Authors
Jeffrey Alan Greene, Lara-Jeane Costa, Alexandra Earl, Kristin Dellinger
University of North Carolina at Chapel Hill

Title
Using Poisson, Negative Binomial, and Zero-Inflated Count Models for Analysis of Self-Regulated Learning Process Data

Summary

Objective
The conceptual and empirical evidence for modeling self-regulated learning (SRL) processes as a dynamic series of events is overwhelming (Winne & Perry, 2000; Zimmerman, 2008). Process data, however, are frequently non-normally distributed and are best analyzed as counts using Poisson, Negative Binomial, and Zero-Inflated models, rather than standard linear regression models (DeMaris, 2004). This presentation will explicate count models and how they can be used to effectively model college students’ SRL while using a computer-based learning environment (CBLE).

Perspective and Theoretical Framework
We utilized an extended version of Winne and Hadwin’s (1998) model of SRL (see Azevedo & Cromley, 2004), to examine how the frequency of SRL processing (e.g., planning, summarization) mediated relations between prior knowledge and academic performance (Zimmerman, 2000). Because this model treats SRL processing as a dynamic process rather than a static disposition, measures must be used that capture data regarding SRL processing as it fluctuates in quantity and quality throughout learning. Think-aloud protocol techniques (Ericsson, 2006) can be used to create counts of how often participants enact specific SRL processes. These count data, while rich in information and conceptually aligned with process models (Winne & Perry, 2000), are problematic to analyze because they require specialized statistical models.

As outlined by DeMaris (2004), distributions of count data are usually non-normal, often because a large proportion of participants have zeroes as scores. In practice, variables with this type of distribution are often transformed into categorical data, perhaps into two groups of participants, one with zero on the variable and another with counts above zero. Categorizing these data results in a loss of power, and ignores a large amount of information about the participants’ scores (MacCallum et al., 2002). In addition, analyzing count data via standard linear regression modeling is not recommended, as often these data violate a number of assumptions of this technique. Both the Poisson and the Negative Binomial regression model are good ways of analyzing count data because their underlying probability distributions more closely approximate the typically right-skewed distributions of count data.

The choice between using a Poisson or Negative Binomial regression model often depends upon the distribution of the data in question. In the Poisson model, it is assumed that the mean and variance of a variable are equal. This assumption, however, is rarely
upheld with count data (DeMaris, 2004). The Negative Binomial regression model is often a more appropriate technique than the Poisson model for analyzing count data because it allows for the variance to exceed the mean, a condition called overdispersion.

Overdispersion can also be due to an unusually large number of zeroes on a count variable (DeMaris, 2004). For example, a SRL strategy, such as drawing, may be used by only a small subset of the sample, resulting in most participants having a zero score for this variable. In such circumstances it is useful to use Zero-Inflated versions of either the Poisson or Negative Binomial regression models because these models distinguish between two types of zeroes in the data. The first type of zero includes participants who are not likely to demonstrate the particular SRL process at any point during their time learning (i.e., “true zero”). For example, participants lacking SRL skills are unlikely to engage in sophisticated strategy use such as coordinating information sources (COIS; see Azevedo & Cromley, 2004). The second type of zero is a result of participants with strong SRL skills simply not using the strategy during the time they are observed (i.e., “occasional zero group”). For example, participants with strong SRL skills are more likely to use COIS, but may not do so during the time they are observed. This participant, if followed during another learning task, very well might use COIS, whereas the previous participant would be unlikely to do so in any learning task. In Zero-Inflated regression, two models are estimated. The first model is a binomial logistic regression of the probability that an individual is in the “true zero” group. The second model is a regular Poisson or Negative Binomial regression model for all participants who do not fall into the first group, and are in the “occasional zero group.” Estimation and interpretation of Zero-Inflated models are the same as described for Poisson and Negative Binomial regression models except that, when looking at individual participants’ predicted scores, one must take into account their likelihood of being in the “true zero” group.

In this presentation we illustrate how these four types of models for count data can be used to analyze SRL processing using data collected in a previous study.

Methods and Data Sources
The abbreviated nature of this paper precludes a detailed description of the methods and data sources used in this study. Briefly, we gathered prior knowledge, think-aloud protocol, and academic performance data from 170 college-students as they used a CBLE to learn about the circulatory system. These data were coded and scored using techniques refined by Azevedo and colleagues (Azevedo et al., 2005; Greene et al., 2009) to generate information regarding participants’ conceptual understanding of the circulatory system and the frequency of their use of 25 SRL processes.

Results
We showed how four different statistical models for count data (i.e., Poisson, Negative Binomial, Zero-Inflated Poisson and Zero-Inflated Negative Binomial) could be incorporated into regression and latent profile analyses. Specifically, our empirical analyses indicated that several SRL processes mediated the relationship between prior knowledge and learning a complex science task with a computer. Participants also
clustered into two groups based on their SRL processing and academic performance, which aligned with Brophy’s (2004) model of schematic and aschematic learners.

**Scholarly Significance of Study**
This presentation contributes to both the methodology of studying SRL and the empirical evidence suggesting that SRL mediates the relation between prior knowledge and academic performance. The techniques outlined in this presentation, as compared to multiple linear regression, will enable researchers to more accurately model SRL processing, increasing the likelihood that future studies will continue to discover valid relations among student characteristics, like prior knowledge, and learning.

**Reference**


