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 $\hat{\beta}_j = \lambda_1 y_1 + \lambda_2 y_2 + ...\lambda_n y_n$.

However, we argue that restricting ourselves to linear estimators is unnecessary and unproductive. 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.

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.

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

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 1 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 1 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.

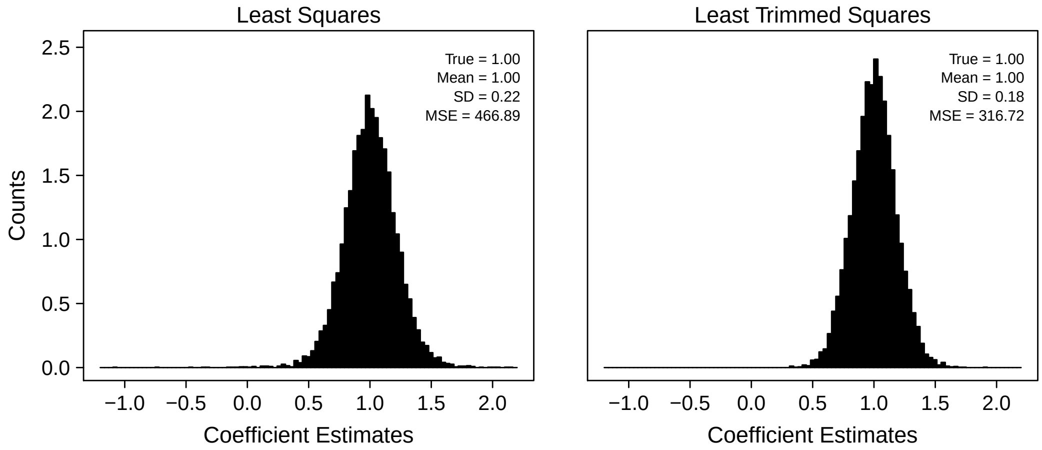

FIGURE 1: caption here

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 distribution 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 non-linearities as evidence that the basic structure of the model could be improved.

[comment on taking logs–common in econometrics.]

[comment on Box-Cox transformation–more general.]

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 MM-estimator.

The MM-estimator, proposed by Yohai (1987). This estimator works extremely well if the assumed normal-linear model holds, though it is not quite as efficient as least squares. On the other hand, when the assumed normal-linear model does not hold, then the MM-estimator might be more efficient than the least squares estimator, perhaps substantially if the deviation is severe. MM-estimation actually combines three of robust estimation into an iterative procedure.

1. Calculate initial coefficient estimates and initial residuals using S-estimation.
2. Calculate the scale of the residuals using M-estimation.
3. Iterate the following until converge:
    a. Update the coefficients using M-estimation with weights determined by the latest residuals.
    b. Calculate new residuals using the latest coefficients.

We now discuss the details of S-estimation and M-estimation. Though the first step of the MM algorithm uses S-estimation, it is convenience to introduce the basic idea of M-estimation first.

M-Estimation

While least squares yields the coefficients that minimize the sum of the squared residuals $\min \sum_{i=1}^{n} \hat{e}_i^2$, M-estimation minimizes some other, less rapidly increasing function of the residuals $\min \sum_{i=1}^{n} \rho(\hat{e}_i)$. Because the solution is not scale invariant, the residuals $\hat{e}_i$ are standardized by a robust estimate of scale $\hat{\sigma}_{(mad)}$, calculated using the median absolute deviation so that $\min \sum_{i=1}^{n} \rho \left(\frac{\hat{e}_i}{\hat{\sigma}_{(mad)}} \right)$.

Replication of Clark and Golder (2006)

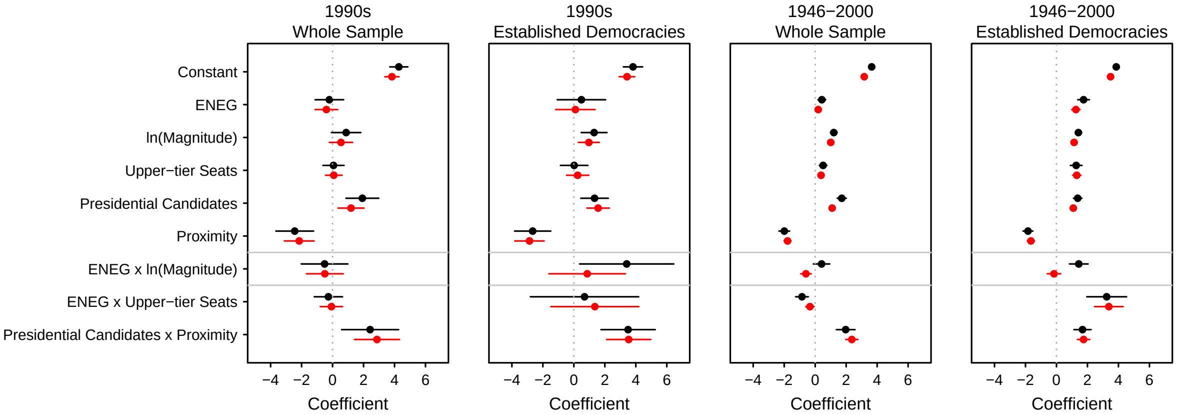

FIGURE 2: Replication of Clark and Golder (2006) using MM-estimation with explanatory variables standardized to have mean zero and standard deviation one-half. The black lines and points show the OLS estimates and 90% confidence intervals and the red lines and points show the MM estimates and confidence 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.

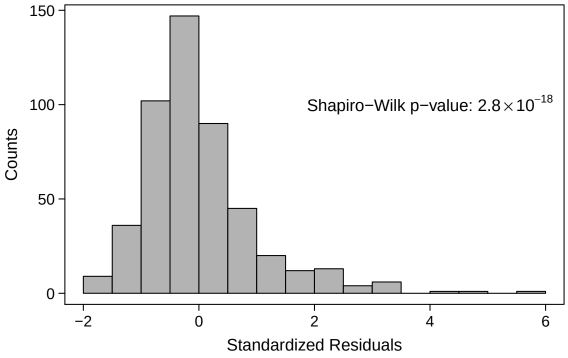

FIGURE 3: A histogram showing the distribution of the residuals from Clark and Golder's (2006) main model. Notice that these residuals do not seem approximately normal. They have a strong skew and heavy tail to the right. For example, one would rarely expect to observe residuals more that three standard 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.

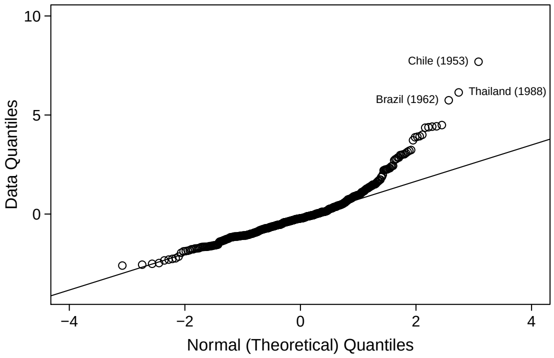

FIGURE 4: A QQ plot showing the deviation of the residuals from normality. If the residuals were approximately normal, then the points in the QQ plot would approximately follow the line. However, notice that the positive residuals deviate sharply from the theoretical expectations. This also suggests that some transformation of the outcome variable might be useful.

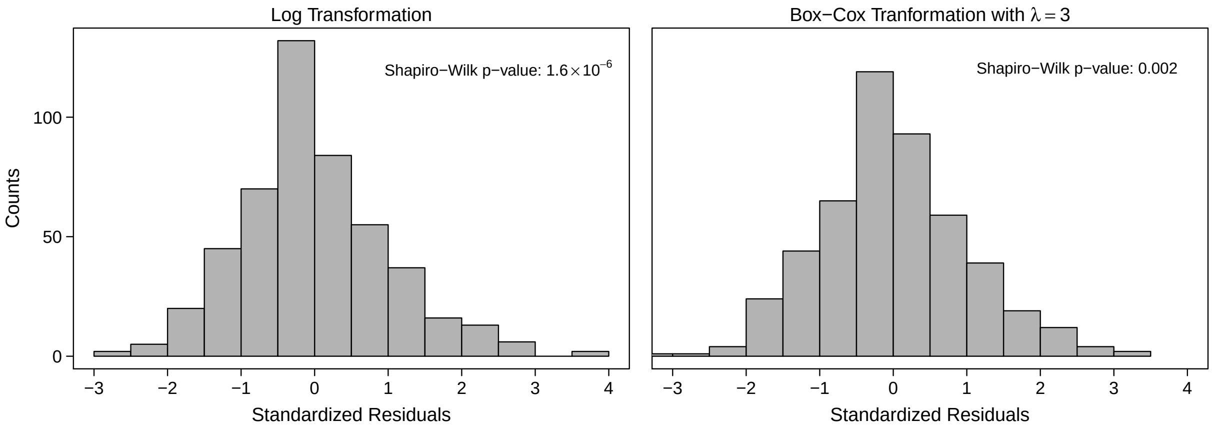

FIGURE 5

Appendix

The Heavy Tails of Electoral Data

A