions for the least squares estimator. Indeed, when the errors follow a normal distribution, the least squares estimator is 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, the standard deviation, and the mean squared error. Table **??** provides the results for the sample size of 100 and **??** provides the results for the sample size of 1,000.

---

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

These result 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 mean squared error about 30% lower than the LS estimate. However, the BW estimator also performs quite well for the Laplace distribution, with a mean squared error about 25% less than the least squares estimator for $N = 100$. For $N = 1,000$ the MSE of the LAD estimator outperforms the LS estimator by about 43% and the BW estimator outperforms the LS estimator by about 27%.

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

However, 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 of about 34% more than the LS estimator for $N = 100$ and about 32% for $N = 1,000$. The BW estimator on the other hand, shows a small *improvement* over the LS estimator, with an MSE of 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 LS estimator is no longer the best estimator.

The normal distribution is the optimal scenario for the least squares estimator and it outperforms the the LAD estimator considerably with the errors are normal. For both $N = 100$ and $N = 1000$, 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. Indeed, the MSE for the BW estimator is only about 6% larger than the MSE for the LS estimator for both

$N = 100$ and $N = 1000$.

Notice that, 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 $t_1 0$ distributions and considerably outperforms the LS estimator for Laplace and t_2 errors and considerably outperforms 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. 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 performance 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., df of 2 to 4), but as the tails grow lighter, the LS estimator quickly begins to outperform the LAD estimator. Indeed, expect of all but very heavy tailed distributions, the LS estimator is considerably more efficient than the LAD estimator.

The BW estimator, though, is a much stronger competitor for the LS estimator. While the LS estimator is more efficient for lighter-tailed distributions (i.e., $df > 10$), the difference is very small except in very small samples. Indeed, for sample sizes of 100 or larger, the LS estimator is about 5% more efficient than the BW estimator *at best*. This simulation again 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.

FIGURE 3: caption here

Recommendation for Applied Researchers

When using the linear model, we suggest that researchers take several steps to ensure that their 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 perhaps if not), carefully examine the residuals. Be careful to check for skewness.

4. If the residuals are not symmetrically distributed, then consider a transformation. The log-transformation has a nice substantive interpretation, so it makes sense as a first cut. If the log-transformation over- or under-corrects the skew, then the Box-Cox transformation parameter should ensure that the residuals are roughly symmetric. This transformation allows the model to represent non-linear relationships implied by the skewness and allows the statistical model to more closely approximate the data.

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.

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. Clark and Golder 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 $\text{ENEP} = \sum_{i=1}^{n} \frac{1}{v_i^2}$, 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.
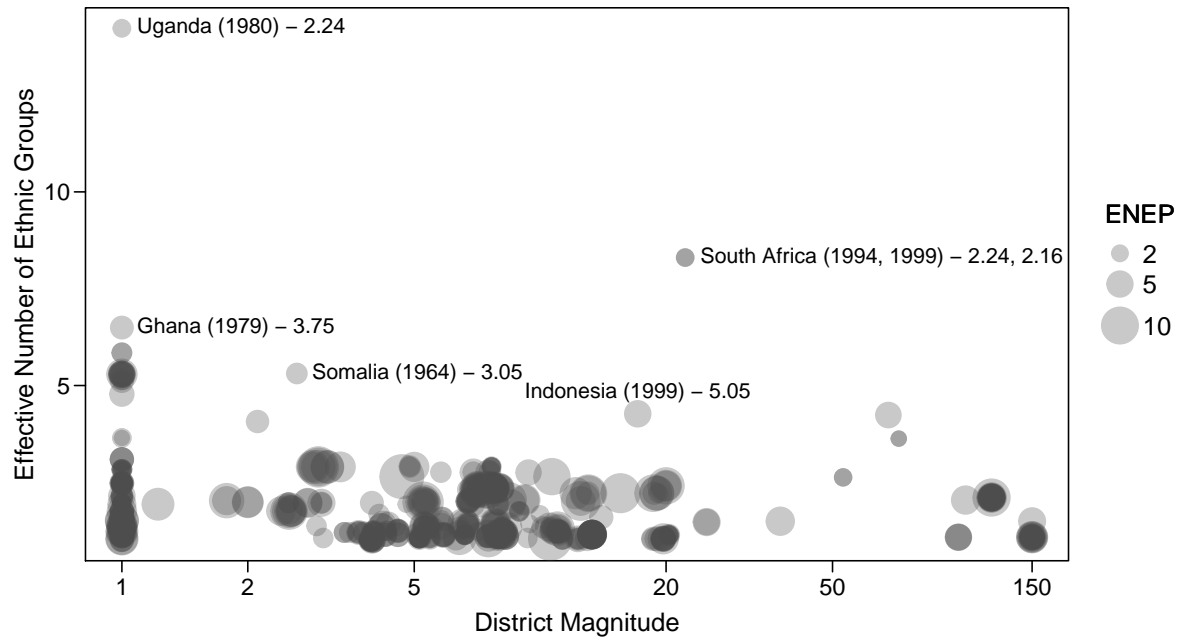

FIGURE 4: This figure shows the distribution of district magnitude (on the log scale) and ethnic heterogeneity, where the point sizes indicate the number of political parties. According to Clark and Golder's hypothesis, large points should lie in the upper-right portion of the plot and small points should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

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

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

The first key coefficient in this analysis is β_1, which summarizes 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.
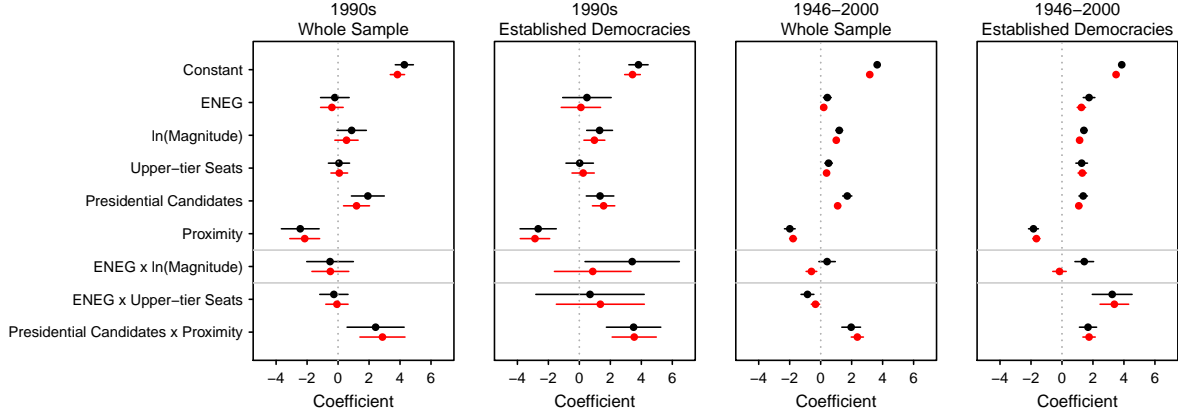

FIGURE 5: Replication of Clark and Golder (2006) using the biweight estimator with explanatory variables standardized to have mean zero and standard deviation one-half. The black lines and points show the least squares estimates and 90% confidence intervals and the red lines and points show the biweight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

Notice that 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 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. In our view, the model for the transformed outcome can capture potentially interesting substantive effects as well.
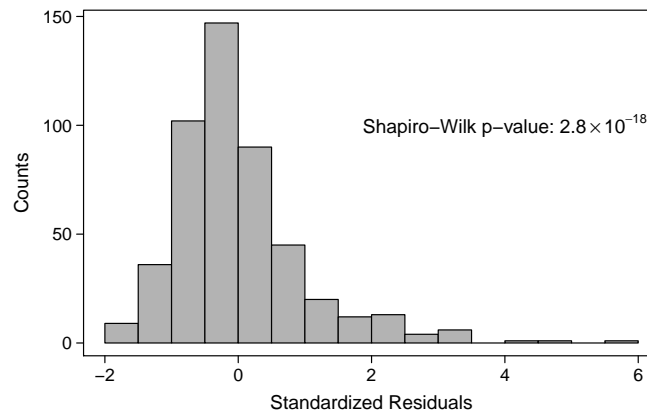

FIGURE 6: A histogram showing the distribution of the residuals from Clark and Golder's (2006) main model. Notice that these residuals do not seem approximately normal. They have a strong skew and heavy tail to the right. For example, one would rarely expect to observe residuals more that three standard 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.
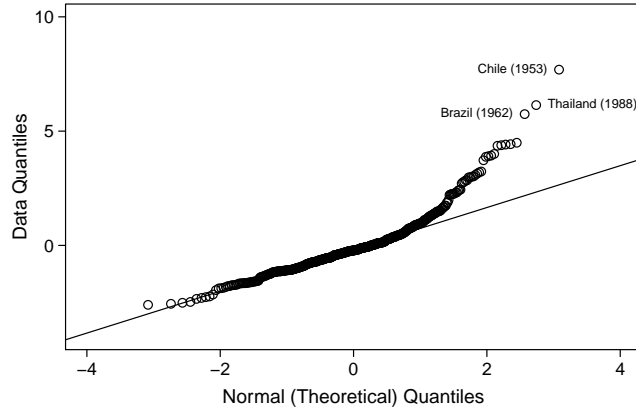
FIGURE 7: A QQ plot showing the deviation of the residuals from normality. If the residuals were approximately normal, then the points in the QQ plot would approximately follow the line. However, notice that the positive residuals deviate sharply from the theoretical expectations. This also suggests that some transformation of the outcome variable might be useful.

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 estimated the model using both a log-transformation and Box-Cox transformation with $\lambda = -\frac{1}{3}$. Figure 8 present the histogram of the residuals from these two regression models. Notice that log-transforming the effective number of electoral parties does not quite remove the skew in the residuals. However, the Box-Cox transformation with $\lambda = -\frac{1}{3}$ provides highly symmetric residuals.
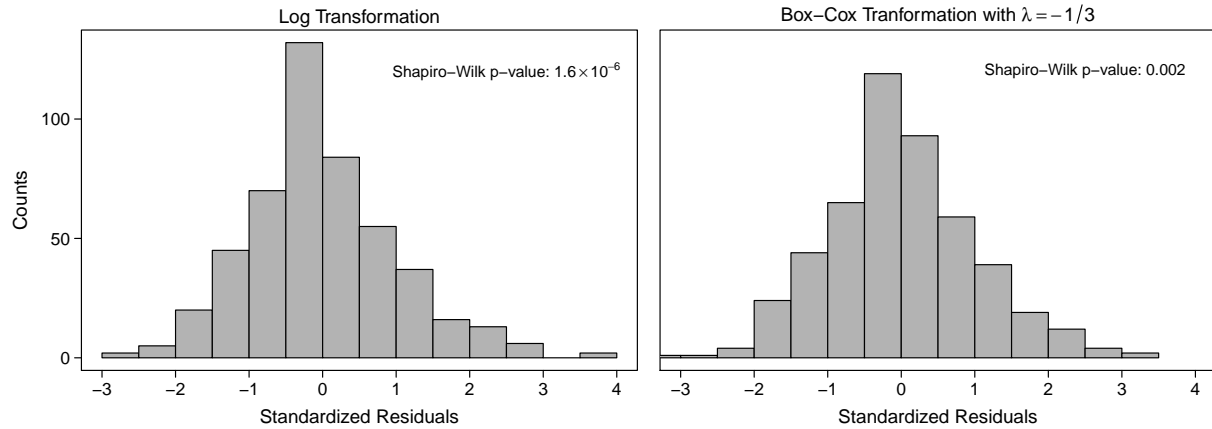
FIGURE 8: Histograms of the residuals after transforming the outcome variable. The left panel shows that the log transformation does not quite remove all of the skew, but the right panel shows that the Box-Cox transformation with $\lambda = -\frac{1}{3}$ creates an approximately symmetric error distribution.

Figure 9 shows the QQ plot for these 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.
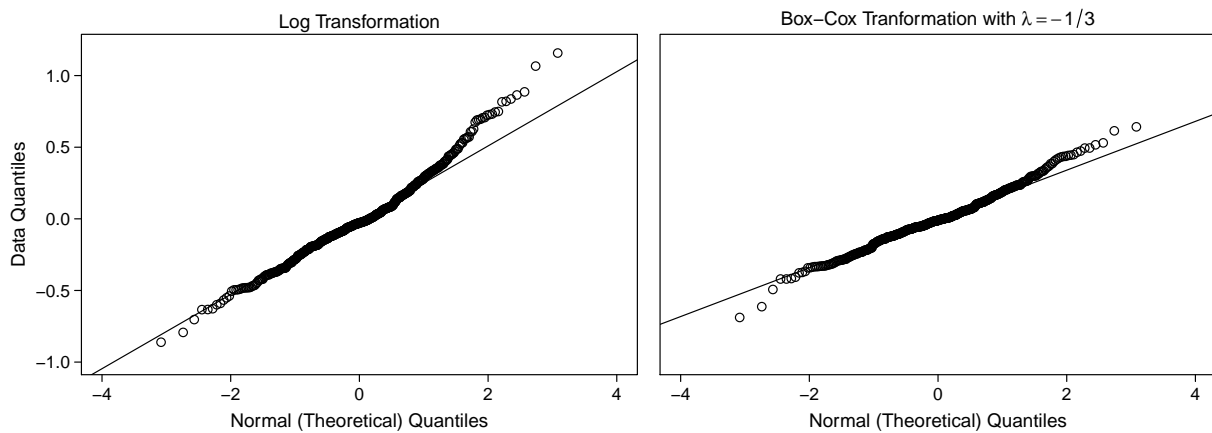

FIGURE 9: QQ plots of the residuals after transforming the outcome variable. The left panel shows that the log transformation does not quite remove all of the skew, but the right panel shows that the Box-Cox transformation with $\lambda = -\frac{1}{3}$ creates an approximately symmetric error distribution, 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. Notice that the data tend to have larger positive and larger negative quantiles than one would expect under a normal distribution. This suggests the residuals have a heavy tail, perhaps resembling a t distribution with degrees of freedom in the range of six to twelve. This also suggests that least squares might not be the most efficient estimator, even after transformation. Indeed, 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. This 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 residuals might actually have heavier tails than this analysis suggests.

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. However, 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) and 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. Note the the effect of district magnitude is nonlinear because it is logged in the regression model but unlogged along the horizontal axis. The upper-left figure replicates Clark and Golder's approach (except of the cluster-bootstrap confidence intervals) and replicates their finding, which they summarize:

> [These results] clearly illustrate that in established democracies, ethnic hetero-geneity 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 account 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 reduces the estimated effect across the range of district magnitude to a statistically insignificant, negligible effect. This figure shows the tremendous gains in efficiency when the errors are approximately normal as opposed to skewed and/or heavy-tailed

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. Notice that this approach reduces effect of social heterogeneity to nearly zero and suggests that these data offer very

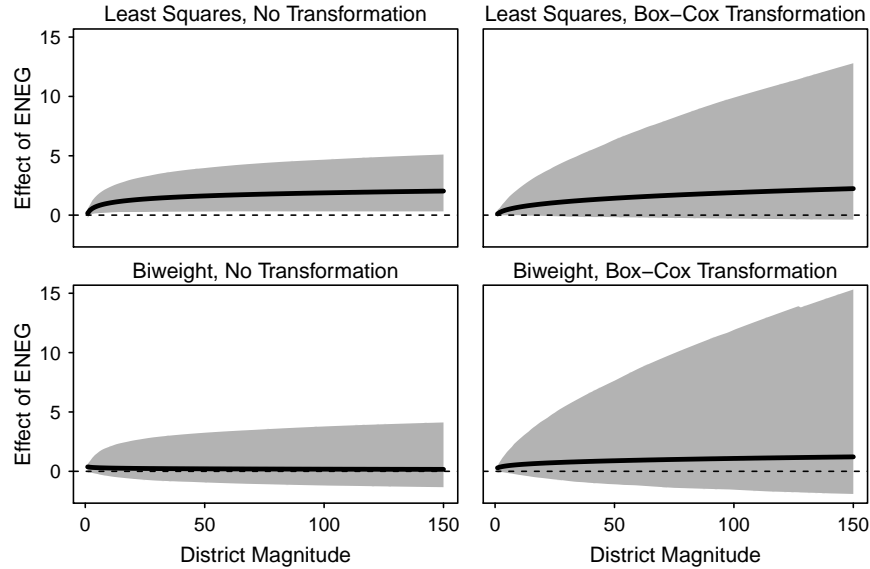little evidence for interaction consistent with the hypothesis.

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.

Because the scale of the estimates differ drastically across the choice to transform the outcome variable or not, 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 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 suggest 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.

	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, using the robust biweight estimator, transforming the outcome variable, or both substantially reduce the amount of evidence that these data offer in favor of the hypothesis. In particular, the biweight estimator suggests that social heterogeneity has small, substantively insignificant effects regardless of the electoral rules. Indeed, the effect is, *at most*, about 1/5 of of a political party per ethnic group, with a estimate of about 1/20 of a political party per ethnic group.

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 substantial interaction simply disappears. These estimates suggest that, in single-member districts, a substantial increase in social heterogeneity increases the median ENEP by about 0.08 [-0.04; 0.15] parties–about half the initial estimate of Clark and Golder. In large, multimember districts (magnitude of 14), the effect again shrinks to 0.08 [-0.06; 0.28]–about 7% of Clark and Golder's initial estimate. This leads to an estimated increase of 0.00 [-0.17; 0.25] in the effect of social heterogeneity as we move from single-member districts to large, multimember districts–an estimate of exactly no interaction.

31

Clark and Golder adopt the convention of many researchers in political science–taking least squares estimates at face value without thinking carefully about the distribution of the errors. Indeed, the Gauss-Markov theorem encourages this convention by noting 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 we reach different substantive conclusions.

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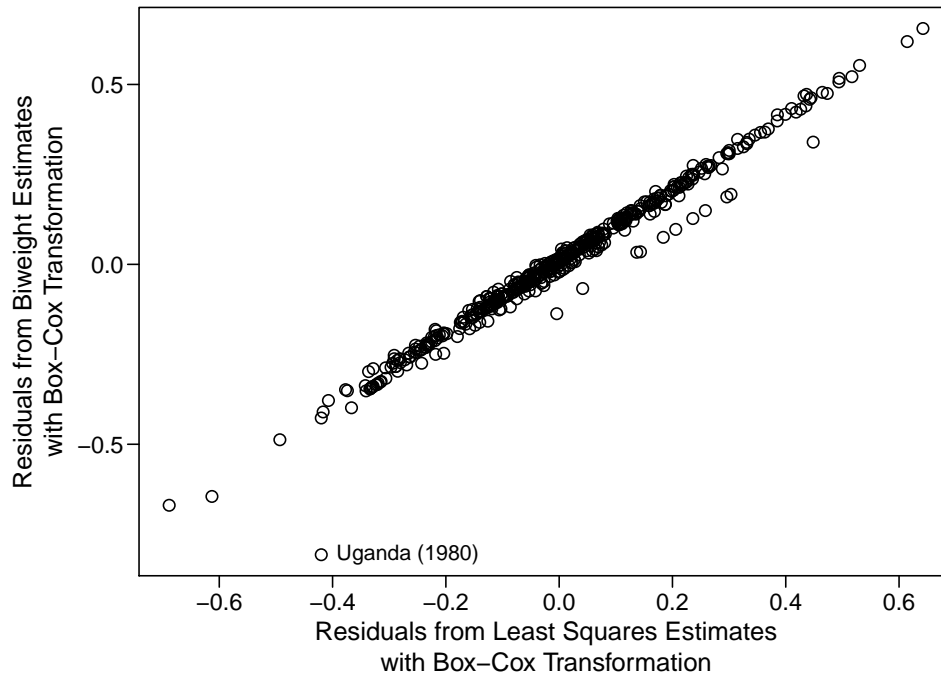
FIGURE 11: This figure shows the relationship between the least squares and biweight estimates after the Box-Cox transformation. Notice that the biweight estimate allows the unusual case of Uganda to stand out from the others.

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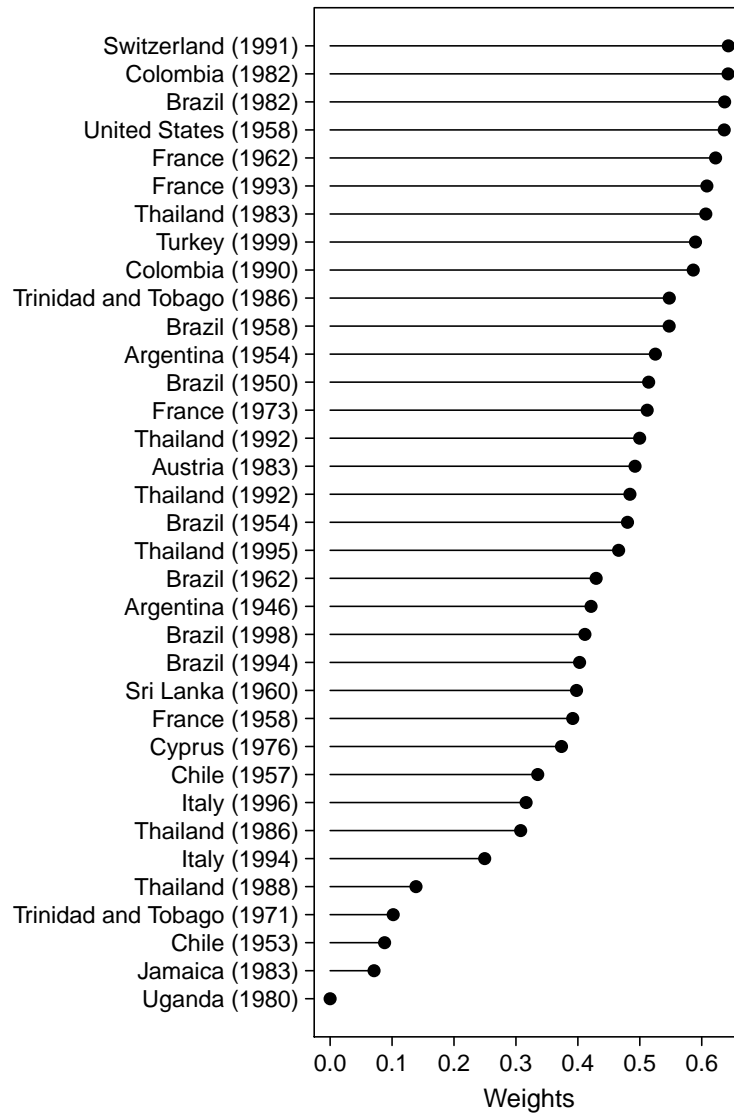
FIGURE 12: This figure shows the final weights implied by the biweight estimator. Because the 1980 election in Uganda is quite different from the remainder of the data, the biweight estimator downweights this case all the way to zero.

Some Implications of the Analysis of the Residuals

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 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 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.

Evidence for the Theory

Perhaps Duverger's logic is not as well-supported empirically as the literature would suggest. Indeed, the evidence offered by the normal linear model seems to hinge on a few cases that seem inconsistent with the majority of the data. However, rejection of Duverger's theory or Clark and Golder's (2006) analysis is premature for three reasons. First, the theoretical logic for Duverger's hypotheses is clear and compelling (CITES). Second, many studies beyond Clark and Golder (2006), including experimental work, find substantial empirical support for the hypotheses. Finally, the biweight M-estimator suggests several potential shortcomings in terms of theory and measures that might currently undermine the evidence for these theories.

Questions about Measuring "Established Democracies"

Room for improvement exists in the measurement of "established democracies." Consider, just to take one example, Uganda in 1980. 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 event a democracy. One common measure offered by Cherub, Gandhi, and Vreeland (2010), does indeed code Uganda as a democracy in 1980. 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, 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. For example, it surely must be the case that countries transitioning to democracy before 1989 (e.g., Uganda) 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.

Questions About the Impact of a Prior Authoritarian Regime

We need additional research into how systems respond to the introduction of additional parties into formerly authoritarian, single-party systems. 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.

Appendix

The Heavy Tails of Electoral Data

A