Political Science Research and Methods (2018), page 1 of 13
doi:10.1017/psrm.2018.34

PSRM

ARTICLE

When BLUE is not best: non-normal errors and the linear model

Daniel K. Baissa[1] and Carlisle Rainey[2*]

[1]Department of Government, Harvard University, 1737 Cambridge St., Cambridge, MA 02138, USA and
[2]Department of Political Science, Florida State University, Room 531B, Bellamy Building, 113 Collegiate Loop, Tallahassee, FL 32306, USA
*Corresponding author. Email: crainey@fsu.edu

Abstract
Researchers in political science often estimate linear models of continuous outcomes using least squares. While it is well known that least-squares estimates are sensitive to single, unusual data points, this knowledge has not led to careful practices when using least-squares estimators. Using statistical theory and Monte Carlo simulations, we highlight the importance of using more robust estimators along with variable transformations. We also discuss several approaches to detect, summarize, and communicate the influence of particular data points.

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

A1: Errors have mean equal to zero.
A2: Errors have a constant, finite variance.
A3: Errors are independent.
A4: Errors follow a normal distribution.

By assuming only A1, we obtain an unbiased, consistent estimator (e.g., Wooldridge 2013, 810, 385, 815–16). However, by assuming A1, A2, A3, and A4, we obtain the best unbiased estimator (BUE), which has the smallest possible variance among the class of unbiased estimators (e.g., Wooldridge 2013, 809–15).

However, political scientists tend to focus on a different property of LS. By the Gauss–Markov theorem, we can remove A4—the assumption of normal errors—and still obtain the best *linear* unbiased estimator (BLUE), which has the smallest possible variance among the class of unbiased, *linear* estimators (e.g., Wooldridge 2013, 809–12). Researchers have primarily justified LS using the Gauss–Markov theorem because it seems to impart desirable small-sample properties without the overly restrictive assumption of normal errors. For example, Berry and Feldman write:

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

quite fortunate that normality is not required for estimation, because it is often very difficult to defend this assumption in practice (1985).[1]

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

The Gauss–Markov theorem has convinced researchers in political science that as long as A1, A2, and A3—the Gauss–Markov assumptions—are met, the distribution of the errors is unimportant. But the distribution of the errors is crucial to a linear regression analysis. Deviations from normality, especially large deviations commonly found in regression models in political science, can devastate the performance of LS compared to alternative estimators. Perhaps even more importantly, regression residuals offer detailed substantive information that researchers often ignore.

In our paper, we emphasize the importance of errors and residuals from a statistical and substantive perspective. We adopt and defend a skeptical perspective toward LS in favor of more robust estimators. We proceed as follows: we (1) clarify the crucial distinction between a linear *model* and a linear *estimator*; (2) explain that the BLUE is not the best estimator unless the errors are normally distributed; (3) highlight powerful, robust alternatives to LS estimators that are unbiased and more efficient for a wide range of substantively plausible error distributions; and (4) provide concrete, practical advice to substantive researchers using linear models.

Is the BLUE the best estimator?

The linear model can be written as $y = X\beta + \varepsilon$. As usual, y is an outcome variable of interest (usually roughly continuous), X a $n \times (k+1)$ matrix containing a single column of ones and k columns holding k explanatory variables, β a $(k+1) \times 1$ matrix of model coefficients, and ε an $n \times 1$ matrix of errors. As usual, the statistical properties of these estimators depend on this model being correct and a full rank X. Researchers in political science commonly estimate this model with LS by minimizing the sum of the squared residuals, such that $\hat{\beta}^{ls} = \arg\min_b S(b)$, where $S(b) = \sum_{i=1}^{n} (y_i - X_i b)^2$. If we assume that the errors ε follow independent and identical normal distributions with mean zero and unknown variance, which we refer to as a "normal linear model," then the LS estimator is the BUE (Casella and Berger 2002, 334–42; Wooldridge 2013, 807–15). This is a powerful result. Under the assumption of normally distributed errors, LS is the most efficient unbiased estimator.

The unreasonable restriction to linear estimators

If we relax the assumption of normality and simply require that the errors have mean zero and constant variance, then the Gauss–Markov theorem guarantees that the LS estimator is the BLUE. 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.

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

At first glance, one might take the linearity restriction in the Gauss–Markov theorem to refer to the structure of the model (i.e., "linear in the parameters"). Indeed, this is the sense in which we use "linear" in the phrase "linear model." However, the "linear" restriction in the Gauss–Markov theorem refers to a different concept—a linear *estimator*, which is a technical condition that has little connection to the substance of the problem. Linearity of the *estimator* requires that the estimates be a linear function of the outcome variable, so that $\hat{\beta}_j = \lambda_1 y_1 + \lambda_2 y_2 + \ldots + \lambda_n y_n$, where the weights λ_i are allowed to depend on X, but not on y. In other words, the Gauss–Markov theorem assumes a linear *model* of the form $E(y|X) = X\beta$, but it also restricts researchers to linear *estimators* of the form $\hat{\beta}_j = \lambda_1 y_1 + \lambda_2 y_2 + \ldots + \lambda_n y_n$.[2]

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

$$y_i = \beta_0 + \beta_1 \log(x_i) + \epsilon_i, \tag{1}$$

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

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i. \tag{3}$$

Indeed, we encourage the use of these models. This collection of linear models illustrates that the linear model can represent a wide range of theoretically relevant relationships, especially when it includes explanatory variables non-linearly. Crucially, though, *there is no statistical reason to restrict ourselves to linear estimators for these linear models*, except for mathematical convenience and computational ease. The only substantive reason to restrict ourselves to linear estimators is if we are willing to assume that we have normally distributed errors. Under the assumption of normal errors, linearity is a reasonable restriction—indeed, the linear estimator is the BUE when the errors are normal. Crucially, restriction to linearity does not get us away from the assumption of normality because the BLUE estimators are the BUE estimators only under the assumption of normal errors. If the errors are not normally distributed, then a researcher can usually do better than the BLUE.

The reasonable consideration of non-linear estimators

Because linearity serves only as an artificial restriction, other unbiased estimators might have smaller variance than the LS estimator. Indeed, in many cases, these alternative estimators possess strongly desirable substantive properties.

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

1. When the normal linear model is exactly correct, the estimator should be approximately unbiased with efficiency comparable to, but less than, LS.

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

2. When the deviation from the normal linear model is small, the estimator should be approximately unbiased with efficiency comparable to, or perhaps greater than, LS.
3. When the deviation from the normal linear model is large, the estimator should have relatively little bias and be much more efficient than LS.

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

To see the importance of this idea in practice, we simulated 10,000 data sets of 50 observations of variables x and y, where the relationship between x and y is given by $y = x + \epsilon$, where ϵ follows a t distribution with three degrees of freedom. 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 LS 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 LS is the BLUE. The left panel of Figure 1 shows the distribution of the estimated slopes using LS.

But we also consider a least trimmed squares (LTS) estimator in which we minimize the smallest 90 percent of the residuals. This method literally discards data. Though it lacks the elegant theory of the LS estimator, the right panel of Figure 1 shows that it is unbiased and more efficient that the LS estimator. The standard deviation of the estimates from the LTS estimator is about 18 percent smaller than the BLUE, and the mean squared error (MSE) is about 32 percent smaller. By any reasonable standard, the LTS estimator is a better estimator than the LS estimator in this example, yet the LS estimator is the BLUE. We obtain this improvement by expanding our focus to non-linear estimators, such as the LTS estimator. In this case, the LTS estimator is clearly non-linear because it places zero weight on the largest 10 percent of the residuals and unit weight on the smallest 90 percent of the residuals.

The relative emphasis on standard errors

There has been much attention in the methodological literature to the sensitivity of standard errors to violations from the assumed model—and substantive scholars have paid attention. White's (1980) seminal paper developing heteroskedasticity-consistent standard errors has received over 20,000 citations, making it one of the most cited papers in economics. Beck and

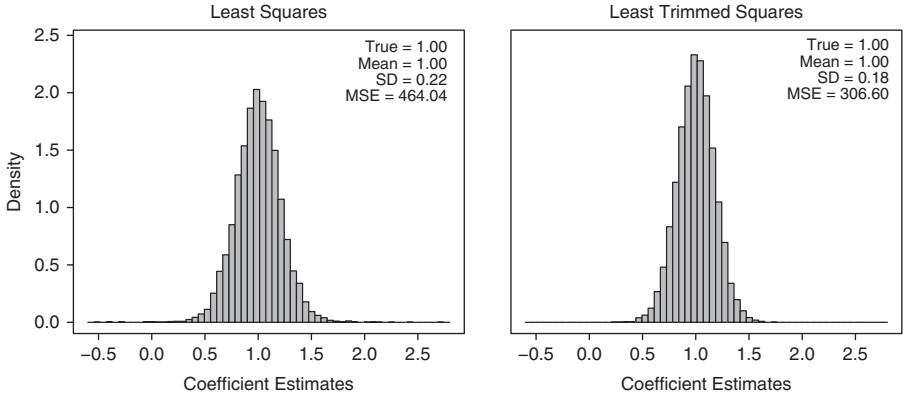

Fig. 1. Histograms of the sampling distribution for the least squares estimator and the least trimmed squares estimator under the true model $y = x + \epsilon$, where $\epsilon \sim t_3$

Note: Despite least squares being the best linear unbiased estimator for the problem, the least trimmed squares estimator is a better estimator. MSE = mean squared error.

Katz's (1995) development of panel corrected standard errors has received over 4300 citations, making it one of the most cited papers in political science.

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

The relative focus on obtaining reasonable standard errors at the expense of reasonable estimates can be seen in Gujarati's (2004) popular econometrics text. 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] (2004, 339).

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

Dealing with skewness: transforming the outcome

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

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

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

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

Second, and perhaps most importantly, skewed residuals from a LS fit indicate model misspecification. While we cannot be certain of the correct model in this situation, we can have confidence that the normal linear model did not produce such data. Sometimes, it is theoretically intuitive that explanatory variables have increasing effects on non-negative outcome variables, such as an individual's annual income. Rather than a college degree increasing one's expected income by \$10,000, perhaps a college degree increases it by 10 percent. If this intuition is correct and a researcher relies on the statistical model $\text{Income}_i = \beta_0 + \beta_1 \text{College Degree}_i + \varepsilon_i$, then the errors will have a strong skew to the right. Simply logging the outcome, or using the model $\log(\text{Income}_i) = \beta_0 + \beta_1 \text{College Degree}_i + \varepsilon_i$, 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 LS estimator by making the assumption of normally distributed errors more appropriate (not to mention the linearity of the model). The performance of LS estimators improves as the error distribution approaches a normal distribution.

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

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

The Box–Cox transformation

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

$$y^{(\lambda)} = \text{BC}(y, \lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{for } \lambda \neq 0 \\ \log y & \text{for } \lambda = 0 \end{cases}, \tag{4}$$

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

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

Mean or median?

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

These interpretational difficulties are not due to the choice to transform the data, but imbedded in the data themselves. The mean $E(\cdot)$ is not a good measure of the location of a skewed variable. While the mean often makes calculations easier, the median offers a better summary of location. The median also has an intuitive interpretation because one-half of the distribution lies above the median and one-half lies below. If a researcher uses $\text{med}(y_{new} \mid X_{new})$ to predict the unknown outcome y_{new} for a known case X_{new}, then she has a 50 percent chance of being too high and a 50 percent chance of being too low.

In addition to the intuitive substantive interpretation of $\text{med}(y \mid X)$, the median has another desirable property. Because the log-transformation is order-preserving, $\text{med}[\log(y) \mid X] = \log[\text{med}(y \mid X)]$, which means that the log *can* easily be undone because $e^{\text{med}[\log(y) \mid X]} = e^{\log[\text{med}(y \mid X)]} = \text{med}(y \mid X)$. Therefore, by adopting $\text{med}(y \mid X)$ and $\frac{\partial \text{med}(y \mid X)}{\partial x_j}$ as the quantities of interest, the researcher

eases the interpretation of the results and can easily move between transformed and untransformed outcomes (e.g., med[log(y)] → med(y)). This holds for the more general case of $y^{(\lambda)}$ as well.

Simulating quantities of interest under transformation

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

1. Estimate the Box–Cox transformation parameter $\hat{\lambda}$ using maximum likelihood. (If the values 1 or 0 fall within the confidence interval, then one may wish to use those values to maintain the direct interpretability of the model coefficients.)
2. Estimate the transformed model $y^{(\hat{\lambda})} = X\beta_{\text{trans}} + \epsilon$ and obtain the estimated model coefficients $\hat{\beta}_{\text{trans}}$ and covariance matrix Σ_{trans}.
3. Choose a hypothetical case or set of cases X_{pred} for which to calculate the quantity of interest. If one is interested in calculating a first difference, it is convenient to use X_{hi} and X_{lo}, where the first-difference $\Delta(y, X_{hi}, X_{lo}) = \text{med}(y \mid X_{hi}) - \text{med}(y \mid X_{lo})$.
4. Following King, Tomz and Wittenberg (2000), for i from one to a large (e.g., 1000) number of iterations n_{sims}:
   a. Simulate $\tilde{\beta}_{\text{trans}} \sim N(\hat{\beta}_{\text{trans}}, \Sigma_{\text{trans}})$.
   b. If interested in the predicted value, then calculate and store $\tilde{Q}_i = \text{med}(y \mid X_{\text{pred}}, \tilde{\beta}_{\text{trans}}) = BC^{-1}(X_{\text{pred}}\tilde{\beta}_{\text{trans}}, \hat{\lambda})$. If interested in the first-difference, the calculate and store $\tilde{Q}_i = \Delta(X_{hi}, X_{lo}, \tilde{\beta}_{\text{trans}}) = \text{BC}^{-1}(X_{hi}\tilde{\beta}_{\text{trans}}, \hat{\lambda}) - \text{BC}^{-1}(X_{lo}\tilde{\beta}_{\text{trans}}, \hat{\lambda})$.
5. Summarize the n_{sims} simulations. The mean or median of $\tilde{Q}$ serves as an estimate of the quantity of interest and the standard deviation of $\tilde{Q}$ serves as an estimate of the standard error. The 5th and 95th percentiles of $\tilde{Q}$ serve as an estimate of the (likely asymmetric) 90 percent confidence interval for the quantity of interest.

Dealing with heavy tails: *M*-estimation

In spite of the scant attention paid to robust estimators in political science, statisticians have developed and refined many robust methods since the seminal work of Box (1953) and Huber (1964). Huber and Ronchetti (2009) provide a detailed review of these developments and Anderson (2008) provides an accessible introduction. Adjudicating among the many robust alternatives to LS is beyond the scope of our paper, but, to fix ideas, we do introduce one robust estimator developed by Beaton and Tukey (1974) in detail which has several desirable properties —the *M*-estimator with Tukey's biweight (BW) function. However, there are many other options: *M*-estimators with other objective functions (e.g., Huber 1973), LMS- and LTS-estimators (Rousseeuw 1984), S-estimators (Rousseeuw and Yohai 1984), and MM-estimators (Yohai 1987).[3]

While LS yields the coefficients that minimize the sum of the squared residuals, so that $\hat{\beta}^{ls} = \text{argmin}_b \sum_{i=1}^{n} (y_i - X_i b)^2$, M-estimators minimize an arbitrary (usually), less-rapidly increasing function of the residuals $\hat{\beta}^{\rho} = \text{argmin}_b \sum_{i=1}^{n} \rho(y_i - X_i b)$. The function $\rho(\cdot)$ is typically chosen to be non-negative, symmetric about zero, and increasing away from zero. For example, Harden and

[3]User-friendly software to estimate models using these methods is available in both R and Stata. For the *M*-estimator with Tukey's BW function, we particularly recommend the rlm() function in the R package MASS (Venables and Ripley 2002) and the rreg command in Stata. For the other robust estimators, we point readers to the R package robustbase (Rousseeuw et al. 2016) and the user command robreg in Stata (Jann 2010).

Desmarais (2011) recommend the least absolute deviation (LAD) estimator (Dodge 1987) such that $\rho(\cdot) = \text{abs}(\cdot)$. However, other estimators offer better performance, particularly when the normal linear model is approximately correct. In particular, we recommend Tukey's BW function (Beaton and Tukey 1974), so that

$$\rho_{bw}(r_i) = \begin{cases} \frac{k^2}{6}\left\{1-\left[1-\left(\frac{r_i}{k}\right)^2\right]^3\right\} & \text{for } |\, r_i \,| \leq k \\ \frac{k^2}{6} & \text{for } |\, r_i \,| > k \end{cases}, \tag{5}$$

where $r_i = y_i - X_i b$ and k is a tuning parameter usually set to 4.685 to ensure good performance under the normal linear model. We refer to the M-estimator using the BW objective function as the "BW estimator." The BW estimator is a compelling alternative to the LAD estimator suggested by Harden and Desmarais (2011) for two reasons. First, the BW 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 receive some weight. Second, the BW estimator is much more efficient than the LAD estimator when the errors are approximately normal.

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

While the theory for M-estimators remains less complete than the theory for LS estimators, M-estimators do have desirable statistical properties. M-estimators are consistent as long as (1) $\rho(\cdot)$ is convex or (2) the errors follow a strongly unimodal distribution (i.e., decreasing away from zero). Because the BW objective function is not convex, we must assume that the errors follow a strongly unimodal distribution, which ensures that the estimates are consistent and distributed asymptotically normal.

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

Estimation

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

1. Start with initial estimates of the coefficients $\hat{\beta}^{(0)}$. The choice of initial estimator is not trivial. In the case of extreme outliers or many parameters, starting with LS might lead the algorithm to a local minimum. We recommend using the LTS method discussed earlier to obtain starting values.
2. Extract the residuals $r^{(0)} = y - X\hat{\beta}^{(0)}$. Use these residuals to estimate the rescaled MAD (median absolute deviation) so that $\hat{\sigma}^{(0)}_{\text{mad}} = \frac{\text{med}\left(|\, r^{(0)} \,|\right)}{0.6745}$.

3. For i from one until convergence:
   a. Using $\hat{\beta}^{(i-1)}$ and $\hat{\sigma}_{\text{mad}}^{(i-1)}$ assign weights w according to the function ρ and denote $\text{diag}(w) = W^{(i)}$.
   b. Calculate $\hat{\beta}^{(i)} = (X'W^{(i)}X)^{-1}X'W^{(i)}y$.
   c. Calculate $\hat{\sigma}_{\text{mad}}^{(i)} = \frac{\text{med}(|y - X\hat{\beta}^{(i)}|)}{0.6745}$
   d. The algorithm has converged when $r^{(i-1)} \approx r^{(i)}$.

If we assume that the errors are symmetrically distributed about zero, then any objective function ρ that is also symmetric about zero (including, e.g., the BW objective function) produces an unbiased estimate of the parameters. But this estimator is *linear* if and only if $\rho(r_i) = r_i^2$. Other choices of $\rho(\cdot)$ might produce better estimators than the BLUE.

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

Monte Carlo simulations

To understand and illustrate how the performance of the BW estimator compares with the common LS estimator, we conducted a Monte Carlo study. For comparison, we included the LTS estimator discussed above and the LAD estimator suggested by Harden and Desmarais (2011). To conduct the study, we simulated from the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = 1$ and the x_i's were generated from independent standard normal distributions. We used six different distributions for the errors, each symmetric and centered at zero.

- *Laplace distribution.* The Laplace distribution has tails that decrease exponentially, but behaves much differently from the normal distribution near zero. Rather than "shoulders," the Laplace distribution has a sharp peak at zero and can be thought of as combining two exponential distributions, one in the positive direction and the other in the negative direction. The LAD estimator is the maximum likelihood estimator when the errors follow a Laplace distribution.
- t_2 *distribution.* The t distribution with two degrees of freedom has very heavy tails. Because the LS estimator weights all points equally (conditional on X), the extreme outliers produced by the t_2 distribution makes LS a very inefficient estimator.
- t_{10} *distribution.* The t distribution with ten degrees of freedom has *slightly* heavier tails than the normal distribution. The t_{10} and normal distributions are so similar that a Shapiro–Wilk test of normality only correctly rejects the null in about 65 percent of repeated samples if 500 observation are simulated from a t_{10} distribution.[4] 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.
- *Logistic distribution.* The logistic distribution has tails similar to the t_{10} distribution—slightly heavier than the normal distribution. Researchers sometimes assume that latent errors in discrete choice models follow a logistic distribution (Train 2009).
- *Uniform distribution.* Simply for comparison, we include a uniform distribution from −1 to 1.
- *Normal Distribution.* The normal distribution yields the optimal conditions for the LS estimator. When the errors follow a normal distribution, the LS estimator has the smallest variance of all unbiased estimators.

[4]One needs about 750 samples to reach 80 percent power.

For the six different error distributions, we simulated 10,000 data sets with 100 observations each and estimated β_1 using the LS estimator, the BW estimator, the LAD estimator, and the LTS estimator. For each condition, we calculated the MSE of the estimate of β_1. To simplify the presentation, we divided the MSE of the BW, LAD, and LTS estimators by the MSE of the LS estimator. Table 1 provides the results. These results show that the MSE varies considerably across the estimators for each error distribution.

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 an MSE about 30 percent lower than the LS estimate. However, the BW estimator also performs quite well for the Laplace distribution, with an MSE about 25 percent less than the LS estimator. The LTS estimator performs the worst of the robust estimators, but the MSE of the LTS estimator is still about 5 percent less than the MSE for the LS estimator.

The t_2 distribution is nearly a worst case for the LS estimator, so all three robust alternatives perform considerably better. The BW estimator has an MSE about 85 percent lower than the LS estimator and the LAD and LTS estimators have an MSE about 83 and 82 percent lower, respectively.

The t_{10} distribution is a much more interesting case, because it is very similar to a normal distribution. In this case, the LAD and LTS estimators have a 33 and 44 percent larger MSE than the LS estimator, respectively. The BW estimator, on the other hand, shows a small *improvement* over the LS estimator, with an MSE about 2 percent *smaller* than the LS estimator. This is crucial because it shows that only a small deviation from normality is required before the BLUE estimator is no longer the BUE estimator.

The logistic and t_{10} distributions are similar and produce similar results. The LAD and LTS estimators perform worse than the LS estimator, with the MSE 24 and 31 percent larger than the LS estimator, respectively. Again, though, the BW estimator shows a small improvement over the LS estimator, with an MSE about 5 percent smaller than the LS estimator.

For the uniform distribution, the LAD and LTS estimators perform much worse than the LS estimator, with an MSE 171 and 288 percent larger, respectively. However, the BW estimator performs similarly to the LS estimator with an MSE only 16 percent larger.

The normal distribution is the optimal scenario for the LS estimator and it outperforms the LAD and LTS estimator considerably when the errors are normal. In this situation, the MSE for the LAD and LTS estimator is 58 and 71 percent larger than the MSE for the LS estimator, respectively. However, the BW estimator performs nearly as well as the LS estimator for normal errors. The MSE for the BW estimator is only about 7 percent larger than the MSE for the LS estimator.

Among the estimators we consider, the BW estimator does not have the smallest variance for the Laplace, uniform, or normal distributions, but it is a *close* second. It does have the smallest variance for the t_2, t_{10}, and logistic distributions. It considerably outperforms the LS estimator for Laplace and t_2 errors and the LAD and LTS estimators for normal and t_{10}, logistic, and uniform errors. Thus, the BW estimator works quite well across a range of error distributions, whereas the LS, LAD, and LTS estimators work well only in particular situations. And even in particular

Table 1. The Mean Squared Error of the Biweight (BW), Least Absolute Deviation (LAD), and Least Trimmed Squares (LTS) Estimators Compared to the Least Squares (LS) Estimator for Six Different Error Distributions With a Sample Size of 100

	Laplace	t_2	t_{10}	Logistic	Uniform	Normal
BW/LS	0.75	0.15	0.98	0.95	1.16	1.07
LAD/LS	0.70	0.17	1.33	1.24	2.71	1.58
LTS/LS	0.93	0.18	1.44	1.31	3.88	1.71

Note: The BW has the best or nearly best performance in each condition, while the LAD and LTS estimators performs quite poorly for the t_{10} and normal distributions, and the LS estimator performs quite poorly for the Laplace and t_2 distributions.

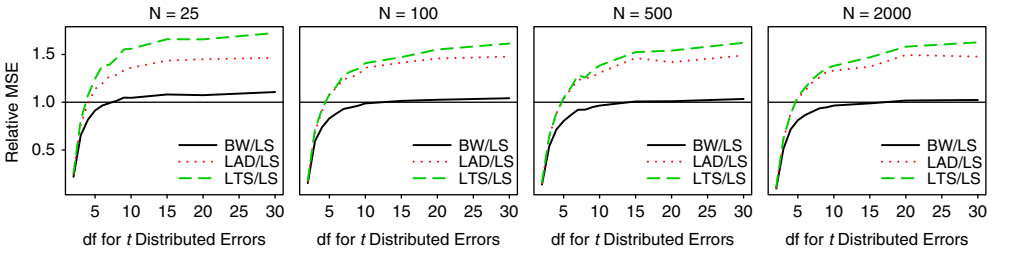

Figure 2. The relative mean squared error (MSE) for the biweight (BW), least absolute deviation (LAD), and least trimmed squares (LTS) estimators compared to the least squares (LS) estimator for *t* distributed errors as the degrees of freedom varies.

Note: For very heavy-tailed distributions (e.g., 2–5 degrees of freedom), the BW, LAD, and LTS estimators significantly outperform the LS estimator. And while the performance of the LAD and LTS estimators significantly worsens as the distribution becomes more normal, the BW estimator remains comparable to the LS estimator.

situations where the LS, LAD, and LTS 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 with degrees of freedom ranging from two to 30 and sample sizes of 25, 100, 500, and 2000. Figure 2 shows the MSE of the LAD and BW estimators relative to the LS estimator.

Notice that the BW, LAD, and LTS estimators perform quite well for very heavy-tailed distributions (i.e., degrees of freedom from two to four), but as the tails grow lighter, the LS estimator quickly begins to outperform the LAD and LTS estimator. For all but very heavy-tailed distributions, the LS estimator is considerably more efficient than the LAD and LTS estimators.

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

Recommendations for applied researchers

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

1. Initially fit the model using LS.
2. As a "robustness check," re-fit the model using a robust alternative, such as the BW estimator.
3. If the inferences change (and even if not), carefully examine the residuals using histograms and QQ plots. Carefully check for skewness.
4. If the residuals are not symmetrically distributed, then consider a transformation. This transformation might be critical because it allows the model to represent non-linear relationships implied by the skewness and allows the statistical model to more closely approximate the data. The log-transformation has a nice substantive interpretation, so it makes sense as a first cut, especially for variables naturally bounded below by 0 or 1. If the log-transformation over- or under-corrects the skewness, then the Box–Cox transformation might create roughly symmetric residuals.
5. Once the residuals are roughly symmetric, re-fit the model using LS and a robust alternative. Especially if the residuals appear to have heavy tails, then the robust estimator

might serve as a more efficient estimator. However, the robust estimator also allows for greater substantive interpretation as well, because it allows unusual cases to stand out.
6. Always pay close attention to the residuals from each model, especially differences, as these can be especially substantively informative.
7. To the extent that some cases seem unusual, especially with robust regression, give these cases careful review. Is it possible that these unusual outcomes are 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?

To illustrate these recommendations, we include a reanalysis of Clark and Golder (2006) as an Online Appendix.

Conclusion

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

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

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

Supplementary materials. To view supplementary material for this article, please visit https://doi.org/10.1017/psrm.2018.34. Replication materials are available at https://doi.org/10.7910/DVN/WZSUS3

Acknowledgments. The authors thank Bill Clark and Matt Golder for making their data available to the authors. An audience at the 2015 Innovations in Comparative Political Methodology Conference at Texas A&M University provided valuable feedback on an earlier draft of this manuscript. The analyses presented here were conducted with R 3.2.2. All data and computer code necessary for replication are available at https://github.com/carlislerainey/heavy-tails

References

Anderson R (2008) *Modern Methods for Robust Regression*. Thousand Oaks, CA: Sage.

Angrist JD and Pischke J-S (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Beaton AE and Tukey JW (1974) The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data. *Technometrics* **16**(2), 147–185.

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

Berry WD and Feldman S (1985) *Multiple Regression in Practice. Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage.

Box GEP (1953) Non-Normality and Tests on Variances. *Biometrika* **40**(3/4), 318–335.

Box GEP and Cox DR (1964) An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B* **26**(2), 211–252.

Casella G and Berger RL (2002) *Statistical Inference* 2nd ed. Pacific Grove, CA: Duxbury.
Clark WR and Golder M (2006) Rehabilitating Duverger's Theory: Testing the Mechanical and Strategic Modifying Effects of Electoral Laws. *Comparative Political Studies* **39**(6), 679–708.
Dodge Y (ed.) (1987) *Statistical Data Analysis Based on the Ll-Norm and Related Methods*. Amsterdam: North-Holland.
Efron B (1981) Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Methods. *Biometrika* **68**(3), 589–599.
Freedman DA (2006) On the So-Called "Huber Sandwich Estimator" and "Robust Standard Error". *The American Statistician* **60**(4), 299–302.
Gujarati DN (2004) *Basic Econometrics* 4th ed. Boston, MA: McGraw Hill.
Harden JJ and Desmarais BA (2011) Linear Models with Outliers: Choosing Between Conditional-Mean and Conditional-Median Methods. *State Politics and Policy Quarterly* **11**(4), 371–389.
Huber PJ (1964) Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* **35**(1), 73–101.
Huber PJ (1973) Robust Regression: Asymptotics, Conjectures, and Monte Carlo. *The Annals of Statistics* **1**(5), 799–821.
Huber PJ and Ronchetti EM (2009) *Robust Statistics* vol. 2nd. Hoboken, NJ: Wiley.
Jann B (2010) 'robreg: Stata Module Providing Robust Regression Estimators'. Available at http://ideas.repec.org/c/boc/bocode/s457114.html
King G and Roberts ME (2014) How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It. *Political Analysis* **23**(2), 159–179.
King G, Tomz M and Wittenberg J (2000) Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science* **44**(2), 341–355.
Krueger JS and Lewis-Beck MS (2008) Is OLS Dead? *The Political Methodologist* **15**(2), 2–4.
Mira A (1999) Distribution-Free Test for Symmetry Based on Bonferroni's Measure. *Journal of Applied Statistics* **26**(8), 959–972.
Mooney CZ and Duval RD (1993) *Bootstrapping: A Nonparametric Approach to Statistical Inference. Quantitative Applications in the Social Sciences*. Newbery Park, CA: Sage.
Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M and Maechler M (2016) 'robustbase: Basic Robust Statistics'. R Package Version 0.92-6. Available at http://CRAN.R-pro ject.org/package=robustbase
Rousseeuw PJ (1984) Least Median of Squares Regression. *The Journal of the American Statistical Association* **79**(388), 871–880.
Rousseeuw PJ and Yohai V (1984) 'Robust Regression by Means of S-Estimators'. In J Franke, W Hardle and D Martin (eds), *Robust and Nonlinear Time Series Analysis*, vol. 26, Lecture Notes in Statistics Springer US, 256–272. NY: Springer.
Train KE (2009) *Discrete Choice Methods with Simulation* 2nd ed. New York: Cambridge University Press.
Venables WN and Ripley BD (2002) *Modern Applied Statistics with S*. New York: Springer.
Western B (1995) Concepts and Suggestions for Robust Regression Analysis. *American Journal of Political Science* **39**(3), 786–817.
White H (1980) A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* **48**(4), 817–838.
Wooldridge JM (2013) *Introductory Econometrics: A Modern Approach* 5th ed. Mason, OH: South-Western Cengage Learning.
Yohai V (1987) High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics* **15**(2), 642–656.

Cite this article: Baissa DK, Rainey C. 2018. When BLUE is not best: non-normal errors and the linear model. Political Science Research and Methods X: 1–13, doi: 10.1017/psrm.2018.34