When BLUE Is Not Best

Non-Normal Errors and the Linear Model

Carlisle Rainey
Assistant Professor
University at Buffalo, SUNY

Daniel K. Baissa
Graduate Student
University at Buffalo, SUNY

Paper, code, and data at
carlislerainey.com/research

Last spring, I taught the linear models class at UB. Before this class, I hadn't thought hard about linear models—all of my work was about binary outcomes.

This paper is just a commentary on the linear regression model as used in practice in political science.

Key Point

Gauss-Markov theorem is an elegant result, but it's not useful for applied researchers.

Key Point

Normality matters.

———————

but first, a little background

Background

$$y_i = X_i\beta + \epsilon_i$$

outcome variable

design matrix

coefficient vector

errors

Technical assumptions:

1. The design matrix is full rank.
2. The model is correct.

Additional assumptions:

1. Errors have mean zero.
2. Errors have constant, finite variance.
3. Errors are independent.
4. Errors follow a normal distribution.

Additional assumptions:

1. Errors have mean zero.
2. Errors have constant, finite variance.
3. Errors are independent.
4. Errors follow a normal distribution.

A1 → consistency

Additional assumptions:

1. Errors have mean zero.
2. Errors have constant, finite variance.
3. Errors are independent.
4. Errors follow a normal distribution.

A1-A4 → BUE

But is there something in between?

Can we get a small-sample result without normality?

Additional assumptions:

1. Errors have mean zero.
2. Errors have constant, finite variance.
3. Errors are independent.
4. Errors follow a normal distribution.

A1-A3 → BLUE

(Gauss-Markov Theorem)

But this is not a powerful result.

The “L” is unnecessarily restrictive and counterproductive.

Linearity in BLUE

Linearity in BLUE

linear model

or

linear in the parameters

Linearity in BLUE

~~linear model~~

or

~~linear in the parameters~~

Linearity in BLUE

linear estimator

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

or

$$\hat{\beta} = My$$

Linearity in BLUE

linearity ≅ easy

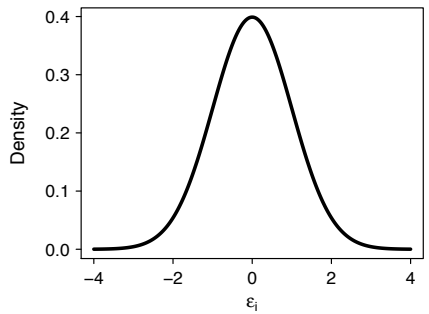
$$\hat{\beta} = My = (X'X)^{-1}X'y$$

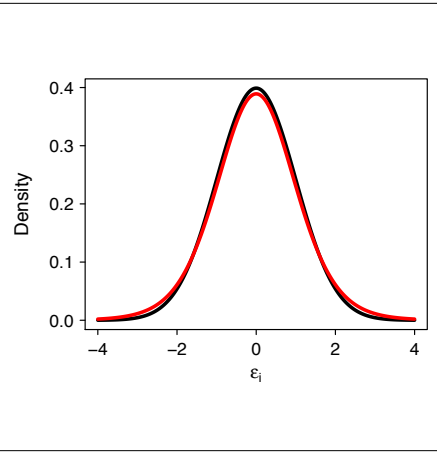
Linearity in BLUE

Question:

BLUE ≅ BUE?

How large of a deviation from normal errors before LS is not approximately BUE?





Restriction to linear estimators makes statistical sense only when errors are normal.

This is the only time when a linear estimator is a good estimator.

Practical Importance

"[Even without normally distributed errors] OLS coefficient estimators remain unbiased and efficient."

–Berry (1993)

"[The Gauss-Markov theorem] justifies the use of the OLS method rather than using a variety of competing estimators."

–Wooldridge (2013)

"We need not look for another linear unbiased estimator, for we will not find such an estimator whose variance is smaller than the OLS estimator."

–Gujarati (2004)

"An important result in multiple regression is the Gauss-Markov theorem, which proves that when the assumptions are met, the least squares estimators of regression parameters are unbiased and efficient."

–Berry and Feldman (1993)

"The Gauss-Markov theorem allows us to have considerable confidence in the least squares estimators."

–Berry and Feldman (1993)

Gauss-Markov has convinced researchers that residuals are not important.

Skewed residuals might indicate misspecification an in any case they make LS inefficient

Least squares is inefficient under heavy tails.

Residuals themselves are informative.

Alternatives

Skewness

log transformation or Box-Cox transformation

but I’m not going to talk about this

Heavy Tails

$$\hat{\beta}^{LS} = \arg\min_b \sum_{i=1}^{n} (y_i - X_i b)^2$$

if you use this objective function, then your estimator works too hard to explain points that fall far from the line.

$$\hat{\beta}^{\rho} = \arg\min_b \sum_{i=1}^{n} \rho(y_i - X_i b)$$

Choose a rho that has roughly the same shape, but goes to infinity more slowly.

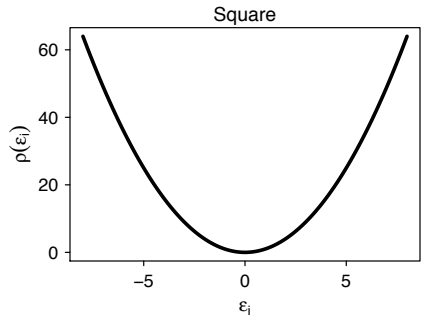
M-estimator

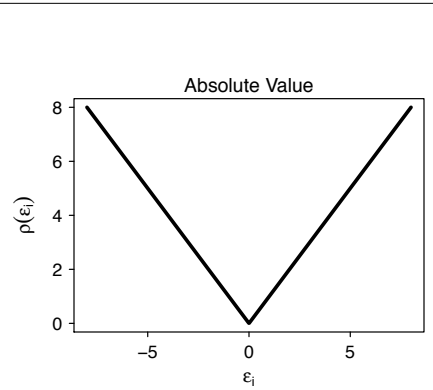
———————————-

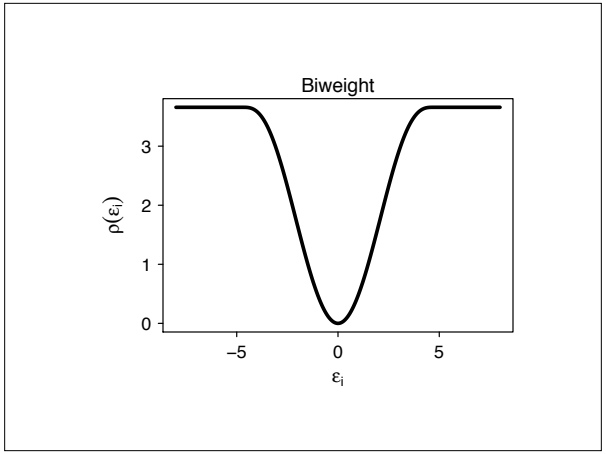
Choose rho to accomplish two goals

Choose function ρ such that the estimator:

1. performs nearly as well as LS for normal errors
2. performs much better than LS for non-normal errors.







This is my preferred robust estimator.

> normal-ish near zero, but flat away from zero.

>

– – – – – – – – – – – – – -

But why would we want to use a robust estimator? Two reasons

Robust estimators are often more efficient than LS.

Robust estimators allow unusual cases to be unusual.

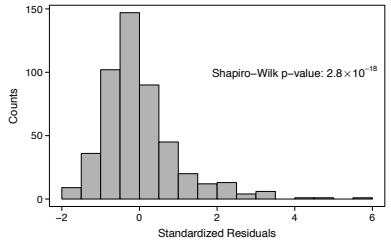
Clark and Golder

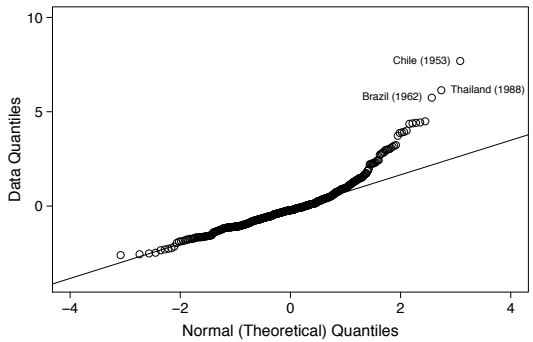
>> Illustrate the kinds of things we can learn by paying close attention to the residuals.

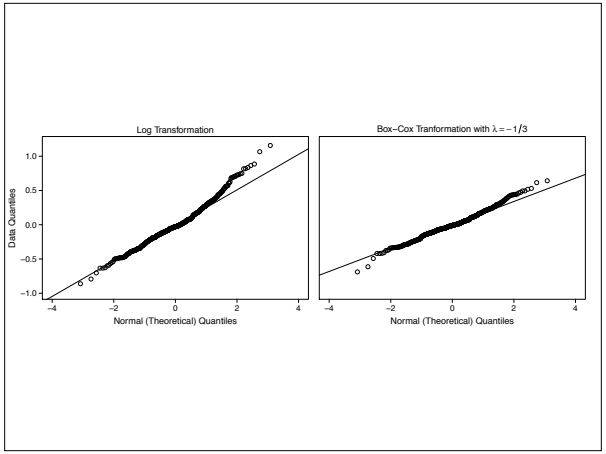
>> District magnitude conditions social heterogeneity.

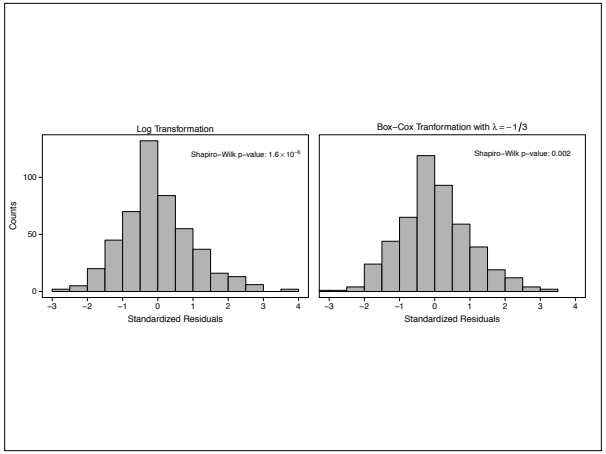
>> When m is low

>> When m is high

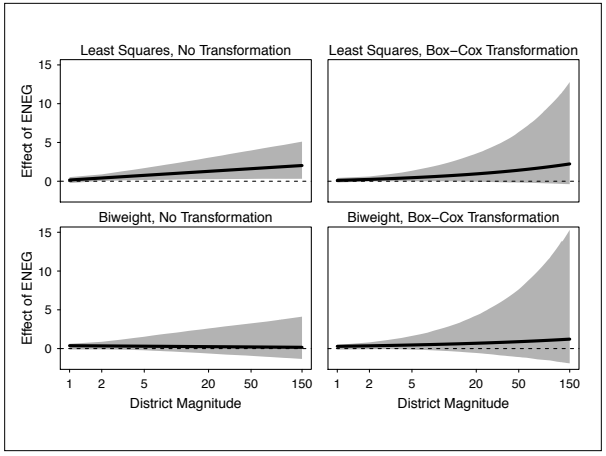


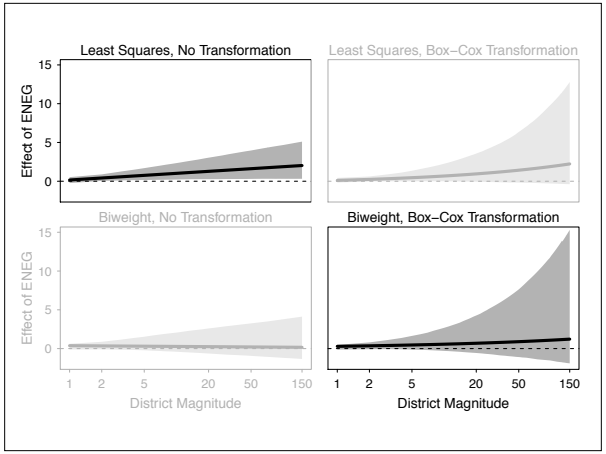






about a t10 or a t6





Substantive Takaways

Substantive Takaways

The theory is wrong.

— — — — — — — — — — — —

I don't think that's the right conclusion for two reasons.

First

Substantive Takaways

The theory is wrong.

We've got lots of evidence in favor of the theory.

- Theoretical
- Observational studies
- Quasi-experiments
- Lab experiments

But also

Substantive Takaways

The theory is wrong.

The estimates are suggest the effects might be smaller *or larger* than Clark and Golder's analysis suggests.

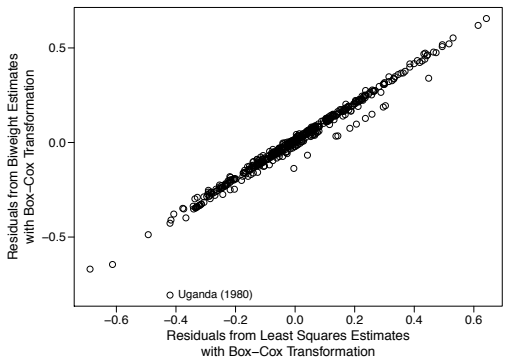
Substantive Takaways

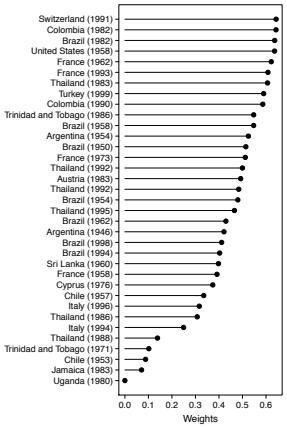
~~The theory is wrong~~.

Substantive Takaways

We can learn from the residuals.

But most importantly..





Substantive Takaways

The 1980 election in Uganda

> British rule until 1962
> 1962 elections -> Milton Obote prime minister.
> Four years later, facing scandal, Obote suspended the constitution, assumed all government powers, declared himself president banned all political parties.

> [Military coup] 1971 Major General Idi Amin ousted Obote.
> until 1979
> 50,000 and 300,000 civilians killed.

Substantive Takaways

What is an “established democracy”?

democracy?

Substantive Takaways

What dynamics lead to equilibrium?

Substantive Takaways

How do these dynamics depend on the prior regime?

Key Points

>> Illustrate the kinds of things we can learn by paying close attention to the residuals.

>> District magnitude conditions social heterogeneity.

>> When m is low

>> When m is high

Point #1

Normality is an important assumption of least squares.

Point #2

Alternatives to least squares often exhibit better behavior for non-normal errors.

Point #3

Researchers can learn much from unusual cases.

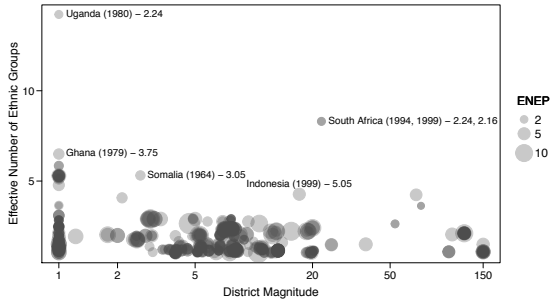
Even More!

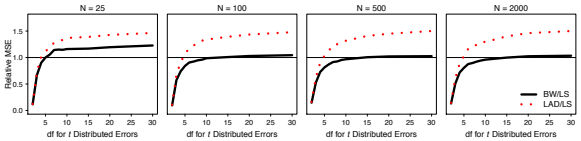
>> Illustrate the kinds of things we can learn by paying close attention to the residuals.

>> District magnitude conditions social heterogeneity.

>> When m is low

>> When m is high





	Mean Squared Error			
	Lapl.	t_2	t_{10}	Norm.
Absolute Performance				
Least Squares	231.072	1571.227	149.507	87.103
Least Absolute Deviation	164.875	305.173	196.751	133.454
Tukey's Biweight	171.136	272.269	145.291	92.514
Relative Performance				
LAD/LS	0.714	0.194	1.316	1.532
BW/LS	0.741	0.173	0.972	1.062

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