

# Divide and Rule: Improving the Use of Ratio Variables

Bernd Beber

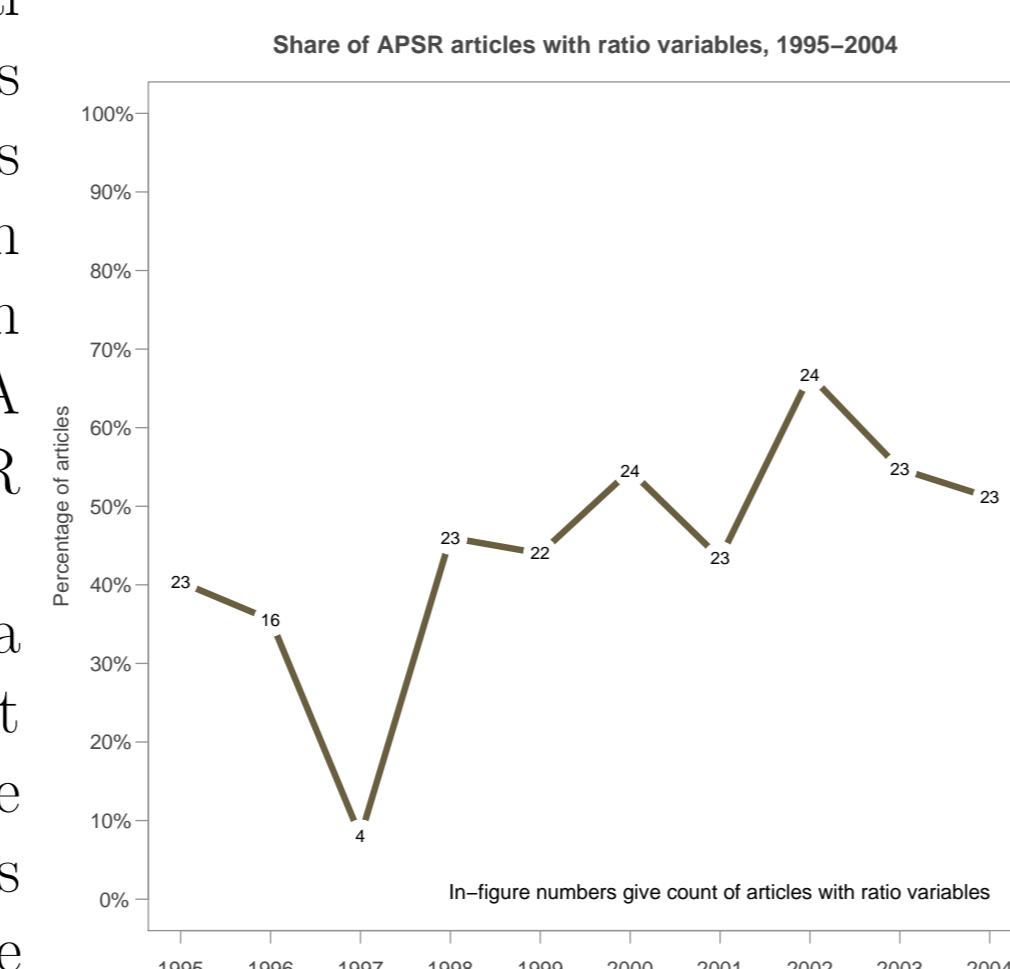
Ph.D. candidate, Department of Political Science, Columbia University<sup>†</sup>

## Abstract

Political scientists frequently employ ratio variables such as income per capita, vote share, or crime rates in their statistical analyses. Using ratios has two key advantages, in particular if a model is specified in ratio form with common components: First, akin to weighted least squares, ratio models can help reduce heteroskedasticity if the common component or scale variable is correlated with the disturbance term in a non-ratio specification. Second, ratio specifications usually require fewer degrees of freedom, which translates into additional efficiency gains. There are drawbacks, however, if a ratio specification is not appropriate, and for at least twenty years political scientists have had advice on properly specifying a model in ratio form: Do not use correlations (which do not model an intercept) with ratio variables that include common components; be wary of ratio variables if the scale variable has severe measurement error; avoid proportions (e.g. urban divided by total population) and use ratios of non-nested components instead (e.g. urban divided by non-urban population); include the inverse of the scale variable as a regressor if it is statistically significant; and verify that the ratio form specification does not in fact induce heteroskedasticity. This poster makes three contributions to this literature. First, it documents the extent to which these pieces of advice have been neglected in articles published in APSR within the last ten years. Second, much of the literature on ratio variables is preoccupied with linear models, while my analysis extends to fractional logit. Third, virtually all of the research in this area assumes that we know the true data-generating model that is to be estimated, but I argue that we often cannot know ex ante whether a specification with ratio variables is appropriate because we are uncertain about how the scale variable enters the data-generating process. The solution I suggest is to first estimate an unrestricted model that is agnostic as to whether the scale variable should enter the model directly as a predictor or by way of a ratio, and then examine the residuals over the scale variable as well as the joint significance of particular groups of coefficients in order to determine whether the model can safely be estimated in a more efficient ratio form. I use Monte Carlo simulations to show that this approach is appropriate in the face of specification uncertainty.

## A Survey of Advice Not Taken

Political scientists often use ratio variables in their statistical work, be it as an outcome variable, covariate, or both. Examples include voter turnout, wells per square mile, military spending as a share of the federal budget, among many others. The first graph shows that the proportion of articles published in the American Political Science Review that contain ratio variables is large. A total of 472 articles—all full-length articles published in APSR from 1995 to 2004—were searched.



Questions about the proper use of ratio variables motivated a sizable literature, particularly in the mid- to late 1980s. We list some key insights from this literature below and examine the extent to which statistical practice in political science reflects these insights. Let  $Y$  be a function of variable(s)  $X$ , where some variable  $Z$  enters the data-generating process either as an independent factor or to scale  $X$  and/or  $Y$ .

- Do not use correlation  $r$  of variables with common components, since  $r(Y/Z, X/Z)$  is obviously non-zero even if  $r(Y, X)$  is not (Firebaugh, 1988). This error was essentially never committed (with the possible exception of an article in 1997 and one in 2004).
- For similar reasons, investigate measurement error in common components (Long, 1980). The first panel of figure 1, which shows a three-year moving average, suggests that measurement error features more heavily when authors use ratio variables, although the difference to other statistical articles is small.

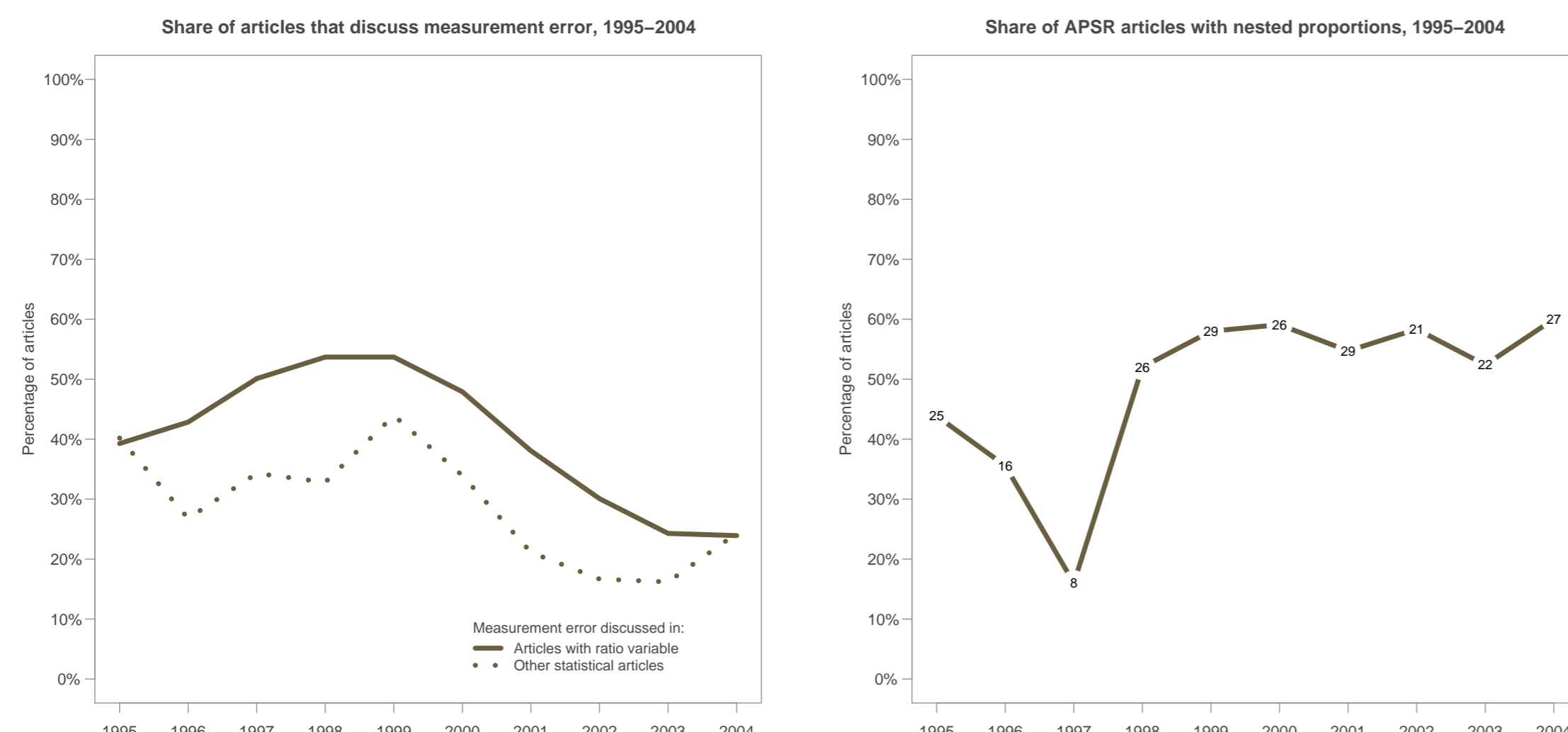


Figure 1

- Avoid proportions in favor of non-nested ratios, which reduces model dependence (Firebaugh and Gibbs, 1985). However, the second panel of figure 1 shows that nested proportions are as popular as other ratio variables.
- Include the inverse scale variable in a model with ratio variables, since omitting it can generate bias if the true data-generating process is in fact  $Y = \alpha + \beta X + \gamma Z + \varepsilon$  (Firebaugh and Gibbs, 1986). Very few articles report doing so, as the first panel in figure 2 shows.

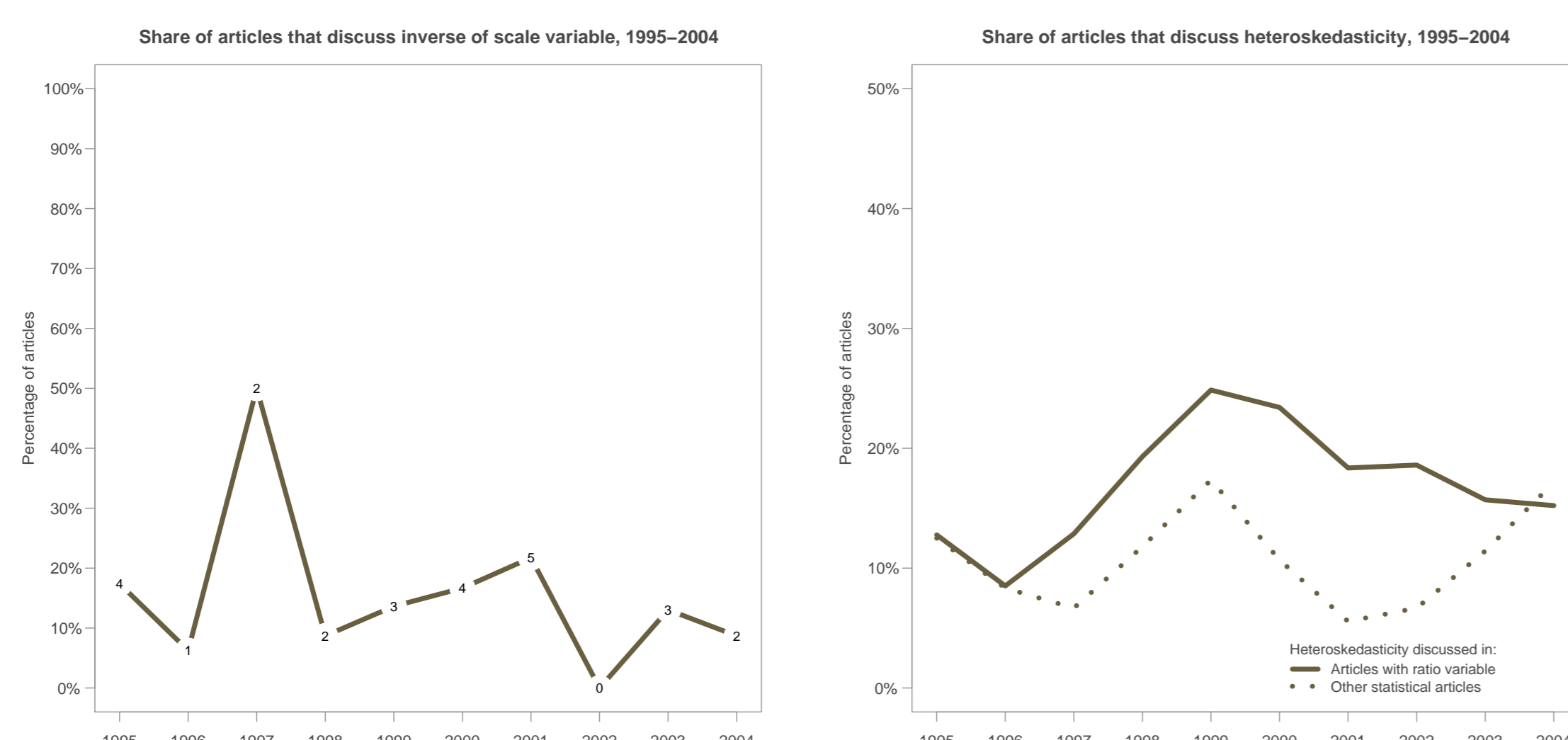


Figure 2

- Check for heteroskedasticity, which could be the result of an improperly specified ratio (that is,  $\varepsilon/Z$ ) (Firebaugh, 1988). The second panel of figure 2 indicates that articles containing ratio variables are (slightly) more likely to test for heteroskedasticity than other papers.

## Modeling in the Face of Specification Uncertainty

The use of ratio variables in regression can help recover degrees of freedom and correspondingly yield efficiency gains, but the fundamental problem is that we are often uncertain about the true specification of the data-generating process, an issue not squarely addressed in this literature. I propose a simple, but effective solution. Suppose we have some variable  $Z$  that could enter the DGP either as an independent factor or as a scale variable (e.g. population size). We will consider four different DGP in particular, for which the linear components are listed in the first column of figure 3. We then estimate the unrestricted model encompassing these DGP, which is the model where  $Z$  is fully interacted:

$$Y = \alpha + \beta X + \delta XZ + \gamma Z + \lambda \frac{X}{Z} + \varepsilon$$

We reject or accept the null of a given DGP on the basis of the estimation results. For example, we accept the null of the first DGP being correct if  $\beta$  and  $\gamma$  are significant;  $\alpha$ ,  $\delta$ , and  $\lambda$  are jointly insignificant; and we find evidence of heteroskedasticity. Simulations shown in figure 3 provide evidence that this approach performs well in recovering information about the data-generating process and in navigating the trade-off between specificity and power (i.e. between minimizing Type I and Type II error).

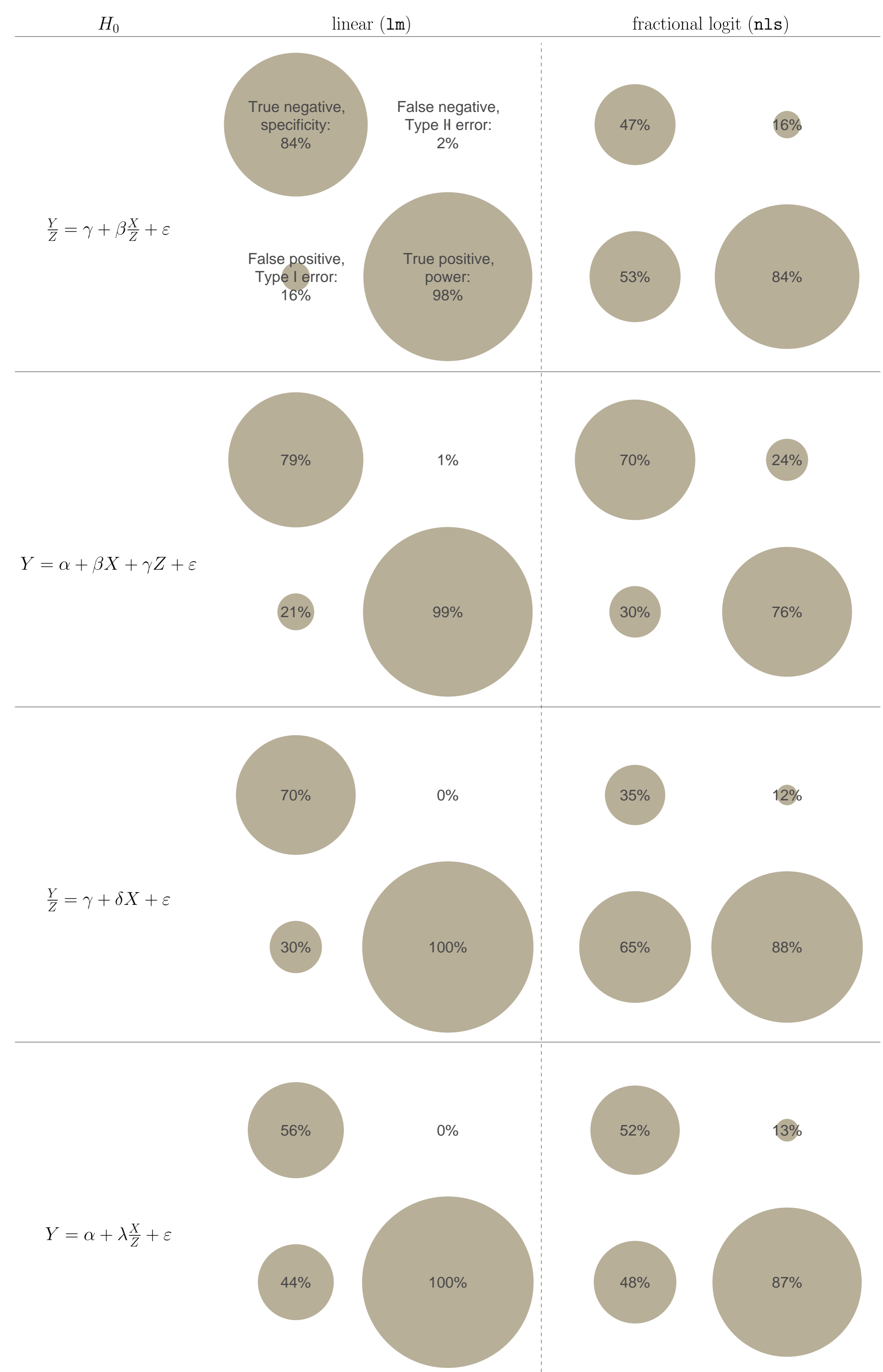


Figure 3

## References

- Glenn Firebaugh. The ratio variables hoax in political science. *American Journal of Political Science*, 32(2):523–535, 1988.
- Glenn Firebaugh and Jack P. Gibbs. User's guide to ratio variables. *American Sociological Review*, 50(5):713–722, 1985.
- Glenn Firebaugh and Jack P. Gibbs. Using ratio variables to control for population size. *Sociological Methods & Research*, 15(1–2):101–117, 1986.
- Susan B. Long. The continuing debate over the use of ratio variables: Facts and fiction. *Sociological Methodology*, 11:37–67, 1980.