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

Key Point

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

Key Point

Normality matters.

Background

$$y_i = \boldsymbol{X}_i \beta + \epsilon_i$$

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?

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.

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 <u>estimator</u>

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

or

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

Linearity in BLUE

linearity ≅ easy

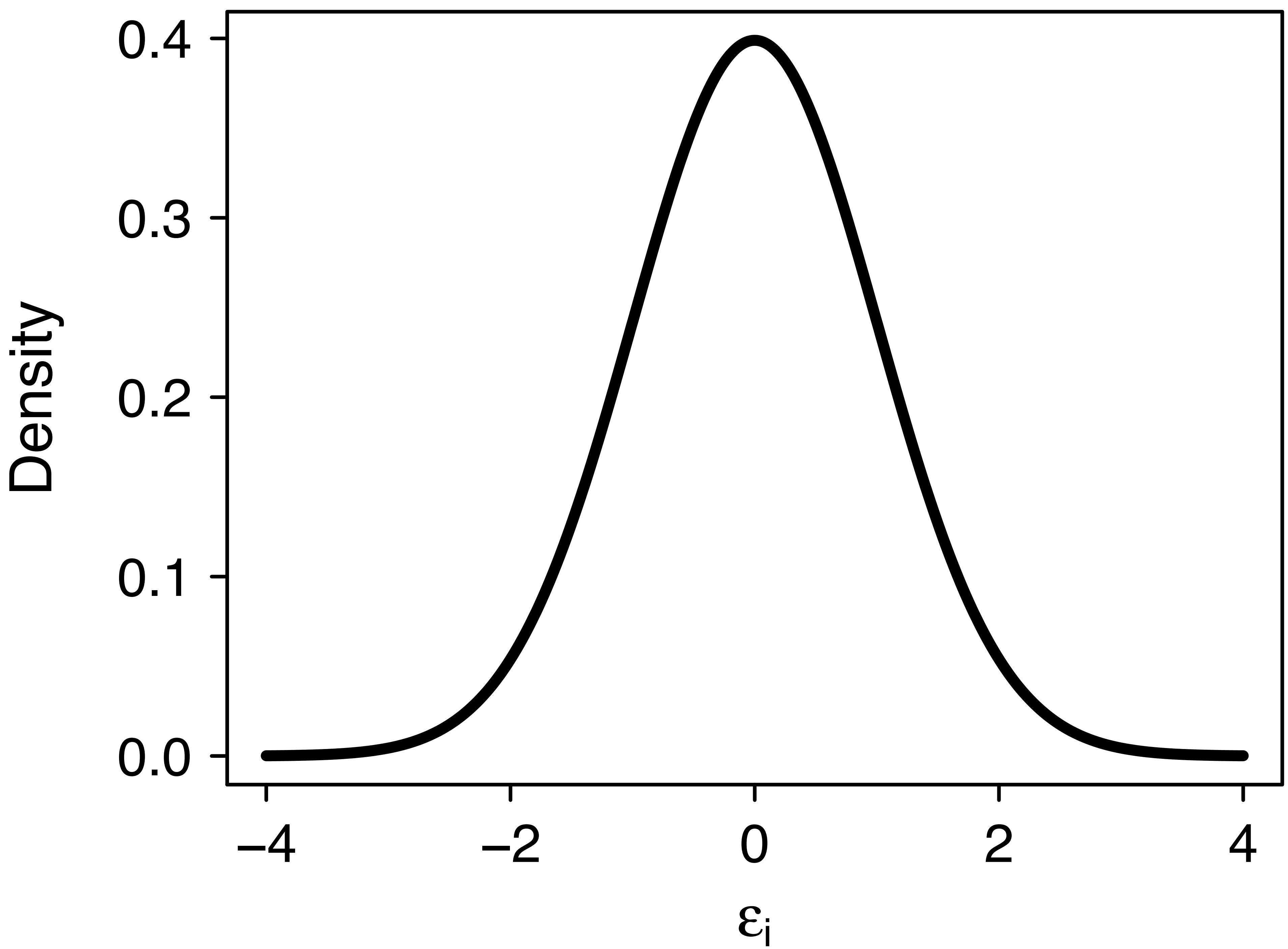
$$\hat{\beta} = My = (X'X)^{-1}X'y$$

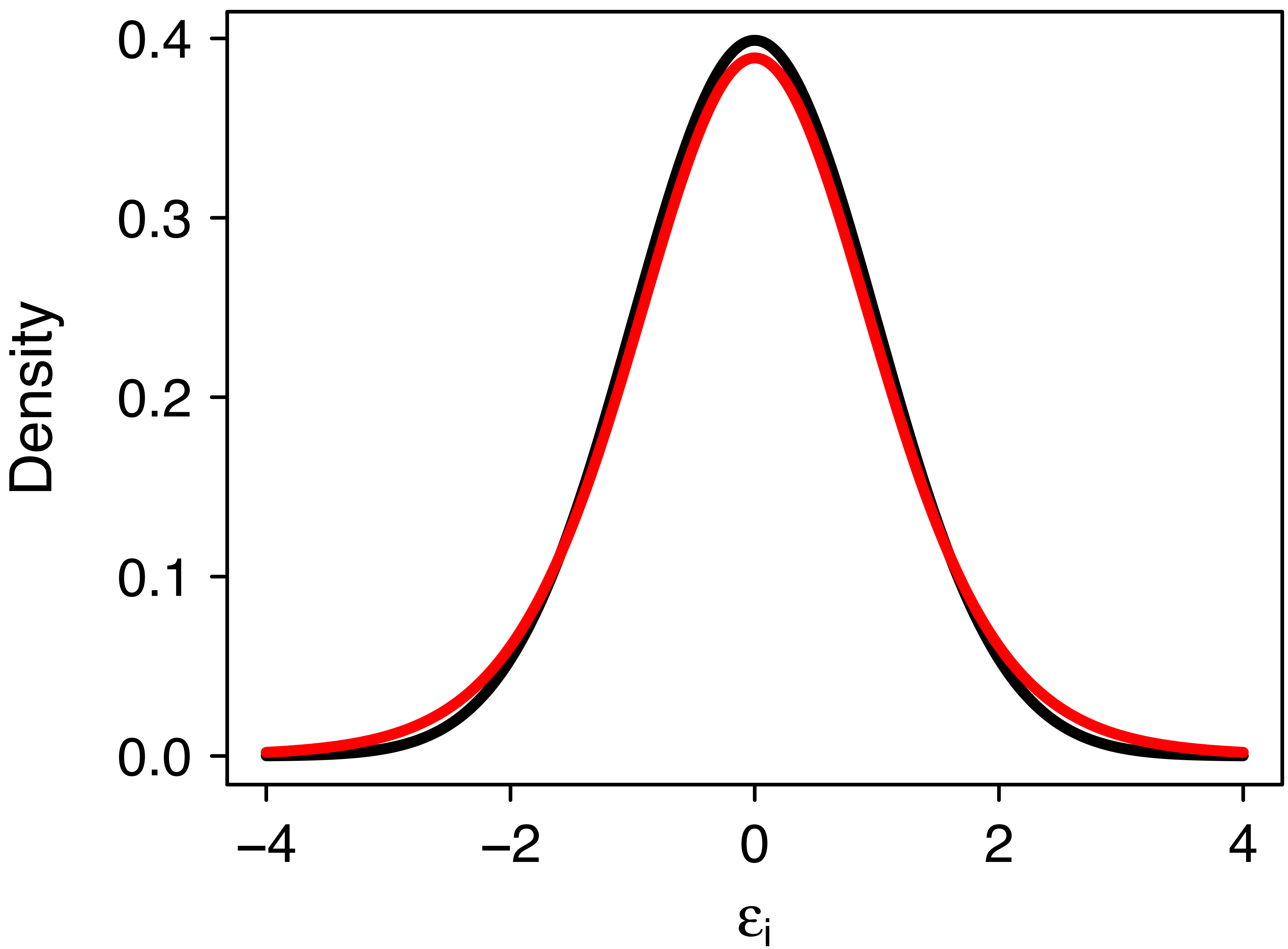
Linearity in BLUE

Question:

BLUE $\cong$ BUE?

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


0.4
0.3
0.2
0.1
0.0
Density
−4
−2
0
2
4
εi



0.0
0.1
0.2
0.3
0.4
Density
−4
−2
0
2
4
εi

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

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.

Alternatives

Skewness

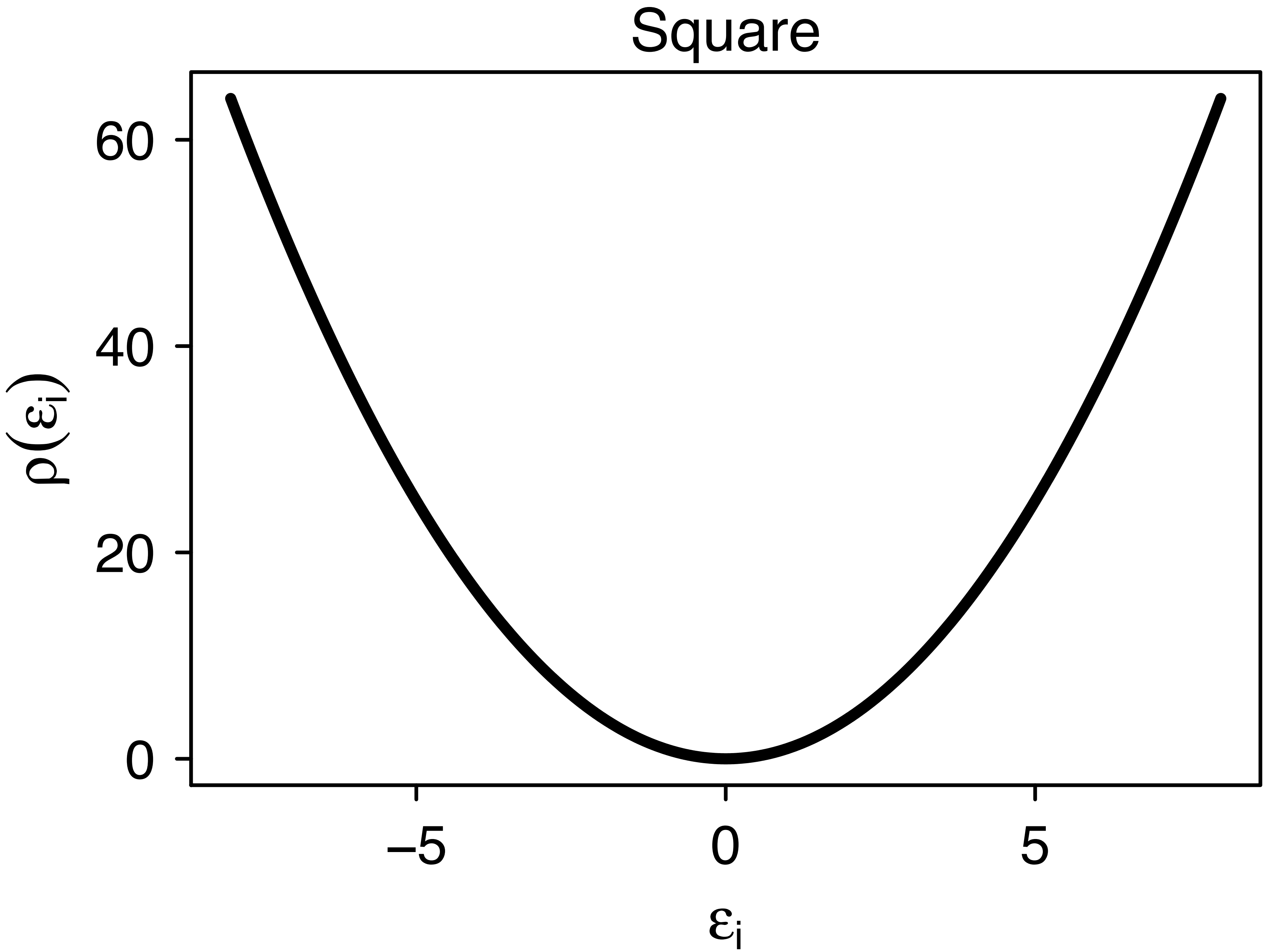
Heavy Tails

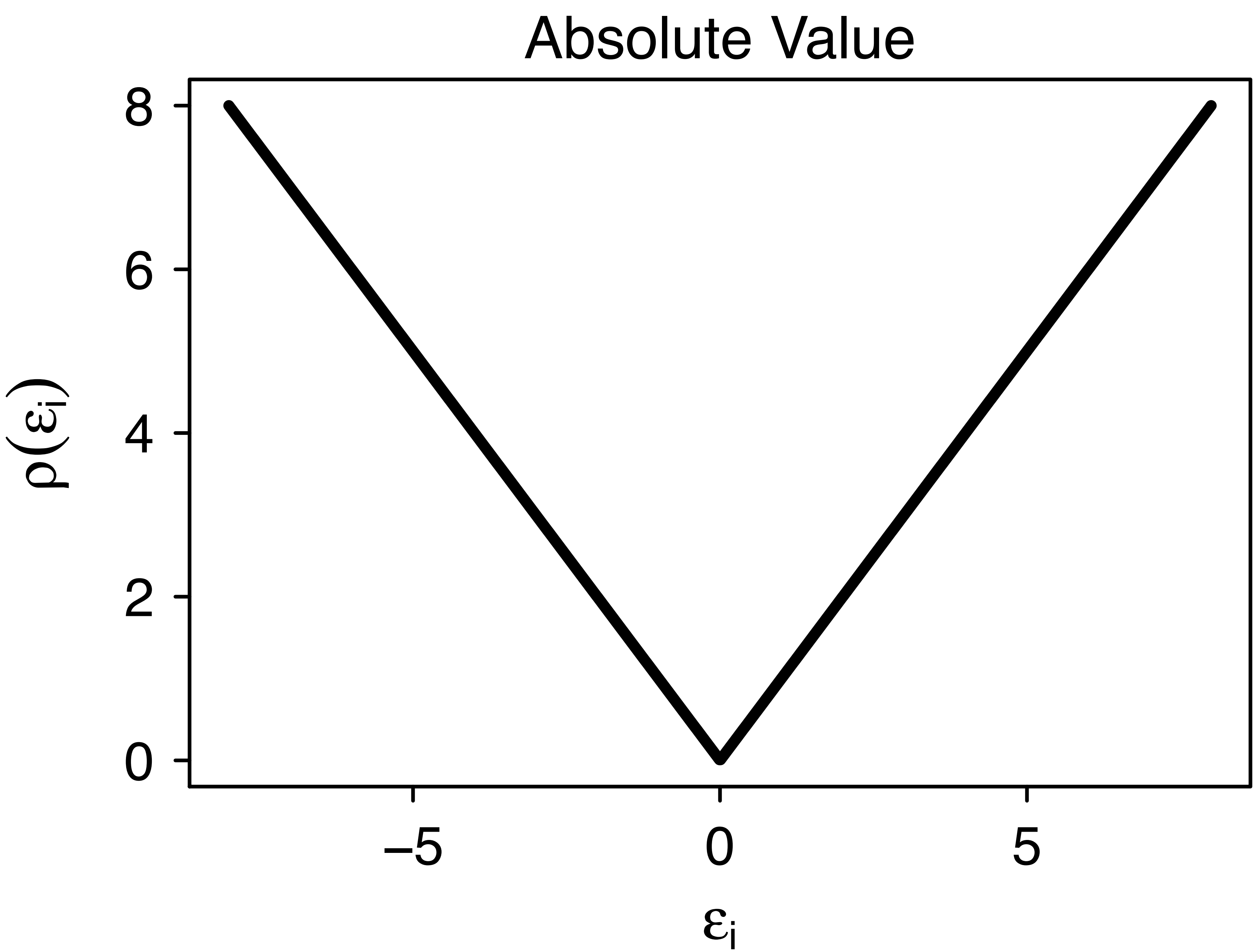
$$\hat{\beta}^{LS} = \arg\min_{b} \sum_{i=1}^{n} (y_i - X_i b)^2$$

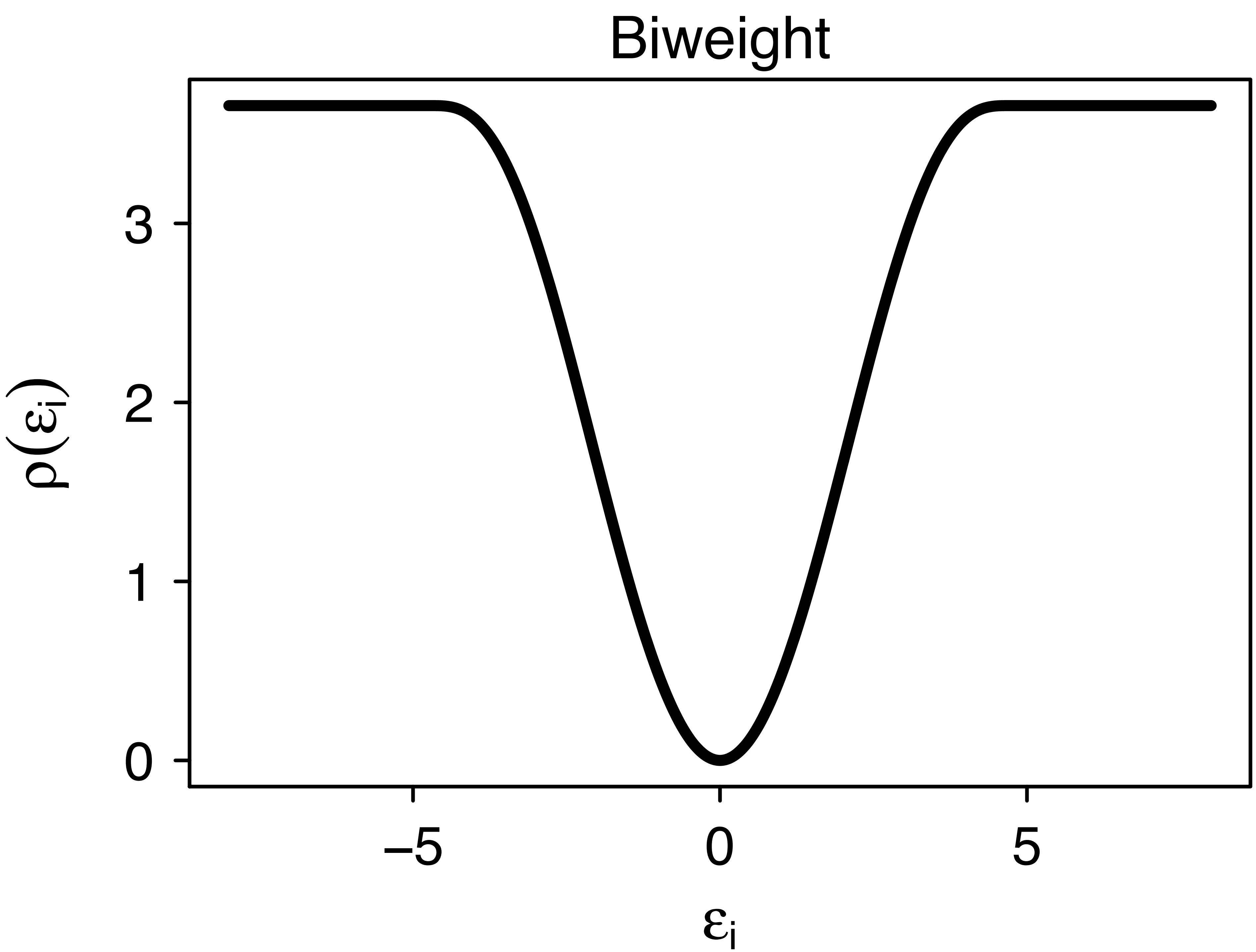
$$\hat{\beta}^{\rho} = \arg\min_{b} \sum_{i=1}^{n} \rho(y_i - X_i b)$$

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.


Square
60
40
20
0
–5
0
5
$\rho(\varepsilon_i)$
ε_i


Absolute Value
8
6
4
2
0
ρ(εi)
–5
0
5
εi

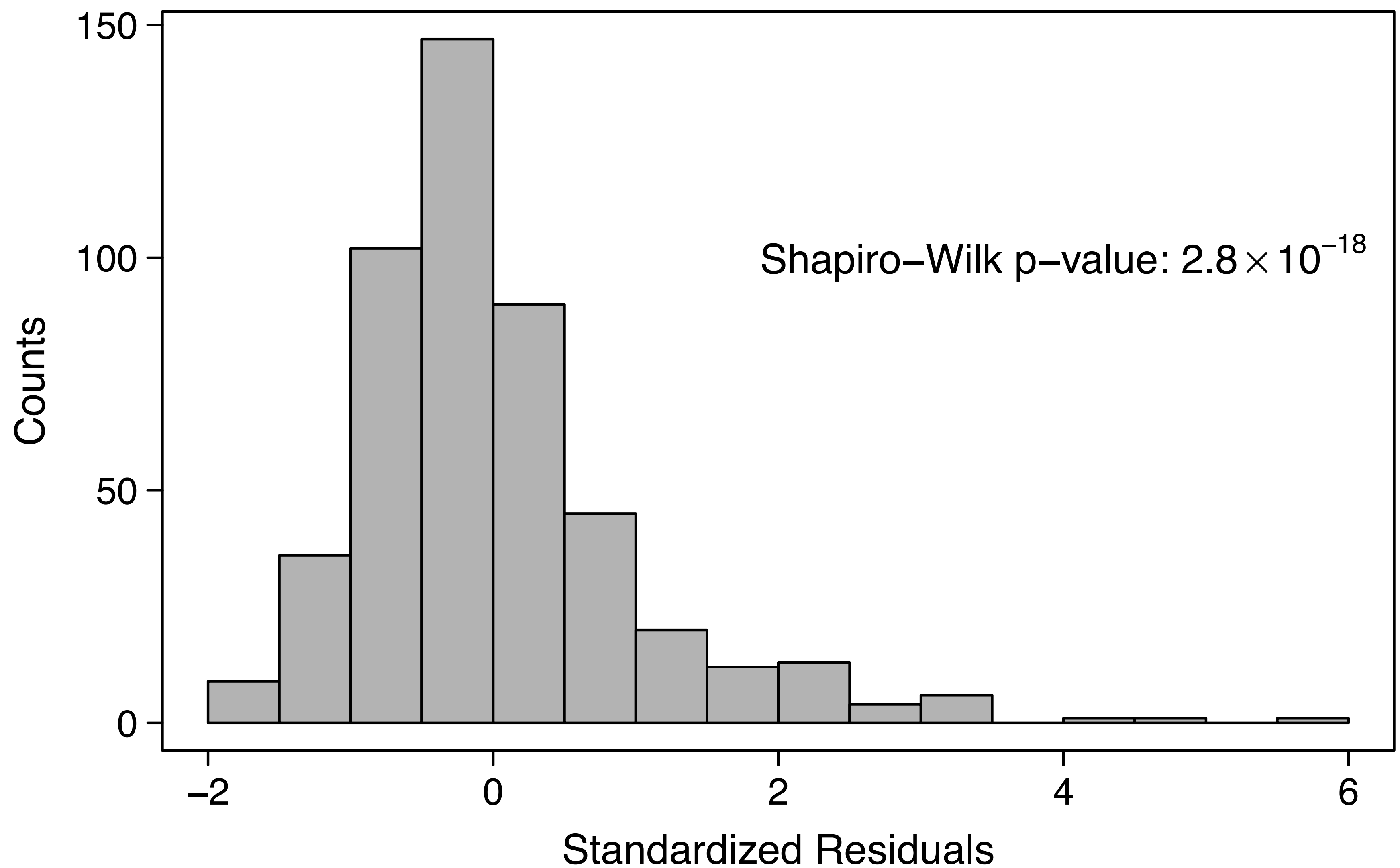


Biweight
ρ(εi)
0
1
2
3
−5
0
5
εi

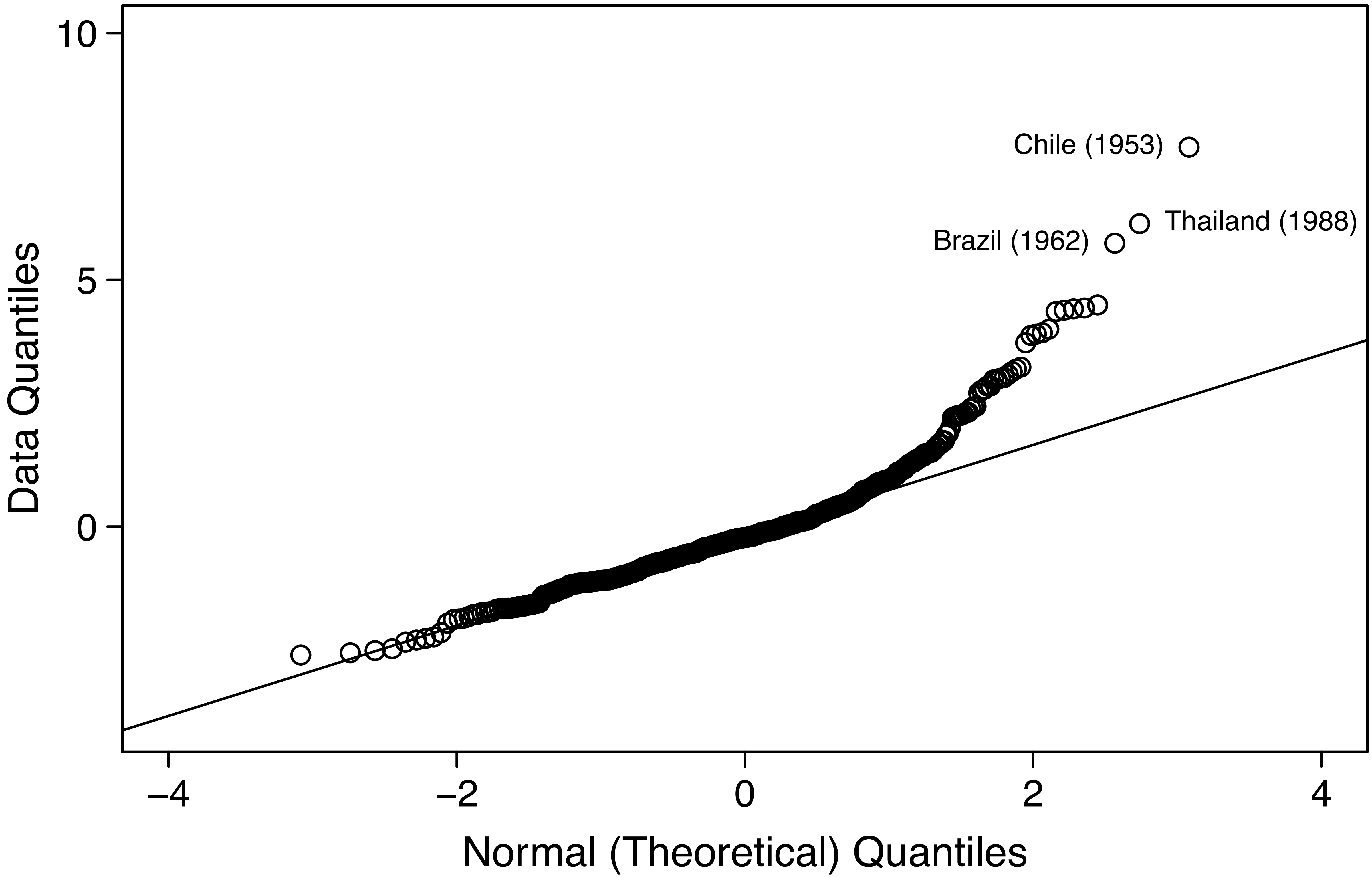
Robust estimators are often more efficient than LS.

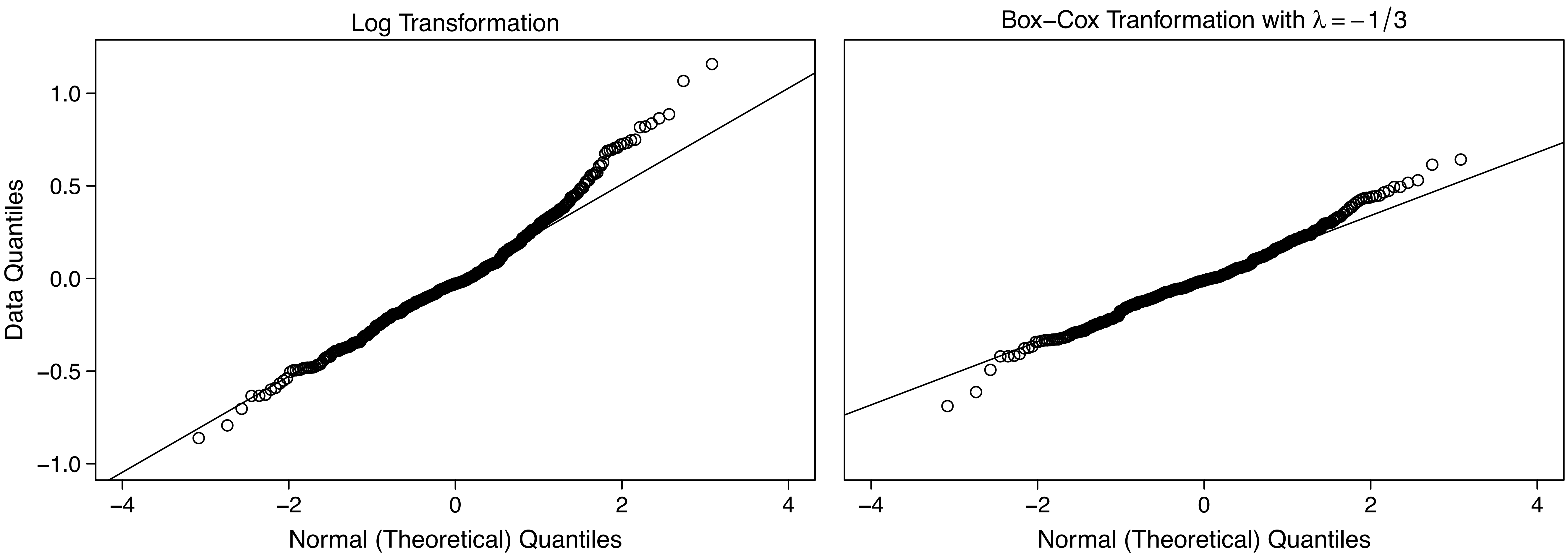
Robust estimators allow unusual cases to be unusual.

Clark and Golder

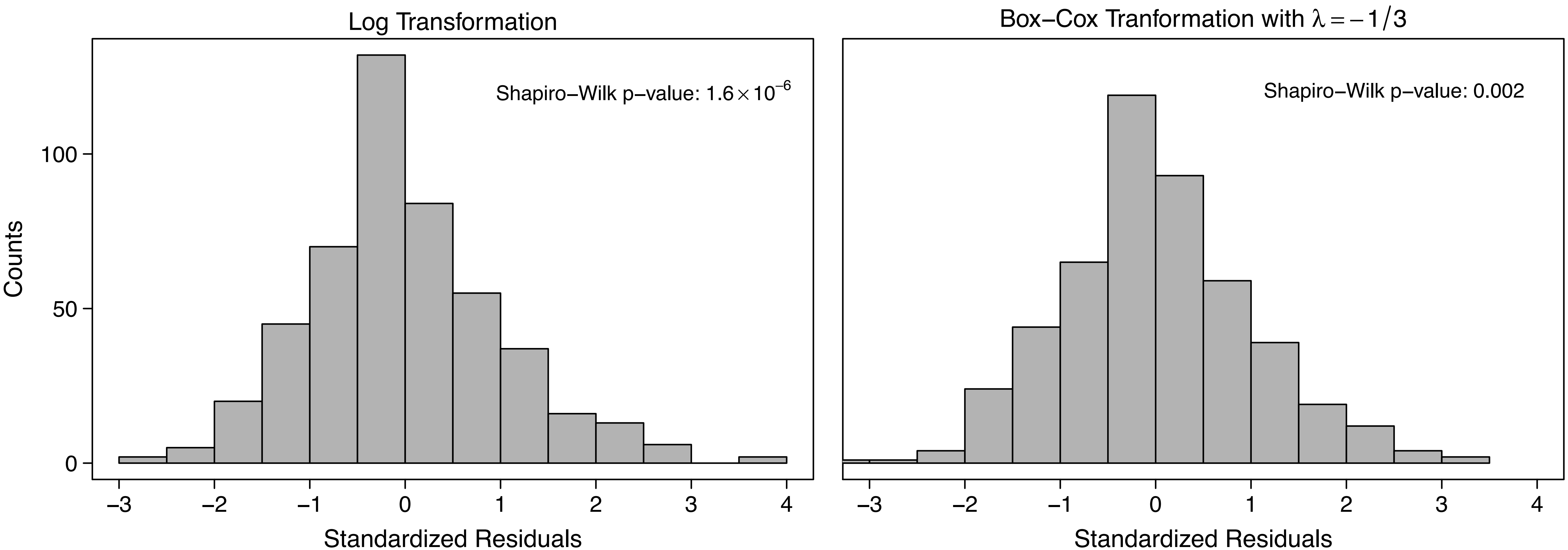


Shapiro−Wilk p−value: 2.8×10^{-18}
Counts
150
100
50
0
−2
0
2
4
6
Standardized Residuals

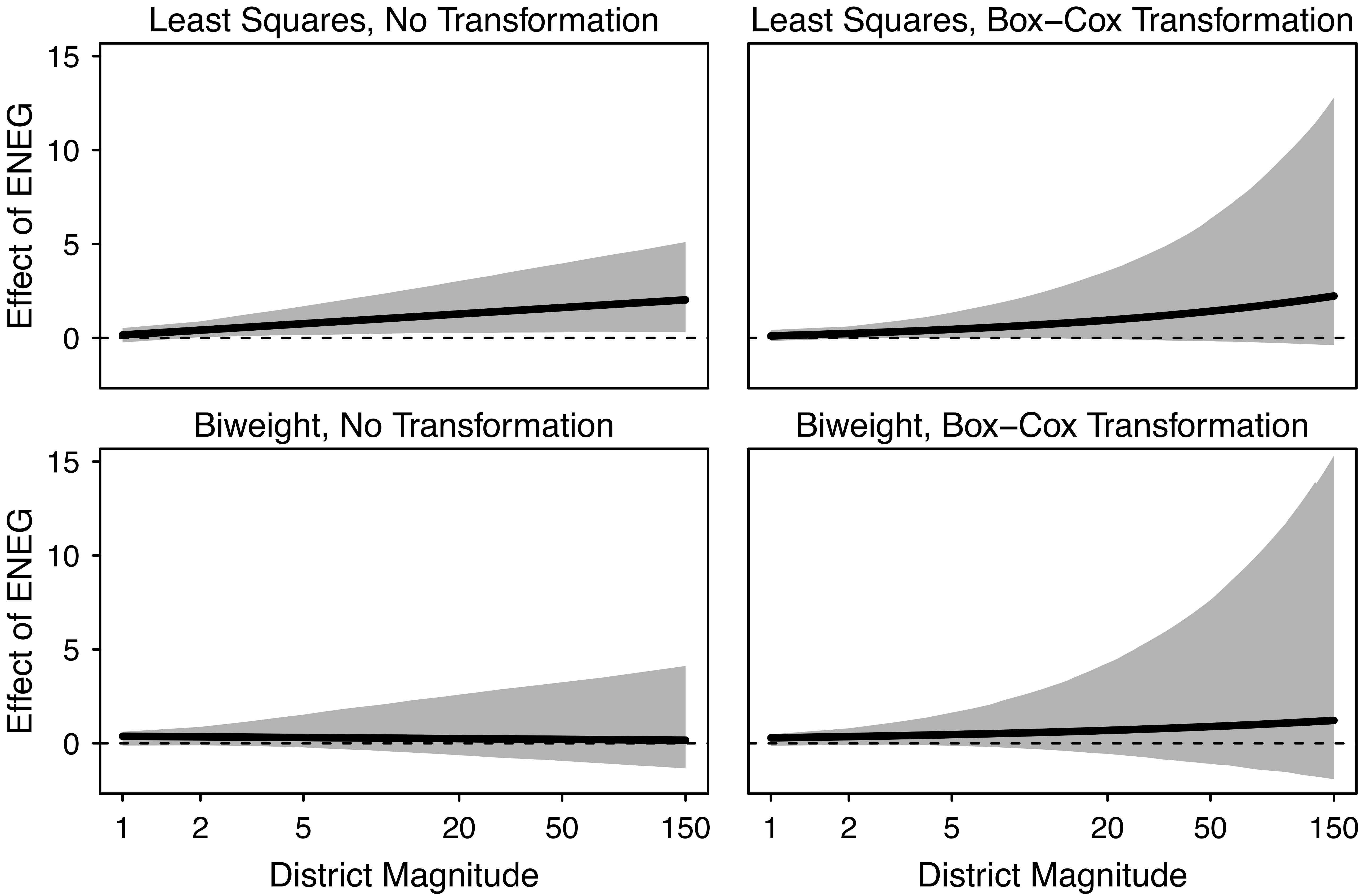

Chile (1953)
Thailand (1988)
Brazil (1962)
10
5
0
−4
−2
0
2
4
Data Quantiles
Normal (Theoretical) Quantiles



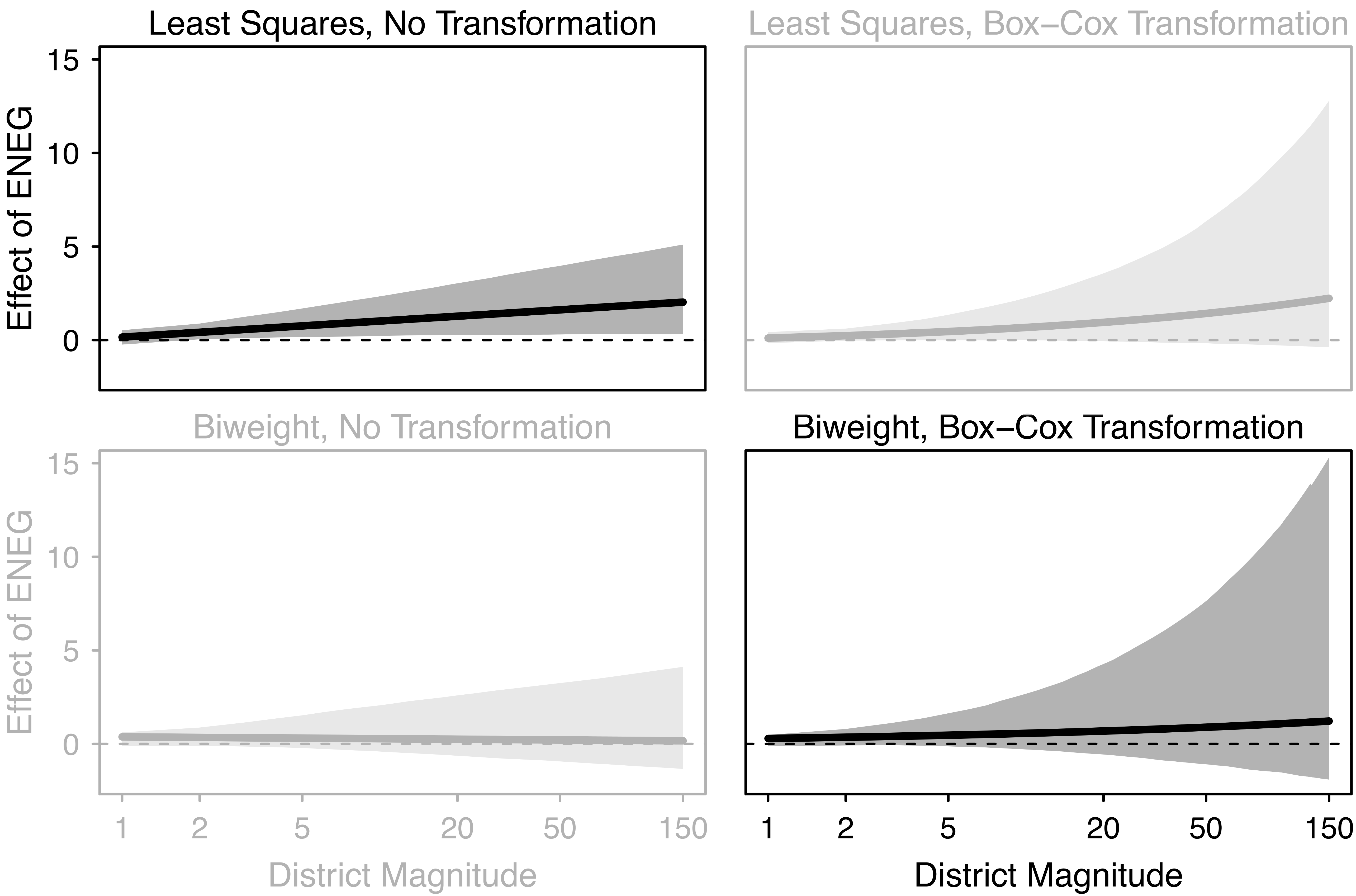
Log Transformation
Box−Cox Tranformation with $\lambda = -1/3$
Data Quantiles
1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4
Normal (Theoretical) Quantiles



Log Transformation
Shapiro−Wilk p−value: 1.6×10^{-6}
Counts
0
50
100
−3
−2
−1
0
1
2
3
4
Standardized Residuals
Box−Cox Tranformation with $\lambda = -1/3$
Shapiro−Wilk p−value: 0.002
−3
−2
−1
0
1
2
3
4
Standardized Residuals



Least Squares, No Transformation
Least Squares, Box−Cox Transformation
Biweight, No Transformation
Biweight, Box−Cox Transformation
Effect of ENEG
15
10
5
0
1
2
5
20
50
150
District Magnitude



Least Squares, No Transformation
Least Squares, Box−Cox Transformation
Biweight, No Transformation
Biweight, Box−Cox Transformation
Effect of ENEG
Effect of ENEG
15
10
5
0
15
10
5
0
1
2
5
20
50
150
1
2
5
20
50
150
District Magnitude
District Magnitude

Substantive Takaways

Substantive Takaways

The theory is wrong.

Substantive Takaways

The theory is wrong.

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

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

Substantive Takaways

The theory is wrong.

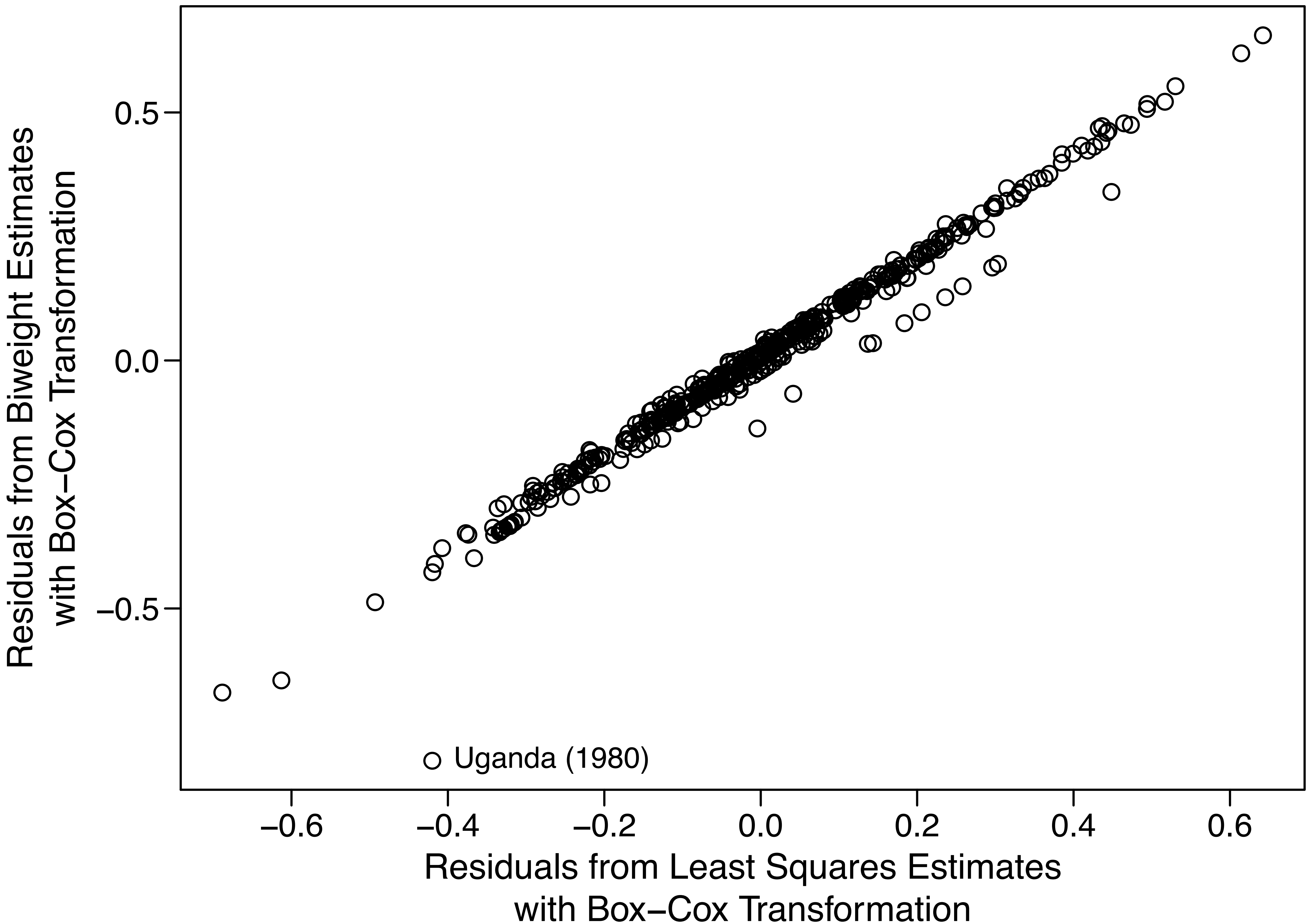
The estimates are suggest the effects might be smaller *or larger* than Clark and Golder's analysis suggests.

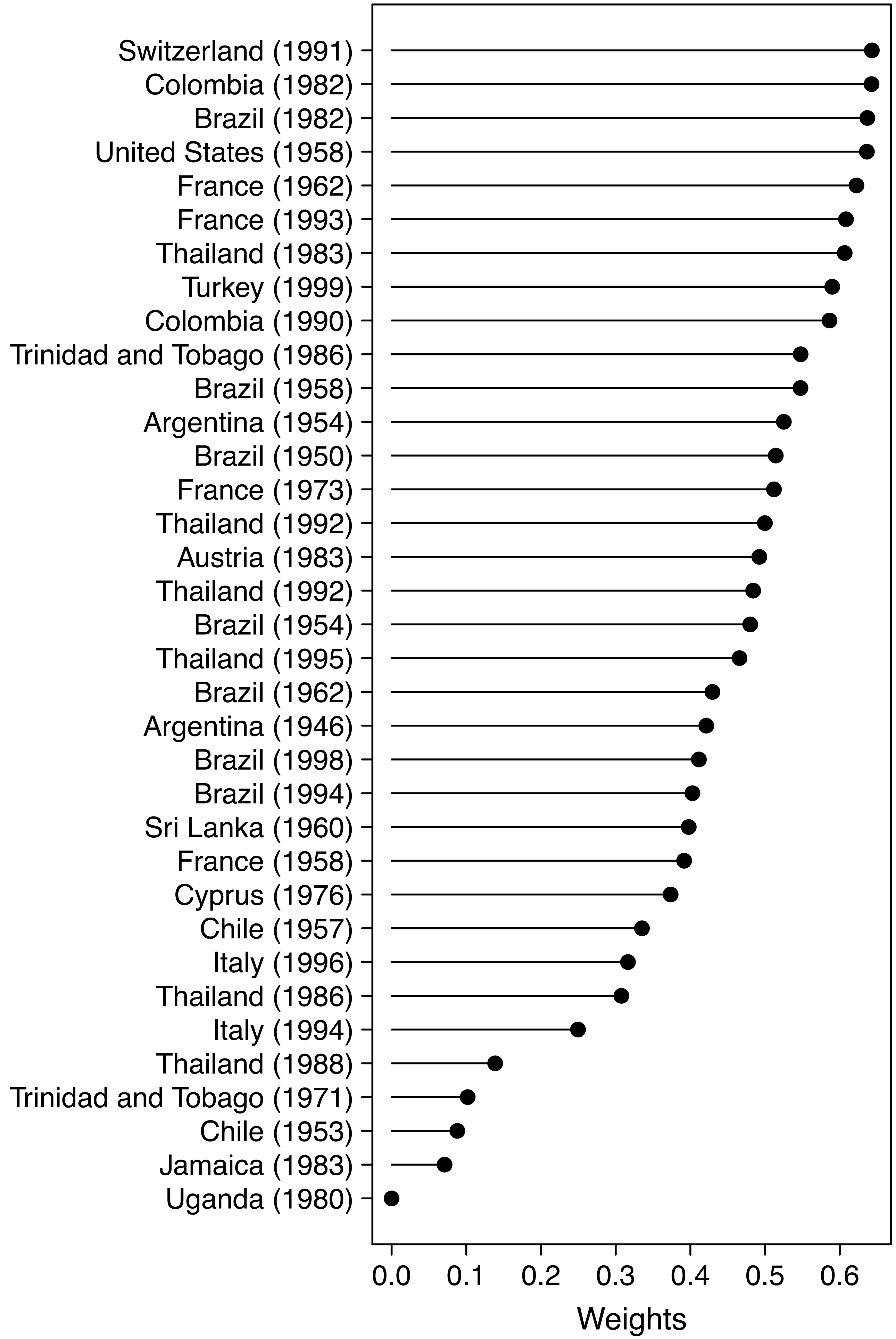
Substantive Takaways

~~The theory is wrong.~~

Substantive Takaways

We can learn from the residuals.


Residuals from Biweight Estimates
with Box−Cox Transformation
0.5
0.0
−0.5
Uganda (1980)
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
Residuals from Least Squares Estimates
with Box−Cox Transformation



Switzerland (1991)
Colombia (1982)
Brazil (1982)
United States (1958)
France (1962)
France (1993)
Thailand (1983)
Turkey (1999)
Colombia (1990)
Trinidad and Tobago (1986)
Brazil (1958)
Argentina (1954)
Brazil (1950)
France (1973)
Thailand (1992)
Austria (1983)
Thailand (1992)
Brazil (1954)
Thailand (1995)
Brazil (1962)
Argentina (1946)
Brazil (1998)
Brazil (1994)
Sri Lanka (1960)
France (1958)
Cyprus (1976)
Chile (1957)
Italy (1996)
Thailand (1986)
Italy (1994)
Thailand (1988)
Trinidad and Tobago (1971)
Chile (1953)
Jamaica (1983)
Uganda (1980)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Weights

Substantive Takaways

The 1980 election in Uganda

Substantive Takaways

What is an “established democracy”?

Substantive Takaways

What dynamics lead to equilibrium?

Substantive Takaways

How do these dynamics depend on the prior regime?

Key Points

Point #1

Normality is an important assumption of least squares.

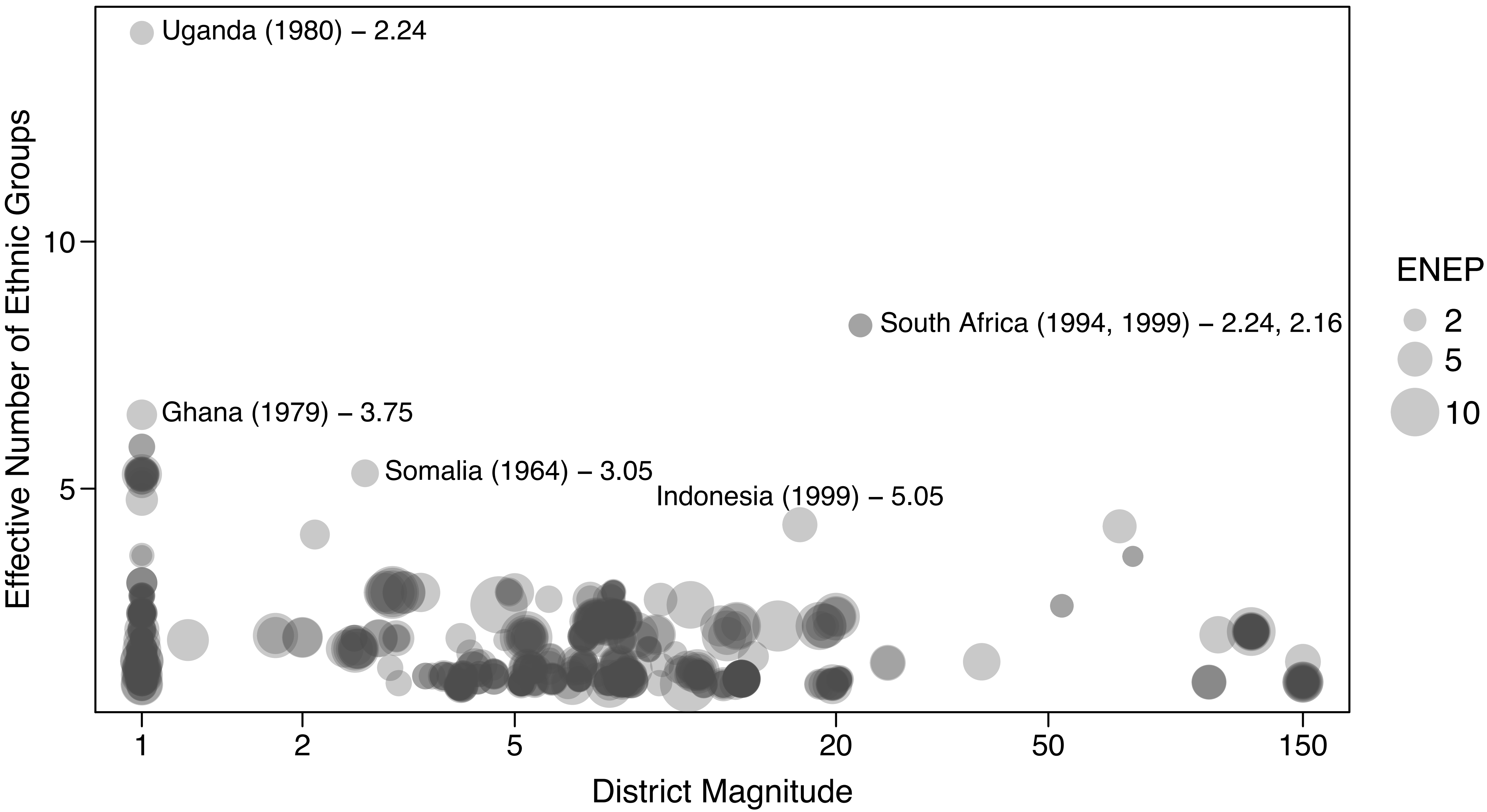
Point #2

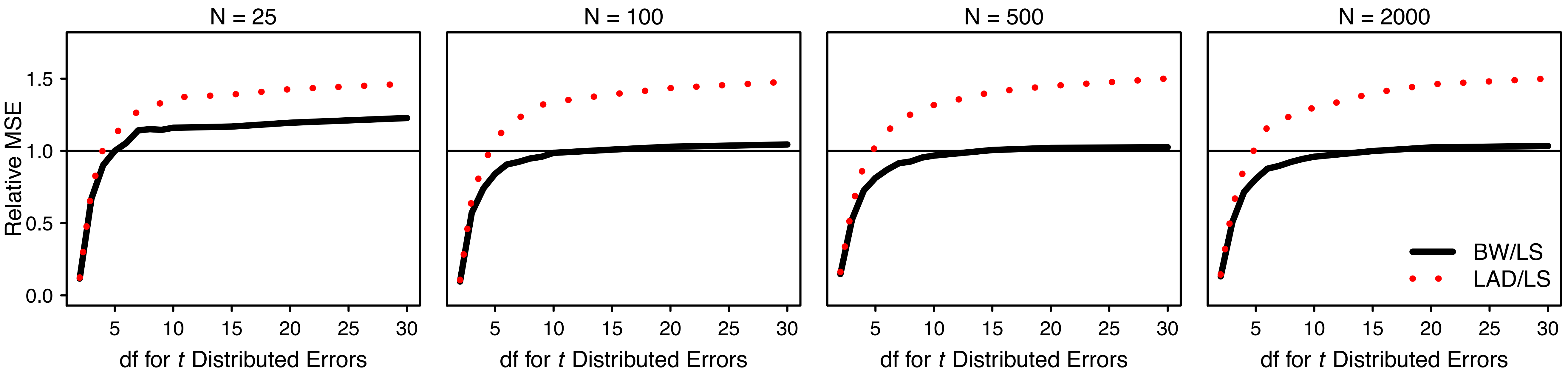
Alternatives to least squares often exhibit better behavior for non-normal errors.

Point #3

Researchers can learn much from unusual cases.

Even More!


Uganda (1980) – 2.24
South Africa (1994, 1999) – 2.24, 2.16
Ghana (1979) – 3.75
Somalia (1964) – 3.05
Indonesia (1999) – 5.05
Effective Number of Ethnic Groups
10
5
1
2
5
20
50
150
District Magnitude
ENEP
2
5
10


N = 25
N = 100
N = 500
N = 2000
Relative MSE
0.0
0.5
1.0
1.5
5
10
15
20
25
30
df for t Distributed Errors
BW/LS
LAD/LS

	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

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