Data Transformations for Social Science Research: Theory and Best Practices[†]

Mark Shadden
Department of Political Science
Pennsylvania State University
mark.shadden@psu.edu

Christopher Zorn
Department of Political Science
Pennsylvania State University
zorn@psu.edu

Abstract

Researchers using regression models of the form $Y = f(\mathbf{X}\beta)$ often use a logarithmic transformation of covariates. Such transformations can be problematic when the covariate contains zero or negative values, and no clear consensus exists on best practices for such circumstances. We evaluate the statistical properties of several commonly-suggested solutions. For circumstances where $X_i = 0$ we show that the most widely-used approaches lead to bias, and that a straightforward, flexible solution is available. For $X_i < 0$, our results underscore the importance of selecting the correct form of the relationship between Y and X.

Transforming Zero Values

Logarithmic transformations of nonnegative covariates are problematic when the variable X contains values of zero. Because the logarithm of zero is undefined, researchers typically resort to *ad hoc* solutions, including dropping observations with $X_i = 0$, adding 1.0, or adding an arbitrary "start" value ϵ; in the latter cases, analysts vary in whether to increment *only* those observations with $X_i = 0$ or *all* observations in the data. A suggested alternative approach is to estimate a model of the form:

$$Y_i = \beta_0 + \beta_1 \ln(Z_i) + \beta_2 D_i + u_i$$

where $Z_i = \ln(X_i + 1)$ and $D_i = 1$ when $X_i = 0$ and zero otherwise; we refer to this as the "plus dummy" specification.

Study Design

We begin by simulating data ($N = 1000$) according to:

$$\begin{aligned} X_i^* &\sim U(-5, 5) \\ X_i &= \exp(X_i^*) \in (0, 150) \\ u_i &\sim N(0, 4) \\ Y_i &= 0 + 2\ln(X_i) + u_i \\ Z_i &= \begin{cases} X_i \text{ if } X_i^* \geq \tau \\ 0 \text{ if } X_i^* < \tau, \end{cases} \\ \tau &\in \{-4.5, -4, -2.5, 0\}; \end{aligned}$$

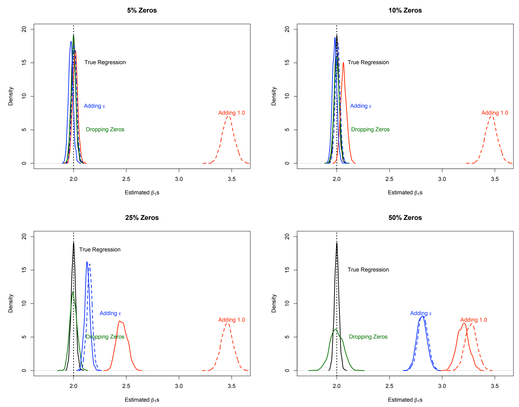
the latter corresponds to covariates with 5% / 10% / 25% / 50% zero values. We then estimate:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \ln X_i \text{ ("true")} && \text{(A.1)} \\ Y_i &= \beta_0 + \beta_1 \ln Z_i \text{ (dropping zeros)} && \text{(A.2)} \\ Y_i &= \beta_0 + \beta_1 \ln(Z_i + 1) \text{ ("add one")} && \text{(A.3)} \\ Y_i &= \beta_0 + \beta_1 \ln(Z_i + \epsilon) \text{ ("add } \epsilon\text{")} && \text{(A.4)} \\ Y_i &= \beta_0 + \beta_1 \ln(Z_i + 1) + \beta_2 D_i \text{ ("plus dummy")}, && \text{(A.5)} \end{aligned}$$

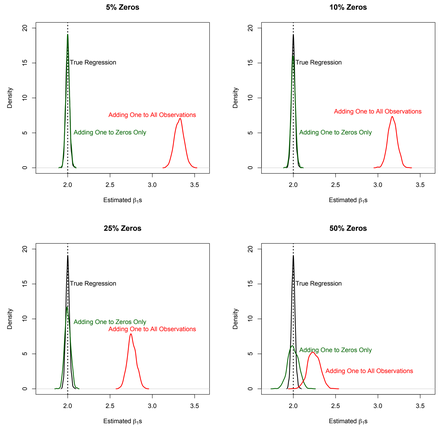
defining $\epsilon = \frac{1}{2}[0 + \min(Z|Z > 0)]$. Each model in (A.3)-(A.5) is analyzed with modifications to both $X_i = 0$ observations only and to all observations in the data.

Results

The figures below present density plots of 1000 estimates of $\hat{\beta}_1$ from (A.1) - (A.4). Black lines plot "true" estimates, green lines are estimated dropping zeros, red lines are analyses adding 1.0, and blue lines are adding ϵ; for the latter two, solid lines indicate modification of $X_i = 0$ observations only, while dotted lines are for modification of all observations. Similar results are found for the bias in the estimated intercepts $\hat{\beta}_0$.



By contrast, estimates from the "plus dummy" model in (A.5), particularly in combination with modification of the "zeros" observations only, exhibits no bias at any level of zero prevalence.

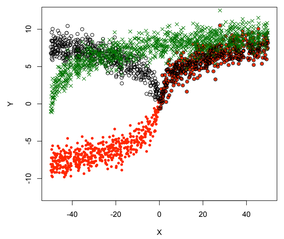


The implications are clear: *When using a log transformation of a covariate with zero values, (a) increment only those observations for which $X_i = 0$, and (b) include an indicator variable for those observations in the model specification.*

Transforming Negative Values

Logging in the negative space is a theoretical problem, not measurement error. One must precisely define the relationship between Y and $\mathbf{X}$ in the positive *and* negative spaces, and correctly choose the corresponding functional form.

Theoretical Relationships



Functional Forms

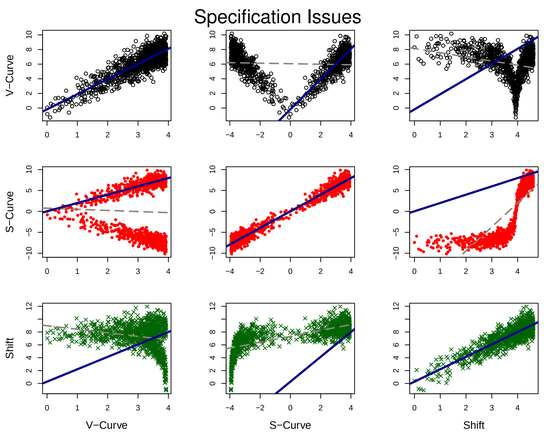
V-Curve: log absolute value.

S-Curve: log absolute value, return sign.

Shift: shift data to a start.

Truncate: drop zeros and negative values.

The figure below illustrates the potential for specification bias. Rows correspond to "true" data generating processes of the forms described above, while columns reflect transformations of each; solid blue lines reflect the correct relationship between Y and X, while dashed grey lines are linear fits.



Simulations

To further investigate, we simulate data ($N = 1000$) according to:

$$X_i^* \sim \begin{cases} U(-5, 95) \\ U(-10, 90) \\ U(-25, 75) \\ U(-50, 50) \end{cases}$$

$$\begin{aligned} X_{i\text{vc}} &= \ln(|X_i^*| + c) \\ X_{i\text{sc}} &= \text{sign}(X_i^*) \times [\ln(|X_i^*| + c)] \\ X_{i\text{shift}} &= \ln(X_i^* + [|\min(X_i^*)| + c]) \end{aligned}$$

$$\begin{aligned} u_i &\sim N(0, 1) \\ Y_{ik} &= 0 + 2\ln(X_{ik}) + u_i \text{ for } k = (\text{vc}, \text{sc}, \text{shift}) \end{aligned}$$

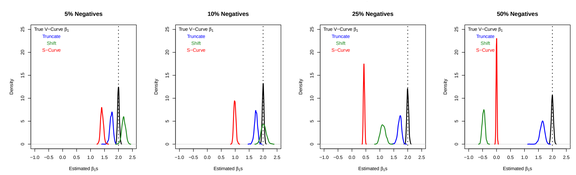
with $c = 0.01$ or 1. We then estimate:

$$\begin{aligned} Y_{ik} &= \beta_0 + \beta_1 \ln X_{ik} \text{ ("true" model)} && \text{(B.1)} \\ Y_{ik} &= \beta_0 + \beta_1 \ln X_i^* \text{ (dropping } X_i \leq 0) && \text{(B.2)} \\ Y_{ik} &= \beta_0 + \beta_1 \ln(X_{ij}),\ j \neq k \text{ (for each alternative model)} && \text{(B.3)} \end{aligned}$$

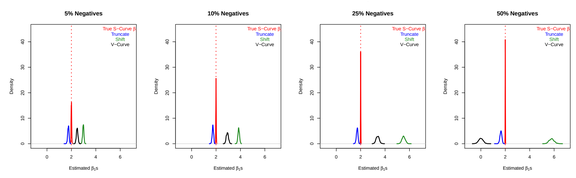
Results

The plots below illustrate the biases in $\hat{\beta}_1$ associated with each (mis)specification described in (B.1)-(B.3) for 5%, 10%, 25% and 50% prevalences of negative values.

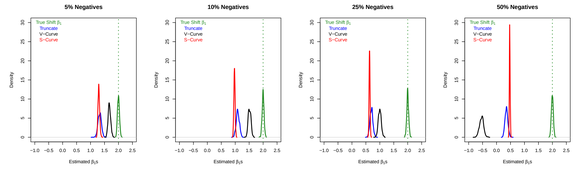
True V-Curve: Bias in Slopes



True S-Curve: Bias in Slopes



True Shift: Bias in Slopes



While the "true" models consistently and accurately recover the underlying parameters, models which truncate (drop) negative-valued observations suffer from attenuation bias, with the largest biases occurring in the "shift" relationship. Mismatching functional forms is seen to lead to biases that are both large and of indeterminate direction.

Summary

While logarithmic transformations of covariates are among the most widely used in applied research, methods for implementing such transformations when variables have non-positive values are both ad hoc and inconsistent. Our simulations demonstrate that, for both zero- and negative-valued Xs, many commonly-recommended approaches can yield substantial biases in parameter estimates. In both instances, the significance of making a correct choice grows in proportion to the fraction of observations with zero and/or negative values of X. In the case of covariates with zeros, the combination of incrementing the zero-valued observations and including an indicator variable for those observations is seen to eliminate the potential bias in more commonly-used methods. For logarithmic transformations of covariates where negative values are present, our findings make clear that correct specification of the $X - Y$ relationship is of paramount importance, and that the magnitude of bias due to misspecification is potentially quite large.

[†]This project is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0750756 and the College of Liberal Arts at Penn State. This work is ongoing; comments are very, very welcome.