Advanced Quantitative Research Methodology, Lecture Notes: Robust Standard Errors[1]

Gary King

Institute for Quantitative Social Science
Harvard University

GaryKing.org

March 9, 2015

What robust SEs are and are not

What robust SEs are and are not

- Are: a way to estimate $V(\hat{\theta})$ with fewer assumptions

What robust SEs are and are not

- Are: a way to estimate $V(\hat{\theta})$ with fewer assumptions
- Are Not: A way to estimate $V(\hat{\theta})$ without any assumptions

What robust SEs are and are not

- Are: a way to estimate $V(\hat{\theta})$ with fewer assumptions
- Are Not: A way to estimate $V(\hat{\theta})$ without any assumptions
- Are: useful when some assumptions are violated, but others are not (Hard to verify)

What robust SEs are and are not

- Are: a way to estimate $V(\hat{\theta})$ with fewer assumptions
- Are Not: A way to estimate $V(\hat{\theta})$ without any assumptions
- Are: useful when some assumptions are violated, but others are not (Hard to verify)
- Are Not: A way to inoculate yourself from criticism

What robust SEs are and are not

- Are: a way to estimate $V(\hat{\theta})$ with fewer assumptions
- Are Not: A way to estimate $V(\hat{\theta})$ without any assumptions
- Are: useful when some assumptions are violated, but others are not (Hard to verify)
- Are Not: A way to inoculate yourself from criticism
- Are: A good test for misspecification

Overview of Robust SEs

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with $1000+/$month more)

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with 1000+/month more)
- Widely misused! RSEs solve one problem, and reveal others

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with $1000+$/month more)
- Widely misused! RSEs solve one problem, and reveal others
- Authors act as if RSEs inoculate them from criticism; in fact, they highlight problems

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with 1000+/month more)
- Widely misused! RSEs solve one problem, and reveal others
- Authors act as if RSEs inoculate them from criticism; in fact, they highlight problems
- Implications for estimates when SEs and RSEs differ:

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with $1000+$/month more)
- Widely misused! RSEs solve one problem, and reveal others
- Authors act as if RSEs inoculate them from criticism; in fact, they highlight problems
- Implications for estimates when SEs and RSEs differ:
  - Best case

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with 1000+/month more)
- Widely misused! RSEs solve one problem, and reveal others
- Authors act as if RSEs inoculate them from criticism; in fact, they highlight problems
- Implications for estimates when SEs and RSEs differ:
  - Best case
    - Some QOIs (β in regression): unbiased but inefficient

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with 1000+/month more)
- Widely misused! RSEs solve one problem, and reveal others
- Authors act as if RSEs inoculate them from criticism; in fact, they highlight problems
- Implications for estimates when SEs and RSEs differ:
  - Best case
    - Some QOIs (β in regression): unbiased but inefficient
    - Other QOIs ($\Pr(Y > 0.6)$): biased

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with $1000+$/month more)
- Widely misused! RSEs solve one problem, and reveal others
- Authors act as if RSEs inoculate them from criticism; in fact, they highlight problems
- Implications for estimates when SEs and RSEs differ:
    - Best case
        - Some QOIs (β in regression): unbiased but inefficient
        - Other QOIs ($\Pr(Y > 0.6)$): biased
    - Worst case:

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with $1000+$/month more)
- Widely misused! RSEs solve one problem, and reveal others
- Authors act as if RSEs inoculate them from criticism; in fact, they highlight problems
- Implications for estimates when SEs and RSEs differ:
  - Best case
    - Some QOIs (β in regression): unbiased but inefficient
    - Other QOIs ($\Pr(Y > 0.6)$): biased
  - Worst case:
    - Misspecification is more widespread

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with 1000+/month more)
- Widely misused! RSEs solve one problem, and reveal others
- Authors act as if RSEs inoculate them from criticism; in fact, they highlight problems
- Implications for estimates when SEs and RSEs differ:
  - Best case
    - Some QOIs (β in regression): unbiased but inefficient
    - Other QOIs ($\Pr(Y > 0.6)$): biased
  - Worst case:
    - Misspecification is more widespread
    - All QOIs are biased

Overview of Robust SEs

- Alternative estimate of the variance matrix (and SEs)
- Consistent even under some forms of model misspecification
- Extremely widely used (most polisci articles using regression; $> 75,000$ cites in Google scholar, with 1000+/month more)
- Widely misused! RSEs solve one problem, and reveal others
- Authors act as if RSEs inoculate them from criticism; in fact, they highlight problems
- Implications for estimates when SEs and RSEs differ:
  - Best case
    - Some QOIs (β in regression): unbiased but inefficient
    - Other QOIs ($\Pr(Y > 0.6)$): biased
  - Worst case:
    - Misspecification is more widespread
    - All QOIs are biased
- A better use for RSEs: as a test for misspecification

Regression Model Variance Specification

Regression Model Variance Specification

- Linear-normal regression model:

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

$$\Sigma = \begin{pmatrix} \sigma^2_{11} & \sigma^2_{12} & \dots & \sigma^2_{1n} \\ \sigma^2_{21} & \sigma^2_{22} & \dots & \sigma^2_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{n1} & \sigma^2_{n2} & \dots & \sigma^2_{nn} \end{pmatrix}$$

Variance matrix unconstrained

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

$$\Sigma = \begin{pmatrix} \sigma^2_{11} & \sigma^2_{12} & \dots & \sigma^2_{1n} \\ \sigma^2_{21} & \sigma^2_{22} & \dots & \sigma^2_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{n1} & \sigma^2_{n2} & \dots & \sigma^2_{nn} \end{pmatrix}$$

Variance matrix unconstrained (includes covariances)

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & 0 & \dots & 0 \\ 0 & \sigma_{22}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{nn}^2 \end{pmatrix}$$

Now assume independence (set covariances to zero)

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

$$\Sigma = \begin{pmatrix} \sigma^2_{11} & 0 & \dots & 0 \\ 0 & \sigma^2_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2_{nn} \end{pmatrix}$$

Now assume independence (set covariances to zero)

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

$$\Sigma = \begin{pmatrix} \sigma^2_{11} & 0 & \dots & 0 \\ 0 & \sigma^2_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2_{nn} \end{pmatrix}$$

Still allows for heteroskedasticity

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

Now assume independence and homoskedasticity

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Now assume independence and homoskedasticity

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j | X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

$$\Sigma = \sigma^2 \underset{n\times n}{I}$$

Standard linear-normal regression model assumptions

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = \beta$

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = \beta$
- Usual estimate of $V(b) = \sigma^2 Q^{-1}$: biased

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = \beta$
- Usual estimate of $V(b) = \sigma^2 Q^{-1}$: biased
- Simulations from the model are wrong! $\rightsquigarrow$ other QOIs are biased.

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = \beta$
- Usual estimate of $V(b) = \sigma^2 Q^{-1}$: biased
- Simulations from the model are wrong! $\rightsquigarrow$ other QOIs are biased.
- True variance:
  $V(b) = V(Q^{-1}X'y) = Q^{-1}X'V(y)XQ^{-1} = Q^{-1}X'\Sigma XQ^{-1} \neq \sigma^2 I$

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = \beta$
- Usual estimate of $V(b) = \sigma^2 Q^{-1}$: biased
- Simulations from the model are wrong! $\rightsquigarrow$ other QOIs are biased.
- True variance:
  $V(b) = V(Q^{-1}X'y) = Q^{-1}X'V(y)XQ^{-1} = Q^{-1}X'\Sigma XQ^{-1} \neq \sigma^2 I$
- Estimating all the unknowns in Σ seems hopeless

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = \beta$
- Usual estimate of $V(b) = \sigma^2 Q^{-1}$: biased
- Simulations from the model are wrong! $\rightsquigarrow$ other QOIs are biased.
- True variance:
  $V(b) = V(Q^{-1}X'y) = Q^{-1}X'V(y)XQ^{-1} = Q^{-1}X'\Sigma XQ^{-1} \neq \sigma^2 I$
- Estimating all the unknowns in Σ seems hopeless
- RSE insight:

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = \beta$
- Usual estimate of $V(b) = \sigma^2 Q^{-1}$: biased
- Simulations from the model are wrong! $\rightsquigarrow$ other QOIs are biased.
- True variance:
  $V(b) = V(Q^{-1}X'y) = Q^{-1}X'V(y)XQ^{-1} = Q^{-1}X'\Sigma XQ^{-1} \neq \sigma^2 I$
- Estimating all the unknowns in Σ seems hopeless
- RSE insight:
  - $V(b) = \underset{k\times k}{Q^{-1}} \underbrace{\underset{k\times n}{X'} \underset{(n\times n)}{\Sigma} \underset{n\times k}{X}}_{k\times k} \underset{k\times k}{Q^{-1}} = Q^{-1} \underset{k\times k}{G}\, Q^{-1}$, with $k << n$

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

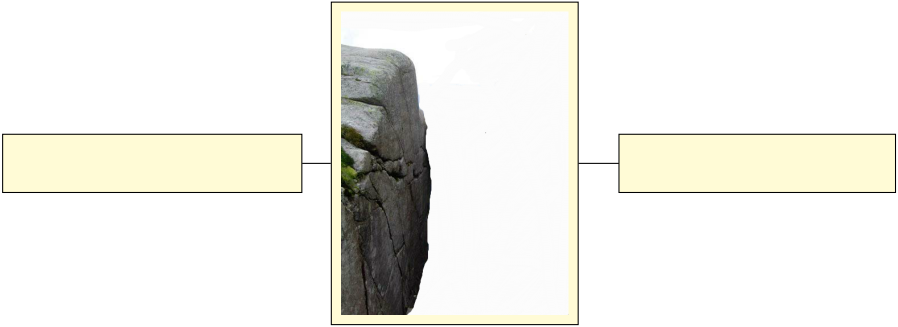
- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = \beta$
- Usual estimate of $V(b) = \sigma^2 Q^{-1}$: biased
- Simulations from the model are wrong! $\rightsquigarrow$ other QOIs are biased.
- True variance:
  $V(b) = V(Q^{-1}X'y) = Q^{-1}X'V(y)XQ^{-1} = Q^{-1}X'\Sigma XQ^{-1} \neq \sigma^2 I$
- Estimating all the unknowns in Σ seems hopeless
- RSE insight:
  - $V(b) = \underset{k\times k}{Q^{-1}} \underbrace{\underset{k\times n}{X'} \underset{(n\times n)}{\Sigma} \underset{n\times k}{X}}_{k\times k} \underset{k\times k}{Q^{-1}} = Q^{-1} \underset{k\times k}{G} Q^{-1}$, with $k << n$
  - Can estimate $V(b)$ by replacing σ_i^2 with e_i^2 in Σ

What if $\Sigma \neq \sigma^2 I$ and we run a regression?

- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = \beta$
- Usual estimate of $V(b) = \sigma^2 Q^{-1}$: biased
- Simulations from the model are wrong! $\rightsquigarrow$ other QOIs are biased.
- True variance:
  $V(b) = V(Q^{-1}X'y) = Q^{-1}X'V(y)XQ^{-1} = Q^{-1}X'\Sigma XQ^{-1} \neq \sigma^2 I$
- Estimating all the unknowns in Σ seems hopeless
- RSE insight:
  - $V(b) = \underset{k\times k}{Q^{-1}} \underbrace{\underset{k\times n}{X'} \underset{(n\times n)}{\Sigma} \underset{n\times k}{X}}_{k\times k} \underset{k\times k}{Q^{-1}} = Q^{-1} \underset{k\times k}{G} Q^{-1}$, with $k << n$
  - Can estimate $V(b)$ by replacing σ_i^2 with e_i^2 in Σ
  - The idea generalizes to any MLE model

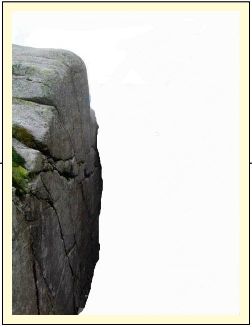
For RSEs to help: Everything has to be Juuuussttt Right

For RSEs to help: Everything has to be Juuuussttt Right



For RSEs to help: Everything has to be Juuuussttt Right

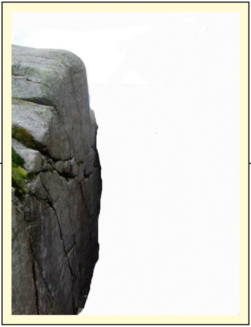
Model Misspecified



For RSEs to help: Everything has to be Juuuussttt Right

Model Misspecified

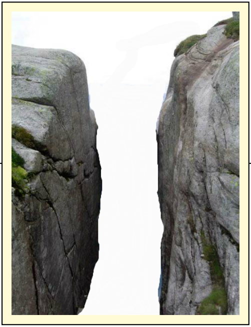
(point estimates biased)



For RSEs to help: Everything has to be Juuuussttt Right

Model Misspecified

(point estimates biased)

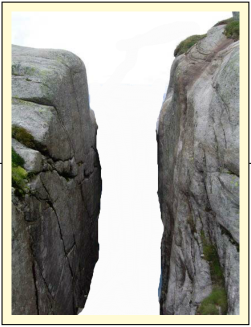


Model Correct

For RSEs to help: Everything has to be Juuuussttt Right

Model Misspecified

(point estimates biased)



Model Correct

(point estimates correct)

For RSEs to help: Everything has to be Juuuussttt Right

Goldilocks Region

Model Misspecified

(point estimates biased)

Model Correct

(point estimates correct)

For RSEs to help: Everything has to be Juuuussttt Right

Goldilocks Region

Model Misspecified

(point estimates biased)

Model Correct

(point estimates correct)

For RSEs to help: Everything has to be Juuuussttt Right

Biased just enough to make RSEs useful,

Goldilocks Region

Model Misspecified

(point estimates biased)

Model Correct

(point estimates correct)

For RSEs to help: Everything has to be Juuuussttt Right

Biased just enough to make RSEs useful,

Goldilocks Region

but not so much as to bias everything else

Model Misspecified

(point estimates biased)

Model Correct

(point estimates correct)

The Goldilocks Region is not Idyllic

The Goldilocks Region is not Idyllic

In the Goldilocks region,

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,
    - the variation in vote outcome,

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,
    - the variation in vote outcome,
    - or vote predictions with confidence intervals.

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,
    - the variation in vote outcome,
    - or vote predictions with confidence intervals.
  - We don't know: the substantive meaning of our results.

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,
    - the variation in vote outcome,
    - or vote predictions with confidence intervals.
  - We don't know: the substantive meaning of our results. How big are they, really?

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,
    - the variation in vote outcome,
    - or vote predictions with confidence intervals.
  - We don't know: the substantive meaning of our results. How big are they, really?
  - We can't check: whether model implications are realistic.

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,
    - the variation in vote outcome,
    - or vote predictions with confidence intervals.
  - We don't know: the substantive meaning of our results. How big are they, really?
  - We can't check: whether model implications are realistic.
- Parts of the model are wrong;

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,
    - the variation in vote outcome,
    - or vote predictions with confidence intervals.
  - We don't know: the substantive meaning of our results. How big are they, really?
  - We can't check: whether model implications are realistic.
- Parts of the model are wrong; why do we think the rest is right?

The Goldilocks Region is not Idyllic

In the Goldilocks region,

- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,
    - the variation in vote outcome,
    - or vote predictions with confidence intervals.
  - We don't know: the substantive meaning of our results. How big are they, really?
  - We can't check: whether model implications are realistic.
- Parts of the model are wrong; why do we think the rest is right?
- If SE$\neq$RSE: find missspecification, fix model, rerun until SE$=$RSE

Replication of Neumayer (ISQ, 2003)

Replication of Neumayer (ISQ, 2003)

- "Multilateral aid flows. . . exhibit a bias toward less populous countries."

Replication of Neumayer (ISQ, 2003)

- "Multilateral aid flows... exhibit a bias toward less populous countries."
- Linear regression: multilateral aid flows on log-population, squared log-population, and control variables

- "Multilateral aid flows... exhibit a bias toward less populous countries."
- Linear regression: multilateral aid flows on log-population, squared log-population, and control variables
- Coefficient on log-population: -3.13

Replication of Neumayer (ISQ, 2003)

- "Multilateral aid flows. . . exhibit a bias toward less populous countries."
- Linear regression: multilateral aid flows on log-population, squared log-population, and control variables
- Coefficient on log-population: -3.13
- Robust SE (0.72) twice classical SE (0.37)

Replication of Neumayer (ISQ, 2003)

- "Multilateral aid flows...exhibit a bias toward less populous countries."
- Linear regression: multilateral aid flows on log-population, squared log-population, and control variables
- Coefficient on log-population: -3.13
- Robust SE (0.72) twice classical SE (0.37)
- Clear evidence of misspecification

Replication of Neumayer (ISQ, 2003)

- "Multilateral aid flows. . . exhibit a bias toward less populous countries."
- Linear regression: multilateral aid flows on log-population, squared log-population, and control variables
- Coefficient on log-population: -3.13
- Robust SE (0.72) twice classical SE (0.37)
- Clear evidence of misspecification
- To correct it, we take the Box-Cox transformation of Y (like a log)

Replication of Neumayer (ISQ, 2003)

- "Multilateral aid flows. . . exhibit a bias toward less populous countries."
- Linear regression: multilateral aid flows on log-population, squared log-population, and control variables
- Coefficient on log-population: -3.13
- Robust SE (0.72) twice classical SE (0.37)
- Clear evidence of misspecification
- To correct it, we take the Box-Cox transformation of Y (like a log)
- Robust SE (0.34) $\approx$ classical SE (0.32)

Transformation of Y: Becomes Normal

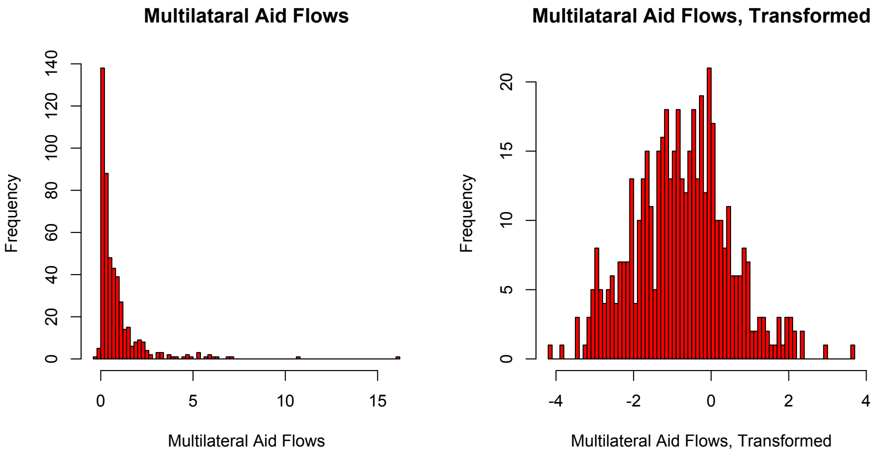
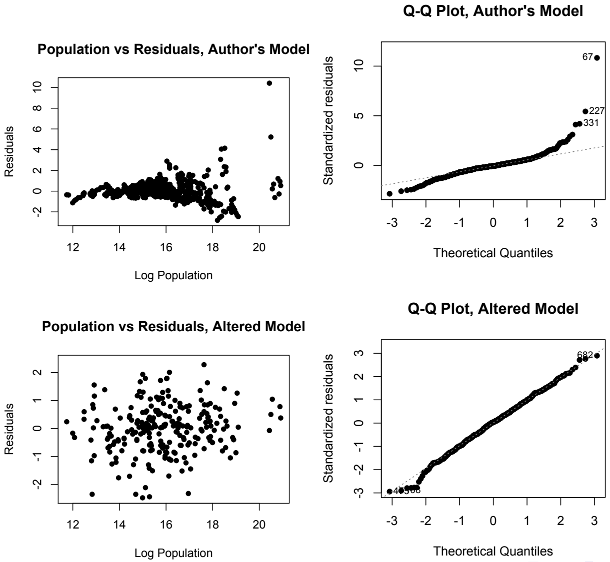


Figure: Distribution of the dependent variable before (left) and after (right) the Box-Cox transformation.

Transformation of Y: Removes Heteroskedasticty



Results: Transformed Model, Opposite Results

The misspecification $\rightsquigarrow$ bias, not inefficiency