The Heavy Tails of Electoral Data

The Importance of Robust Estimators[*]

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, and vote share, among 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 important 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 their conclusions depend on several influential data points. Removing these data or using a robust estimator substantially weaken 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/meaningful-inferences .

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

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

Introduction

Our goal 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.

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 minimum variance unbiased estimator (MVUE).

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. However, this should provide little comfort to researchers because their 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, 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 term "linear regression." However, the "linear" restriction in the Gauss-Markov Theorem refers to a highly technical and obscure statistical criterion that requires that the estimates be a linear function of the outcome variables, so that $\hat{\beta}_j = \lambda_1 y_1 + \lambda_2 y_2 + ...\lambda_n y_n$, so that 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 $\frac{\partial S(b)}{\partial b} \equiv 0$ and solve for the vector b. Noting that $\frac{\partial S(b)}{\partial b} = \sum_{i=1}^{n} 2(y_i - X_i b)(-X_i) \equiv 0$ 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

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

$X'(y - Xb) = 0 \Rightarrow (X'X)b = X'y \Rightarrow b = (X'X)^{-1}X'y$ 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, we argue that 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 $y_i = \beta_0 + \beta_1\sqrt{(}x_i) + \epsilon_i$, $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, or $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i$. Indeed, 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*, except for mathematical convenience, and there are substantive reasons to reject this restriction. For example, if the researcher is aware that one case has a unusually large outcome variable (conditional on the explanatory variables), then the researcher might wish to weight that case less than the other, more typical cases so that one atypical case does not exert several times more impact on the estimates than other, typical cases. Indeed, substantive researchers might wish to attach zero weight to extremely unusual cases because these cases might be due to a different substantive process.

A Substantive Argument Against Linear Estimators

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 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 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.).

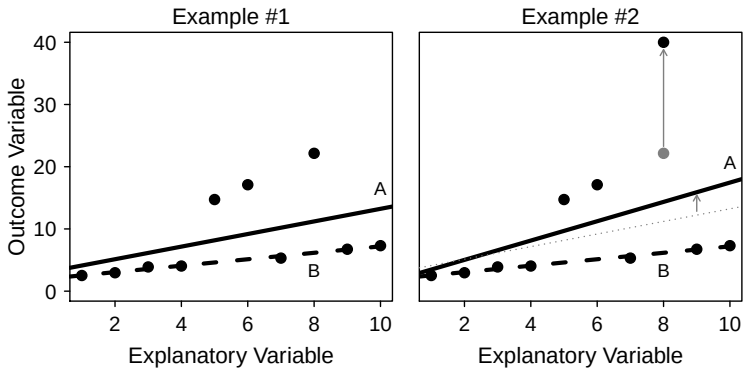

FIGURE 1: caption here

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.

It is not often appreciated that if the errors do not follow independent and identical normal distributions, then the OLS is no longer the MVUE–other estimators might outperform OLS.

Many researchers simply assume a statistical model for which a MVUE is easily available for little or no substantive reason. Knowing that the assumed model (e.g., 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 for the assumed model and small, plausible deviations.
2. Excellent efficiency under the assumed model, though perhaps not the best possible efficiency.
3. Excellent efficiency under small deviations from the assumed model
4. Reasonable efficiency and bias in typical sample sizes with large deviations from the assumed model.

Mathematically, this suggests that applied researchers should not necessarily restrict themselves to unbiased estimators or minimum variance estimators under an assumed model. Instead, a more desirable criterion might be the mean squared error of the estimate under a wide range of deviations from the assumed model. 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 slightly 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. 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 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 estimate, 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. This improvement is dropped 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.

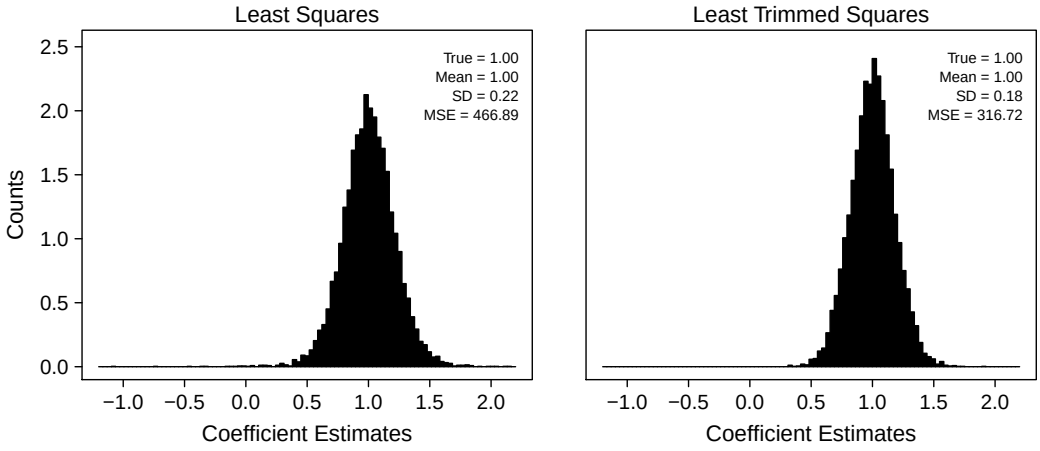

FIGURE 2: caption here

While the statistical properties of the least squares under the assumed normal-linear model are extremely well developed (e.g., MVUE), these properties are not nearly as well developed from robust alternatives. Some asymptotic results are available, but the closed-form theory is generally much weaker.

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) introduction to panel corrected standard errors has received over 4,300 citations, making it one of the most cited political science papers ever.

On the other hand, there has been scant attention paid 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. 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, Demaris 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].

Another popular textbook, Wooldridge (????) does briefly discuss robust estimators, though in a much more limited manner than his discussion of robust standard errors. Angrist and Pischke (2008), though, devote an entire chapter to robust standard errors and completely ignore robust estimation of model coefficients.

Dealing with Skewness: Transforming the Outcome

Skewed error distribution create two problems for the linear model. First, because 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 take skewed error distributions as a lack of model fit. For example, it is 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%. We consider skewed error distributions as evidence that the basic structure of the model could be improved.

Points that are unusual prior to transformation might be quite typical after transformation.

Even if we remain indifferent toward the theoretical implications of skewed error distributions, we must remain cautious about the statistical implications. The performance of least squares estimators improves as the error distribution approaches a normal distribution. Further, the theoretical properties of alternative, more robust estimators, such as the M-estimator we discuss below, depend on a symmetric error distribution distribution.

It is quite common in disciplines such as economics, for example, to log-transform non-negative outcome variables by default. The motivation is that 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.

While we agree with the spirit of this suggestion, we have much more precise empirical methods for choosing *whether* and *how* to transform the outcome variable y. Box and Cox (1964) proposed the Box-Cox transformation

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

.

In this case, the model becomes $y^{(\lambda)} = X\beta + \epsilon$. The researcher can easily use maximum likelihood to obtain estimates $\hat{\lambda}$ and standard errors for the transformation parameter λ. 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 and should be addressed using simple transformations that are easy to implement.

However, 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 $E(y|X)$ 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–$\frac{\partial E[\log(y)|X]}{\partial x_j}$ is 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 $e^{E[\log(y)|X]} \neq E(y|X)$.

However, in the context of skewed distributions, the mean $E(\dot{)}$ 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 there is a one-half of the distribution lies above the median and one-half lies below. This mean that there is approximately a one-half chance that $y|X$ falls above $\text{med}(y|X)$ and a one-half changes that $y|X$ falls below.

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., $\text{med}[\log(y)] \rightarrow \text{med}(y)$). This holds for the more general case of $y^{(\lambda)}$ in addition to $\log(y)$.

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 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 these 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.

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 $\hat{\beta}^{\rho} = \arg\min_b \sum_{i=1}^{n} \rho(y_i - X_i b)$. For example, Harden and Desmarais (2011) recommend the least absolute deviation (LAD) estimator that chooses $\rho(\dot{}) = abs(\dot{})$. However, other estimators offer similarly robust alternatives. 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 $r_i = y_i - X_i b$.

[Discuss the desirable substantive properties of bw, such as fitting the majority of the data well and ignoring, rather than simply down weighting unusual data.]

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. 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 $\hat{\sigma}_{mad} = \frac{\text{med}\left(|y - X\underline{\ }|\right)}{0.6745}$. Dividing by 0.6745 makes $\hat{\sigma}_{mad}$ a consistent estimator of the standard deviation of normal errors.

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 $\hat{\sigma}^{(0)}_{mad} = \dfrac{\text{med}\left(|y - X\hat{\beta}^{(0)}|\right)}{0.6745}$.
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 $\hat{\sigma}^{(i-1)}_{mad}$ 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 $\hat{\sigma}^{(i)}_{mad} = \dfrac{\text{med}\left(|y - X\hat{\beta}^{(0)}|\right)}{0.6745}$
   d. The algorithm has converged when $r^{(i-1)} \approx r^{(i)}$.

[Discuss some properties of M-estimators, unbiased for symmetric error distributions, consistent.]

[Discuss standard errors for m-estimators.]

Monte Carlo Simulations

To understand and illustrate how the performance of the biweight M-estimator, 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_10 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 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.

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

	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	1.001	1.010	0.999	1.000	0.145	0.567	0.113	0.112	210.907	3218.668	128.015	126.091
Least Absolute Deviation	1.001	0.999	0.999	0.999	0.121	0.157	0.131	0.141	146.219	247.484	170.921	197.565
Tukey's Biweight	1.001	0.999	0.998	0.999	0.126	0.149	0.112	0.116	159.560	221.130	124.870	133.664
Relative Performance												
LAD/LS	1.000	0.990	1.000	1.000	0.833	0.277	1.155	1.252	0.693	0.077	1.335	1.567
BW/LS	1.000	0.989	0.999	1.000	0.870	0.262	0.988	1.030	0.757	0.069	0.975	1.060

TABLE 1: Summarizes of the Monte Carlo simulations for four different error distributions with a sample size of 100.

	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 Table 1, except with a sample size of 1,000 rather than 100.

- *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 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 least squares estimator, the median absolute deviation estimator, and the biweight 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.

These result show that all three estimators are essentially unbiased regardless of the error distribution and sample size. Efficiency, however, varies considerably across the estimators.

The MAD estimator is the MLE when the errors follow a Laplace distribution, so as we might expect, the MAD performs well for Laplace errors, with a mean squared error about 30% lower than the least squares error distribution. However, the biweight 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 MAD estimator outperforms the LS estimator by about 43% and the biweight estimator outperforms the LS estimator by about 27%.

The t_2 distribution is nearly a worst case for the least squares estimator, so both robust alternatives perform considerably better. The MAD estimator is has a mean squared error of about 92% less than the least squares estimator and the biweight estimator has a MSE about 93% less than the LS estimator for $N = 100$ and 15% and 14% 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., $N = 500$). In this case, the MAD estimator has a MSE of about 34% more than the LS estimator for $N = 100$ and about 32% for $N = 1,000$. The biweight 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 $N = 1,000$.

The normal distribution is the optimal scenario for the least squares estimator and it outperforms the the MAD estimator considerably with the errors are normal. For both $N = 100$ and $N = 1000$, the MSE for the MAD estimator is about 57% larger than the MSE for the LS estimator. However, the biweight estimator performs nearly as well as the LS estimator. Indeed, the MSE for the biweight estimator is only about 6% larger than the MSE for the LS estimator for both $N = 100$ and $N = 1000$.

Notice that while the biweight estimator was not the most efficient estimator for the Laplace and normal distributions, it was a close second. It was the most efficient estimator for the t_2 and t_10 distributions and considerably outperforms the LS estimator for Laplace errors and considerably outperforms the MAD estimator for normal errors. Thus, the biweight estimator works quite well across a range of error distributions, whereas the LS and MAD estimators work well only in particular situations.

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 MAD and biweight estimators relative to the LS estimtator.

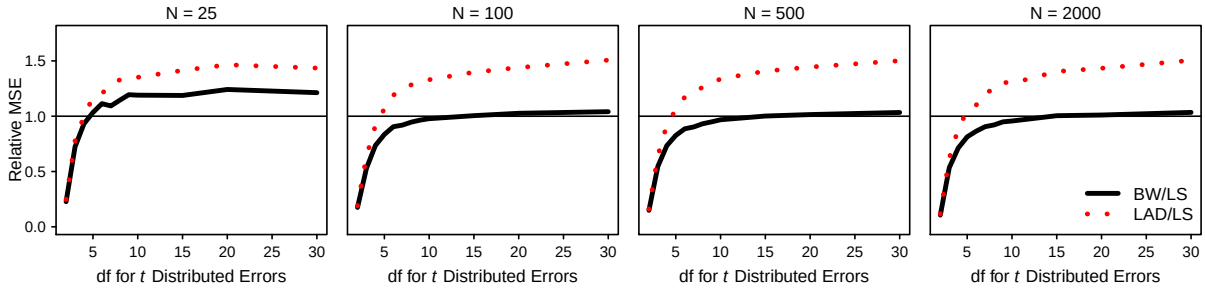

FIGURE 3: caption here

Notice that the MAD and biweight estimators perform quite well for very heavy tailed distribution (i.e., df of 2 to 4), but as the tails grow lighter, the least squares estimator quickly begins to outperform the MAD estimator. Indeed, expect of all but very heavy tailed distributions, the LS estimator is considerably more efficient than the MAD estimator.

The biweight estimator, though, is a much stronger competitor for the least squares 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 *at best*. This simulation again suggests that the biweight 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.

Replication of Clark and Golder (2006)

Clark and Golder (2006, p. 694) 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.

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).

[Summarize G and G's measures of ENEG, ENEP, and Magnitude.]

To test this hypothesis, Clark and Golder fit the following regression 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 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. For these initial estimates, we make no attempt to account for the clustered nature of the data in calculating the standard errors, but do supply the usual 90% confidence intervals to serve as a lower-bound on the uncertainty. Figure 4 presents these estimates and confidence intervals.

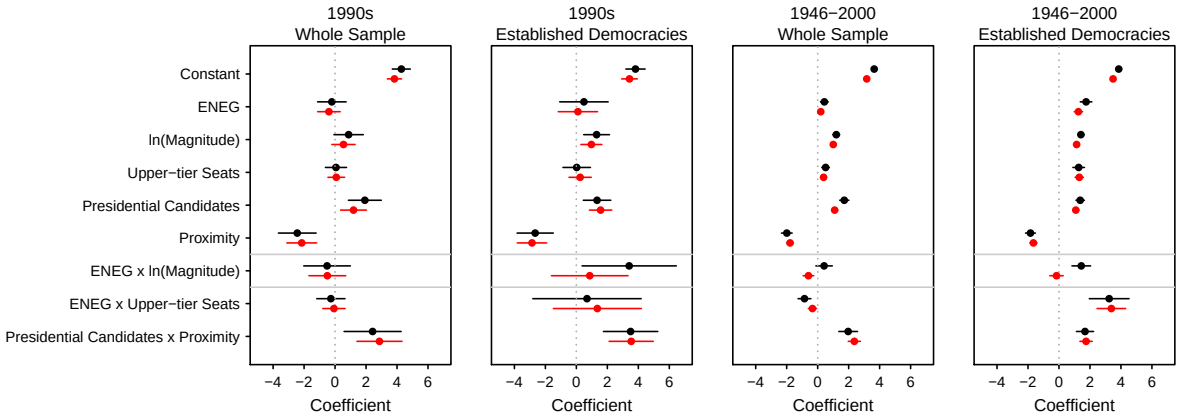

FIGURE 4: Replication of Clark and Golder (2006) using MM-estimation with explanatory variables standardized to have mean zero and standard deviation one-half. The black lines and points show the OLS estimates and 90% confidence intervals and the red lines and points show the MM estimates and confidence 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.

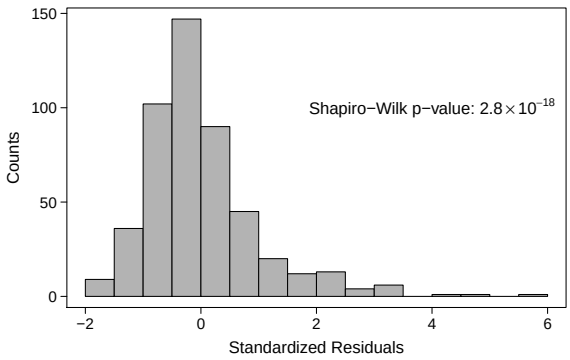

FIGURE 5: A histogram showing the distribution of the residuals from Clark and Golder's (2006) main model. Notice that these residuals do not seem approximately normal. They have a strong skew and heavy tail to the right. For example, one would rarely expect to observe residuals more that three standard 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.

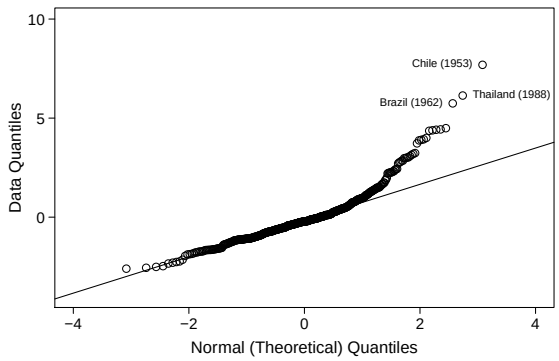

FIGURE 6: A QQ plot showing the deviation of the residuals from normality. If the residuals were approximately normal, then the points in the QQ plot would approximately follow the line. However, notice that the positive residuals deviate sharply from the theoretical expectations. This also suggests that some transformation of the outcome variable might be useful.

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 only established democracies. These results seem to depend on the choice of estimator. In cases such as this, it is especially important to carefully examine the residuals.

For the remainder of our analysis, we focus on the estimates from large sample of countries from 1946-2000 that includes only established democracies. Figure 5 presents the histogram of the residuals from the least-squares estimates and Figure 6 presents the QQ plot for these residuals. Both figures indicate a substantial skew to the right. While this does not necessarily lead to biased estimates, it does, in our view, suggest that the linear model would be more appropriate for a transformed outcome variable. If the transformation makes the errors more closely approximate a normal distribution, then the least squares estimators will be more efficient. In our view, 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 estimated the model using both a log-transformation and Box-Cox transformation with $\lambda = -\frac{1}{3}$. Figures **??** 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.

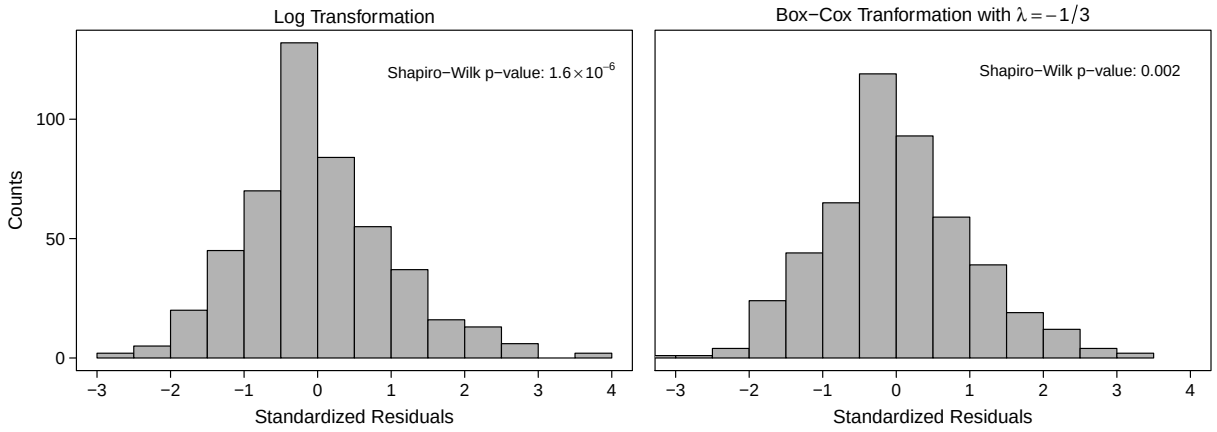

FIGURE 7

However, although the transformation removes much or all of the skew, the residuals retain tails that are slightly heavier than the tails of the normal distribution. This 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, the maximum likelihood estimate of the degrees of freedom is about ten. Recall that the biweight estimator is a (slightly) more efficient estimator for t_{10} distributed errors. However, the least squares estimates *assumed* normally distributed errors and thus the residuals might actually have heavier tails than this analysis suggests.

[Discuss the QQ-plots and the heaviness of the tails.]

But what to we learn from carefully considering whether the residuals match the assumptions of the normal linear model?

- *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.
- *Room for improvement exists in the measurement of "established democracies."*

- *Room for improvement exists in the measurement of "social heterogeneity."*
- *We need a stronger theory about the dynamics with which systems reach an equilibrium number of parties.*
- *We need additional research into how systems respond to the introduction of additional parties into formerly authoritarian, single-party systems.*

Appendix

The Heavy Tails of Electoral Data

A