What do you do when faced with analyzing student ratings from 1 to 5 for 3 instructors in 3 classes? Aside from questioning the validity of students assessing instructor capability other than for the income generated by high enrollments, and that this is but a toy example, many naive analysts turn to the well-known analysis of variance (ANOVA) using score as a response and instructor and class (eg., time series, linear models, etc.) as influencing variables to make comparisons among the instructors and classes. While each instructor may teach more than one class, no two instructors share the same class.

As a straightforward example, let us assume 240 students submit ratings for their instructors and and classes. Further, let us assume the ratings are based upon a 1, 2, 3, 4, 5 point scale. While ordinarily the class sizes would be of varying numbers, we will assume the balanced situation of 10 students for each class and instructor combination. A descriptive form of the ANOVA model is

$$\text{score} = \text{mean} + \text{instructor} + \text{class} + \text{instructor} \times \text{class} + \text{error}. \tag{1}$$

The mean is the average of all the students' ratings. The terms instructor, class, and the interaction of instructor with class, assuming each instructor has taught each class, partition the variance in the scores to reveal how much of the overall variance is accounted for by these terms, and how much remains as error (residuals).

This model appears perfectly reasonable, so why do seasoned analysts turn to methods such as those we present in *Handbook for Advanced Modeling: Non-Gaussian and Correlated Data*?

Consider the nature of the response, score. As mentioned above, it ranges from 1 to 5 in steps of what appears to be unity. There are three issues with model error we must consider when using a 5-point scale as a response: 1) a Gaussian probability distribution, (2) homogeneity of variance and (3) the independence, viz., no error value-to-value correlation. As we shall see below, there are models to accommodate any or all violations of these three considerations.

The first point requires that the errors follow a Gaussian distribution. This requirement allows the use of least squares estimation of the model parameters, an efficient method to obtain the maximum likelihood function for asymptotically unbiased estimates of the parameters. A 5-point rating response must assume the difference between each level has the same weight, or that the weight is known so compensation may be obtained. However, there are no guarantees that any two students have the same sense of distance between the levels, and hence there is unlikely to be continuity throughout the scale. It is possible to discover how each rater considers the between-level differences, but finding rater distance weights is intractable in nearly all situations. Hence, we may assume the ANOVA residuals are most likely to be non-Gaussian.

Secondly, that students aren't likely to think of the level differences consistently suggests that ANOVA residuals will most probably possess non-constant variance. If this is the case, many try to find weights form homogeneous variance of the errors, but we consider this to be quite artificial and not generally reproducible from one data set to another.

Thirdly, as more than one student sat the same class or enjoyed an instructor for more than one class, the ratings must be assumed to not be independent, and thereby carrying this lack of independence to the residuals such that they have value-to-value correlations. Other than using models designed to account for this correlation, there are few reproducible remedial methods.

We now consider a model that accounts for the issues of using a 5-scale as a response. The model is as easily constructed as the familiar ANOVA in the commonly used statistical packages. The model we consider is the ordered multinomial regression model. Actually, there are three versions available to the analyst depending on what information is desired. We refer you to Chapter 5 of our book for full explanations, but in this treatment we will assume the analyst interested in assessing the odds of obtaining "higher" values over "lower" values. For example, we assess the odds of obtaining a 4 or a 5 over values of 1, 2, or 3, and all other combinations. Hence, we now describe cumulative logistic model.

The cumulative logistic model, which we will call the cum logit model for convenience, forms the odds ratio of obtaining a 4 or a 5 versus having a a lower score of 1, 2, or 3. The ratio for this specific cut level at 3 in a cum logit model is constructed as:

$$\frac{\pi_{i1} + \pi_{i2} + \pi_{i3}}{\pi_{i4} + \pi_{i5}} = \exp\left(\beta_{03} + \beta_{1,2}x_{1,2,i} + \beta_{1,3}x_{1,3,i} + \beta_{2,2}x_{2,2,i} + \beta_{2,3}x_{2,3,i}\right). \tag{2}$$

Note we have removed the interaction term for simplicity, but if present, a slope and interaction term is generated for each of the two levels of instructor and class. In Eq. 2, $\pi_{ij},\ j = 1, 2, 3, 4, 5$ are probabilities of student i giving a rating j, the numerator of the response ratio is the categories of 1 to 3 for student i, the denominator is the categories of 4 and 5 for student i, the parameter β_{03} the intercept associated with the rating cut level at 3, $\beta_{k,l},\ k = 1, 2$, and $l = 1, 2$, are the slope parameters to be estimated for the predictors $x_{1,1,i}$ is the lth level of instructor, and $x_{2,l,i}$ is the lth level of class, each for student i. Note that the first level of instructor and class is the reference level, and hence, does not appear explicitly in the model. Most statistics packages manage the levels in the model. For 5 ratings levels there are 4 equations that can describe the 4 odds ratios that can be formed depending upon at which level of the cut point, in this case cut levels 2, 4, and 5.

For our example, $\exp\beta_k$ gives the expected multiplicative change in the odds of the 1, 2, 3, rating over the 4, 5 rating for a unit step change in the respective k predictors. For example, instructor 2 may have twice the odds of obtaining a rating of 4 or 5 than that of instructor 1. Similarly, the odds that class 2 obtains a score of 4 or 5 may be 25% less than that of class 3.

So, how is using an ordinal multinomial model superior to ANOVA? Using the probabilities and odds modeling methods account for inherent heterogeneous variation in how students rate instructors and classes without arbitrary assumptions for relating the category levels. Categorical response models allow for nonlinear relationships between the probabilistic responses and the associated predictor variables. Further, the assumptions for estimating parameters in ANOVA are not necessary for categorical response models and hence artificial adjustments such as using weights that attempt to satisfy the ANOVA assumptions are unnecessary. In terms of predictions, models of odds and probabilities will always produce sensible predicted probabilities associated with each outcome category, which does not hold for linear models. Finally, the implementations in the various statistical packages in common use make using categorical models just as straightforward as using ANOVA.

See additional blog responses.