

Mythical Statistical Models and scale-dependent drivers*

Dawn C. Parker, Dept. of Environmental Science and Policy

September 7, 2006

What if we wanted to create a statistical model to understand the “driving forces” of composition (social vs. natural science) within an educational institution?

Our first question might be: *What determines the probability that a student has a social vs. natural science emphasis at the department level?* Our model will have one *endogenous* element: whether the student is social vs. natural science. It will have several *exogenous* elements. Some are fixed among all students in the department, so we don’t include them in our model (the department and college that the student is in). We start by assuming a single *independent or explanatory variable/driving force* for science type: the number of class of each type the student has taken as an undergrad. Our model is:

$$p(S|N) = f(\# \text{ classes each type in undergrad})$$

Our second question might be: *What determines the probability that a student has a social vs. natural science emphasis at the college level?* Our *endogenous* element is still the same: whether the student is social vs. natural science. We also have the same number of exogenous elements, but now one more of them varies at the scale at which we are analyzing the problem (which department the student is from). (Note that the scale at which the model operates—the student—remains the same.) So, we add a dummy variable for which college the student is from as an additional independent variable/driving force. (What kind of effect will this dummy have, given the information we have?) The new model is:

$$p(S|N) = f(\# \text{ classes each type in undergrad; department dummy})$$

This model may indicate that the department is the most important driving force—that it appears statistically more important than the number of classes.

Finally, we may want to ask a different question: *What determines the probability that a student will be in Environmental and Natural Resources vs. GeoStuff?* We have some priors about potential explanatory variables/driving forces, and create a model that includes:

$$p(ENR|GS) = f(\begin{array}{l} \# \text{ classes each type in undergrad;} \\ \# \text{ faculty each type in department,} \\ \text{required classes of each type,} \\ \text{whether student has a natural or social science orientation)} \end{array})$$

*Copyright 2006 Dawn Parker.

Wise students may notice a potential problem here—the dependent variable (endogenous element) for the previous model is also an independent variable/exogenous element for this model. This means that the two models form a system of equations, or that in the real world, we believe that the two elements are determined together.