Appendix

Modeling with Skewed Distributions

Citation data display at least two unique features that complicate efforts to model them. First, it is well-established in the literature that citation data tend to be highly-skewed (in the positive direction), with many papers receiving zero or very few citations and a small number of papers receiving a large number of citations (Thelwall and Wilson 2014). The skewed distribution of our article citation variable meant that it was inappropriate for us to model it using a conventional linear regression model, such as ordinary least squares.

Poisson and negative binomial models are two regression methods that are commonly used to model count variables (such as the number of times a paper has been cited) that are positively skewed. Poisson and negative binomial regression differ in their assumptions about the distribution of the dependent variable. Poisson regression is appropriate if the dependent variable is equi-dispersed, while negative binomial regression is appropriate if the dependent variable is overdispersed. Popular indicators of overdispersion are the Pearson and $\chi^2$ statistics divided by the degrees of freedom, so-called “dispersion statistics”. If these statistics are greater than 1.0, a model is said to be overdispersed (Hilbe, 2011: pages 88 and 142). By these measures, we found our citation count variable to have overdispersion, so we chose to use a negative binomial model.

Dealing with Multi-level Data

The second unique feature of citation data is that they display a nested (or multi-level) structure, in the sense that papers are “nested” within authors and within journals. By default, regression models are typically set up based on an assumption that the observations in the dataset are independent of each other. This assumption is violated, however, when two or more papers
by the same author are included in a citation data study and/or when two or more papers from the same journal are included. Papers written by the same author (and/or published in the same journal) will share something in common with each other that they do not share with papers from different authors or journals, which can have implications for the number of citations the papers receive. For example, if some authors or journals are more well-known and respected than others, there might be an “author effect” or a “journal effect” that contributes to their papers receiving more citations on average than papers written by other authors or published in other journals, other things being equal.

The current best statistical approach to modeling nested data is multilevel modeling (MLM), also called hierarchical modeling. While MLM has been around for at least a few decades, planning researchers have only just begun to use it (Ewing et al., 2015; Tian et al., 2015). MLM accounts for dependence among observations, in this case the dependence of article citation counts on the characteristics of authors and journals. If this dependence is not accounted for in the model, the standard errors of the regression coefficients will typically be underestimated and the coefficient estimates will be inefficient. MLM overcomes these limitations, accounting for the dependence among observations and producing more accurate coefficient and standard error estimates (Raudenbush and Bryk, 2002). For example, in the now voluminous travel behavior literature, travel characteristics of individuals or households are ordinarily modeled in terms of both individual socioeconomic characteristics and neighborhood built environmental characteristics (see Ewing and Cervero 2001). When multiple cases are drawn from the same neighborhoods, the resulting regression coefficients will be inefficient and standard errors of coefficients will be underestimated. MLM overcomes these limitations,
accounting for the dependence among individuals residing in a given place and producing more accurate coefficient and standard error estimates.

When an outcome varies systematically in two dimensions (such as citation counts that vary with authors and journals), and random effects are present, the resulting data structure is best represented by a cross-classified random effects model (see Raudenbush and Byrk 2002, ch. 12). A cross-classified random effects model is a special class of multi-level model in which lower-level units are nested within two or more higher-level units. The two higher levels in this study are journals and authors. A cross-classified random effects model allows us to study the effects of journal differences while controlling for author effects, and the effects of author differences while controlling for journal effects.

The MLM models estimated in this study can be characterized as pairs of linked statistical models. At the lower level, article citations are modeled as a function of article characteristics plus a random error. Thus, for each combination of author and journal there is a regression equation that describes the association between article characteristics and article citations for that journal and author. At the higher level, the intercept and coefficients are conceived as outcomes, and are modeled in terms of journal and author characteristics plus random effects.

In some MLM models, only the intercepts are allowed to vary across (e.g.) journals and authors, while all regression coefficients are treated as invariant (fixed) across journals and authors. These are often termed "random intercept" models, to denote that only the intercept randomly varies. In other MLM models, the regression coefficients are allowed to randomly vary as well. These are often termed "random coefficient" models.
In this study, all models initially assumed the random intercept form. Only the intercept term in the journal- and author-specific model was allowed to vary; and all journal- and author-specific coefficients were taken as fixed. Then, this assumption was relaxed, and coefficients were allowed to vary as a function of journal and author characteristics, effectively permitting interactions between journal/author and article characteristics.

The variances of coefficients across journals and authors were seldom significant, and interactions between journal/author and article characteristics were seldom significant and never sufficiently large to appreciably affect the relationships between article characteristics and article citations. Hence, we only report results for random intercept models.

References


