Paper 248–2009

Learning When to Be Discrete: Continuous vs. Categorical Predictors

David J. Pasta, ICON Clinical Research, San Francisco, CA

ABSTRACT

Some predictors, such as age or height, are measured as continuous variables but could be put into categories ("discretized"). Other predictors, such as occupation or a Likert scale rating, are measured as (ordinal) categories but could be treated as continuous variables. This paper explores choosing between treating predictors as continuous or categorical (including them in the CLASS statement). Specific topics covered include deciding how many categories to use for a discretized variable (is 3 enough? Is 6 too many?); testing for deviations from linearity by having the same variable in the model both as a continuous and as a CLASS variable; and exploring the efficiency loss when treating unequally spaced categories as though they were equally spaced.

INTRODUCTION

Early in your statistical training, whether it was formal or informal, you probably learned that variables have a "level of measurement" of nominal, ordinal, interval, or ratio. The popularization of this rubric goes back at least to the 1950s (see Blalock 1979 section 2.2 and the references mentioned there). A nominal variable is a classification for which there is no ordering (although sometimes there is a partial ordering): the values are just "names" and are not to be interpreted quantitatively even if they are numbers. The values of an ordinal variable can be put into a unique order, but the distance between values cannot be quantified. For an interval variable, the distance between values can be quantified but the "zero" is arbitrary, so we cannot talk about one value as being "twice as big" as another. Finally, the highest achievement for a variable is to be a ratio variable: both the distances between values and their ratios can be quantified.

It may surprise you to learn that this method of characterizing variables is not, in fact, generally accepted by statisticians. Yes, it has some value as a pedagogical tool and it provides some common language for discussing what sorts of analyses might make sense. However, it ignores important distinctions within categories, including whether a nominal variable has a partial ordering and whether a ratio variable arises as a count or a proportion. Much can be (and has been) written on this topic; a good starting place is Velleman and Wilkinson (1993). For the purposes of this paper we will emphasize a very practical distinction that arises in the analysis: will the variable be treated as continuous or as categorical?

We will refer to variables as continuous even though it is easy to argue that no variable being analyzed in a digital computer is truly continuous, as measurements are recorded with finite precision. What we really mean is that we're treating the variable as a measure of an underlying continuous or approximately continuous value and we are willing to treat the differences between values as quantitative. Thus it is meaningful to talk about the effect of "a one-point increase" in the value of X or for that matter "a 0.3-point increase". This is one place where it may be important to distinguish among subdivisions of continuous variables. If variable X is a count, we would probably want to talk only about whole-number increases in the value of X; if it is a proportion, we would only want to talk about increases that were less than 1. What we are calling continuous variables are referred to by others as quantitative, metric, interval-scaled, or other similar terms. The important thing to remember is that for continuous variables we are treating each unit change as having the same effect.

When we do not want to treat the differences between values as quantifiable, or at least not uniformly quantifiable, we treat the variable as categorical. In SAS® procedures, this means including the variable on the CLASS statement. The values represent categories. It will be important to know whether those categories are unordered (nominal), partially ordered, or fully ordered (ordinal). It is even possible for the fully ordered variables to be interval or ratio – for example, if it represents numerical ranges of income – but what is important for our purposes is that we want to estimate the effect of each value separately. Thus the effect of moving from one category to another may differ depending on the categories. These variables are also referred to as discrete, but we use the term categorical because it is in broad use and because even variables treated as continuous are measured discretely.

A WORD ABOUT BINARY VARIABLES

Binary variables are those that take on exactly two values, such as 0 and 1 or True and False or Male and Female. For analysis purposes, they can be considered either continuous or categorical. In general it doesn't matter which way you think about them. However, it can have implications for computational algorithms, for parameterizations of models, and for interpretations of results. There are circumstances where it matters a great deal whether you are treating a binary variable as continuous or categorical, such as when you are adjusting for it in a linear model and you are calculating least squares means (LSMEANS). Specifically, putting a binary variable in a CLASS statement affects (1) the parameterization and therefore (2) the interpretation of the results; it also affects (3) the calculation of the least squares means (LSMEANS) and also (4) the interpretation of the OBSMARGIN option on LSMEANS. ***Generally***, it is safer to treat binary variables as categorical than to treat them as continuous, although there are times when you will want to treat them as continuous.

SHOULD MY VARIABLE BE CONTINUOUS OR CATEGORICAL?

At first blush, it seems easy to tell which variables should be continuous and which should be categorical. There are, however, many gray areas and even situations where you are quite sure it may turn out that others have a different point of view. My experience is that the decision at times appears to hinge on the analytic techniques people are most familiar with. Someone who works with lots of survey data and is very comfortable with categorical variables is eager to treat household income (measured to the nearest thousand) as a categorical variable by dividing it into groups. Another analyst, working almost exclusively with continuous variables, might be eager to take household income (as recorded in broad ranges) and make it a continuous variable. How much difference does it make? Are there clear situations that go one way or the other?

First, the easy direction: Any continuous variable can be made into a categorical one – or a set of categorical ones – by "discretizing" it. You define categories and use the continuous value to determine the appropriate category for each measurement. Why would you want to do that? Don't you lose information that way? How can that ever be a good idea?

It is true that if the variable in question has an exactly linear relationship with the outcome, you do lose information by making a continuous variable into a categorical one. Furthermore, instead of estimating a single coefficient (1 degree of freedom, or df) you need to estimate K coefficients if your variable has K categories, which represents K-1 df. (You use up only K-1 degrees of freedom because of the inherent redundancy of classification – if you know an observation is not in any of the first K-1 categories, it must be in the Kth category. Put another way, the proportion of observations in the categories must add up to 1. Therefore as long as there is an intercept term in the model, or another categorical variable, the number of degrees of freedom is equal to the number of categories minus 1.) On the other hand, what if the relationship is not precisely linear? Treating the variable as continuous allows you to estimate the linear component of the relationship, but the categorical version allows you to capture much more complicated relationships.

What about the other direction? Does it ever make sense to take a categorical variable and treat it as continuous? Indeed it does. In fact, I would argue that it is nearly always worthwhile at least examining the linear component associated with any ordinal variable. Even if you want to keep a variable as categorical, it is worth understanding the extent to which the relationship is linear. It is, in general, a more powerful approach to analyzing ordinal variable to treat them as continuous and to fail to consider that possibility may cause many useful relationships to be overlooked. The article by Moses et al. (1984) is positively eloquent on the subject.

One concern often expressed is that "we don't know that the ordinal categories are equally spaced." That is true enough – we don't. But we also don't "know" that the relationship between continuous variables is linear, which means we don't "know" that a one-unit change in a continuous variable has the same effect no matter whether it is a change between two relatively low values or a change between two relatively high values. In fact, when it's phrased that way -- rather than "is the relationship linear?" -- I find a lot more uncertainty in my colleagues. It turns out that it doesn't matter that much in practice – the results are remarkably insensitive to the spacing of an ordinal variable except in the most extreme cases. It does, however, matter more when you consider the products of ordinal variables.

I am squarely in the camp that says "everything is linear to a first approximation" and therefore I am very cheerful about treating ordinal variables as continuous. Deviations from linearity can be important and should be considered once you have the basics of the model established, but it is very rare for an ordinal variable to be an important predictor and have it not be important when considered as a continuous variable. That would mean that the linear component of the relationship is negligible but the non-linear component is substantial. It is easy to create artificial examples of this situation, but they are very, very rare in practice.

Are there situations where even I would insist on keeping a variable as categorical? As tempting as it might be for some people to put an order on race/ethnicity or religious affiliation, except in rare cases that is inadvisable. There are certainly situations where objects have been grouped by unspecified criteria and part of the object of the analysis is to understand those groupings – those need to be considered nominal, not ordinal. Genetic mutations might also be nominal, although often there is a partial ordering associated. You can probably think of some other examples from your own experience. In general, though, truly nominal (not even partially ordered) variables are infrequent in practice.

Just as uncommon, in my view, are continuous measures where you are certain that the effect is linear (a one point change has the same impact no matter on the scale it occurs). In fact, other than the limiting case of binary variables (where there is not enough information to detect nonlinearities) no good examples come to mind outside of the physical sciences. So I see the world as pretty much shades of gray. There are many variables might be treated either as continuous (linear) or as categorical and many fewer that should definitely be treated one way or another.

AN EXAMPLE: TESTING FOR DEVIATIONS FROM LINEARITY

I mentioned testing for deviations from linearity. How do you do that? It's actually pretty easy, but it leads to output that people find a little odd-looking at first. For any ordinal variable, (1) but the ordinal variables in the CLASS statement, (2) make an exact copy that will not be in the CLASS statement, and (3) include both variables in the MODEL statement. For example, you might have a variable measuring education called EDUCAT with K categories. You can create L_EDUCAT (L for Linear), and include both in the model. What happens? L_EDUCAT will have 0 degrees of freedom and 0 Type III effect (it doesn't add any information after the categorical EDUCAT is included). EDUCAT will be a test of deviations from linearity with K-2 degrees of freedom – 1 lost to the overall constant, and 1 lost to the linear effect L_EDUCAT. There are some details to watch out for, best expressed by looking at some SAS output.

```
EDUCAT categorical with typical labels
Dependent Variable: y
                                              Sum of
Source                      DF               Squares     Mean Square    F Value    Pr > F
Model                        4            21702.2880       5425.5720       2.24    0.0707
Error                       95           230398.3776       2425.2461
Corrected Total             99           252100.6656

R-Square     Coeff Var      Root MSE        y Mean
0.086086      33.46631      49.24679      147.1533

Source                      DF             Type I SS     Mean Square    F Value    Pr > F
educat                       4           21702.28797      5425.57199       2.24    0.0707

Source                      DF           Type III SS     Mean Square    F Value    Pr > F
educat                       4           21702.28797      5425.57199       2.24    0.0707
```

```
                                                 Standard
Parameter                      Estimate              Error    t Value    Pr > |t|
Intercept                    136.6563385 B      8.32422640      16.42      <.0001
educat    HS grad             -2.3539316 B     14.86171676      -0.16      0.8745
educat    college grad        35.1031661 B     13.59340479       2.58      0.0113
educat    less than HS         2.6127789 B     15.99531606       0.16      0.8706
educat    post college        21.0818184 B     15.19788858       1.39      0.1686
educat    some college         0.0000000 B       .                .         .
```

Is that pretty? Well, not really. The reference category is "some college" and the order is, shall we say, not exactly natural. One quick solution to that is to use numbered labels. Most of the output is the same until you get to the parameter estimates.

EDUCAT categorical with numbered labels

```
Dependent Variable: y
                                        Sum of
Source                      DF         Squares     Mean Square    F Value     Pr > F
Model                        4      21702.2880       5425.5720       2.24     0.0707
Error                       95     230398.3776       2425.2461
Corrected Total             99     252100.6656

R-Square     Coeff Var      Root MSE         y Mean
0.086086      33.46631      49.24679       147.1533

Source                      DF       Type I SS     Mean Square    F Value     Pr > F
educat                       4     21702.28797      5425.57199       2.24     0.0707

Source                      DF     Type III SS     Mean Square    F Value     Pr > F
educat                       4     21702.28797      5425.57199       2.24     0.0707
```

```
                                                   Standard
Parameter                        Estimate              Error    t Value    Pr > |t|
Intercept                      157.7381569 B     12.71546586      12.41      <.0001
educat    1 less than HS       -18.4690395 B     18.66120207      -0.99      0.3248
educat    2 HS grad            -23.4357501 B     17.69917942      -1.32      0.1886
educat    3 some college       -21.0818184 B     15.19788858      -1.39      0.1686
educat    4 college grad        14.0213477 B     16.64845280       0.84      0.4018
educat    5 post college         0.0000000 B       .                .         .
```

Well, that's certainly easier to follow. Now the reference category is the highest education (post college) and the categories are ordered. We've got a p-value of 0.071, which is borderline. What happens if we treat education as a continuous variable? All we need to do is omit it from the CLASS statement.

EDUCAT continuous

```
Dependent Variable: y
                                       Sum of
Source                      DF         Squares     Mean Square    F Value     Pr > F
Model                        1      10457.6803      10457.6803       4.24     0.0421
Error                       98     241642.9853       2465.7447
Corrected Total             99     252100.6656
```

```
R-Square     Coeff Var      Root MSE        y Mean
0.041482      33.74458      49.65627      147.1533

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
l_educat                     1     10457.68028     10457.68028       4.24    0.0421

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
l_educat                     1     10457.68028     10457.68028       4.24    0.0421

                                         Standard
Parameter          Estimate                 Error    t Value    Pr > |t|
Intercept       121.1955784           13.54728807       8.95      <.0001
l_educat          8.4005599            4.07910241       2.06      0.0421
```

That gave us a p-value of 0.042, so we have a statistically significant linear trend. But are the deviations from linearity statistically significant? This is the moment we've been waiting for.

EDUCAT continuous and categorical

```
Dependent Variable: y
                                            Sum of
Source                      DF             Squares     Mean Square    F Value    Pr > F
Model                        4          21702.2880       5425.5720       2.24    0.0707
Error                       95         230398.3776       2425.2461
Corrected Total             99         252100.6656

R-Square     Coeff Var      Root MSE        y Mean
0.086086      33.46631      49.24679      147.1533

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
l_educat                     1     10457.68028     10457.68028       4.31    0.0405
educat                       3     11244.60769      3748.20256       1.55    0.2078

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
l_educat                     0         0.00000               .          .         .
educat                       3     11244.60769      3748.20256       1.55    0.2078

                                                             Standard
Parameter                         Estimate                      Error    t Value    Pr > |t|

Intercept                      227.8448953 B              73.98734261       3.08      0.0027
l_educat                       -14.0213477 B              16.64845280      -0.84      0.4018
educat     1 less than HS      -74.5544302 B              59.07208796      -1.26      0.2100
educat     2 HS grad           -65.4997931 B              42.86841897      -1.53      0.1299
educat     3 some college      -49.1245137 B              26.32351518      -1.87      0.0651
educat     4 college grad        0.0000000 B                .                .          .
educat     5 post college        0.0000000 B                .                .          .
```

The overall p-value is the same as it was originally (0.071), and as promised the L_EDUCAT variable has 0 degrees of freedom in the Type III section. The EDUCAT variable has 3 degrees of freedom and a p-value of 0.21, indicating a lack of statistical significance. That is, the deviations from linearity are non-significant. I will leave it as an exercise for the reader to figure out how to manipulate the parameter estimates from this run to get the values from the first two. It's a good way to make sure you understand what SAS is doing.

"DISCRETIZING" A CONTINUOUS VARIABLE INTO A CATEGORICAL VARIABLE

Suppose you have a variable measured as continuous but you are concerned about the possibility of nonlinear effects. The nonlinearity might be simple curvature (say, a quadratic or exponential curve upwards or downwards). It might be a floor or threshold effect or a ceiling effect. Or it might be something more complicated, where the relationship is not monotonic. How should you go about creating the categorical version?

You should start by making use of any available substantive knowledge. If the variable is nonnegative and you think zero values might be qualitatively different from the positive values, be sure to make the zero category separate. Similarly, if a variable takes on both negative and positive values it is likely appropriate to have categories consisting entirely of negative or entirely of positive values. (This does not preclude the possibility of including some small negative and/or small positive values along with zero; in many situations it makes sense to include values close to zero in a "near zero" category.) You might also know that a certain value is often used as a boundary for one reason or another – with ages, there are good reasons to use 65 in U.S. studies, for example.

You should also take account of the empirical distribution of the variable when creating categories. As interested as you might be in a category, if there is very little data there you won't be able to say much about it. By the same token, having a category with few values might be used to isolate a small number of aberrant cases. In the absence of other overriding considerations, it's often sensible to define your categories so they all have about the same number of observations.

How many categories should you create? That's easy: as many as you need but no more. Unless you have prodigious amounts of data and a very complicated relationship, it's hard to imagine needing even as many as 10 categories. In fact, in practice 3 or 4 categories often carry most of the information and there's rarely a reason to go beyond 5. Some theoretical and simulation work has shown that in a variety of situations there is little to gain by going beyond 5 categories. That same work has shown that for symmetric variables the optimal spacing is also symmetrical around the median but not evenly spaced. However, the efficiency loss from using equal spacing is minimal. Some of the early work on this is in Cox (1957), which includes the optimum groups for the Normal case. The paper by Cochran (1968) provides a broader view of the subject, including some non-Normal cases. Section 3.4 of Cochran (1983) provides a brief and fairly accessible summary of both papers.

NUMBER OF CATEGORIES	OPTIMUM SIZES	VAR (OPTIMAL)	VAR (EQUAL)
2	50, 50	1.57	1.57
3	27, 46, 27	1.23	1.26
4	16, 34, 34, 16	1.13	1.16
5	11, 24, 30, 24, 11	1.09	1.11

I tend to use five categories, with approximately 10, 25, 30, 25, 10 percent in each category, which is fairly easy to remember. (If you remember the 10 and 25 at the edges, you can get the 30 by subtraction.) With four categories, think of it as approximately 1/6, 1/3, 1/3, 1/6. Remember that as have "half as much in the end categories" and you can reproduce that one, too. (A similar rule would apply to three categories, producing 25, 50, 25.)

How much difference does it make how many categories you use and whether they are approximately equal categories or you go to the trouble to use near-optimum sizes? For an INCOME variable and various discretized versions, the correlations range from .83 to .95 if you use at least three categories – a difference, but less than many people expect.

Correlations with original INCOME variable	
11 categories	0.95
7 categories	0.91
5 categories equal	0.88
5 categories unequal	0.90
4 categories equal	0.86
3 categories unequal	0.86
3 categories equal	0.83
2 categories equal	0.72

MORE ON BINARY VARIABLES

There are also situations where there are practical advantages to creating one or more binary variables – dare we say binarizing? – from a continuous variable. You may want to know if having any positive value is the important thing rather than the specific amount. Or you might believe that there is a threshold effect but you are not sure at exactly which value it occurs. You might create several binary variables with different thresholds so that you can compare their predictive power. In another circumstance, you might create a cumulative set of vinary variables that could be combined to represent an ordinal variable with unequal spacing. You could then test the equality of coefficients to determine how close to equally-spaced the variables were. An example of this sort of coding would be to create from a continuous variable X one variable called Z1 that is 1 when X is 1 or more and 0 otherwise, another variable called Z2 that is 1 when X is 2 or more and 0 otherwise, and another variable called Z3 that is 1 when X is 3 or more and 0 otherwise. The sum of Z1, Z2, and Z3 would be a step function with steps at 1, 2, and 3. By including all three variables with possibly different coefficients, you'd be allowing the steps to be of different sizes. If you included just a single variable Z that took on values 0, 1, 2, 3 and treated it as continuous, you'd be assuming each of the steps had the same magnitude. You could treat Z as categorical and you'd have the same model as if you included the Z1, Z2, and Z3 variables. In both cases you would have used 3 df (4-1=3 for the Z variable and 1+1+1=3 for the 3 Z1, Z2, Z3 variables). One advantage of the binary coding is that it is somewhat easier to explicitly test the equality of the step sizes. An obvious disadvantage is the need to create multiple variables. For more on coding of binary variables and piecewise constant (and piecewise linear) variables, see Pasta (2005)

TREATING AN ORDINAL VARIABLE AS CONTINUOUS

When you want to treat an ordinal variable as categorical, you can just include it in a CLASS statement or create a series of binary ("dummy") variables for each of the categories. If you want to treat it as continuous, first be sure that the coding puts them in the correct monotonic order. This is true even if you ware going to include them in the CLASS statement but plan to calculate statistics that use the ordered values. You need to be careful about whether the values are ordered by unformatted (internal) value or formatted value. Do you recognize the sequence EFFNOSSTTT? How about OTTFFSSENT? Does it help to mention the sequence 8,5,4,9,1,7,5,10,3,2? (Write down the English for the numbers from one to ten and sort them alphabetically.) It's also common for "don't know" or "unsure" to be away from its natural ordinal "home" in the middle of a Likert scale. For more on how categories are ordered in the CLASS statement and how the "reference" category is determined, see Pritchard and Pasta (2004).

Once you're sure you have the ordinal variable coded correctly, you can just include it in the model (not in the CLASS statement) to treat it as continuous. But how do you know if you've captured "most" of the explanatory power this way? The relatively simple solution is to create an identical copy of the ordinal variable with a different name and work with both variables. You might consider adopting a naming convention for such variables. Over the years, I have tended to add a "C" or "C_" to the front of the copied variable … but then I forget whether that C stands for Categorical or Continuous. Maybe better would be to use L_ to represent Linear (I like to reserve D_ for date variables).

Consider a series of Likert-scale questions with seven response categories: disagree strongly, disagree moderately, disagree slightly, neither agree nor disagree, agree slightly, agree moderately, agree strongly. If these are coded from 1 to 7 (or 7 to 1), they can be used as they stand. (You might want to consider subtracting 4 from the values so they range from -3 to +3 to make it easier to interpret the results, but that is not material here.) If the original variables were named Q21-Q28 (being questions 21 through 28 of a questionnaire), you might create exact copies named L_Q21-L_Q28. Then for each variable you could include both the original (categorical) version and the linear version in the same model! Won't that be redundant? Yes, it will, and the linear version will show up with zero degrees of freedom and not statistical test of significance. The Type III statistical test of the categorical version will have one less degree of freedom than usual and it will be testing the deviation from linearity – whether the remaining K-2 df have statistically significant explanatory power. If the categorical version is statistically significant, that tells you there is a significant non-linear component and it makes sense to omit the linear version of the variable. If the categorical version is not statistically significant by whatever criterion you shoose to use, that means that the linear component carries the explanatory power and the categorical variable can be dropped. (Of course, there's no guarantee that the linear version will be significant after dropping the categorical variable – that still needs to be tested.)

UNEQUAL SPACING OF ORDINAL VALUES

Although variables are generally very insensitive to variations in the spacing between values, there are times when you want to use a special spacing or at least test whether it is substantially different from equal spacing. This naturally occurs when the categories are ranges of a continuous measure, as is often done on a questionnaire or interview (for example, age ranges or ranges of household income). It is often common that counts of frequent events might be clearly unequally spaced – the categories might be 0, 1, 2-3, 4-6, 7-10. Such a variable might be coded according to the midpoint of the categories; in this example, you would use 0, 1, 2.5, 5, 8.5. But what if you also had a category of "11 or more"? You might allow that value to have its own categorical variable (see the next section) or you might try to assign a value. The assigned value might be based on data from another population with more detailed data collected or it might be based on a theoretical distribution, or it might be the next value that "feel right." In this example, I could easily argue for using 13. Where does lucky 13 come from? The first differences of the assigned values are 1, 1.5, 2.5, 3.5 … and 4.5 feels "right" for the next gap and 8.5+4.5=13.

Another approach that I have found useful (and not entirely arbitrary) is to use the harmonic means of the endpoints to define the mean. The harmonic mean is the inverse of the average of the inverses. For the last (open-ended) category that goes from X to infinity, this means going "halfway to infinity" by averaging 1/X and 0 and taking the inverse. This means, simply, using two times the value of X as the "midpoint" of the upper interval. In the count example, that would be 2*11=22 and in the case of an income category in thousands that ended with "over 200" using a value of 400.

When a variable has an equally-spaced version and a carefully-spaced version that are about equally good as predictors, how do you decide which to use? This may depend on the ease of discussion and interpretation of the results. Consider income categories. It may be easier to talk about the effect of "each 1,000 dollars increase in income" or "each 10,000 dollars increase in income" than to talk about "each one category increase in income." A similar issue arises with age categories. In other cases, the categories will be at least as easy to talk about.

When an ordinal variable is created from an originally continuous variable and you conclude that the nonlinear component is negligible, you're faced with a choice: use the original variable or the one in categories? The answer is the usual one: it depends on which is easier to talk about.

ANOTHER EXAMPLE: ALTERNATIVE SPACING OF EDUCATION

Consider the education variable we used earlier. What happens if we look at alternative spacings when we are using it as a predictor? Simulated data were generated for a relationship with education where the five categories took on the values 1, 2, 8, 20, and 21. That is about as uneven a distribution as you might expect. Two cases were considered: one where the relationship was strong and one where the relationship was much weaker. The results can be summarized in a correlation matrix.

```
             yeduc1      yeduc2      educat      eduyrs      eduval

yeduc1       1.000

yeduc2       0.257       1.000
              0.0098

educat       0.883       0.204       1.000
              <.0001      0.0421

eduyrs       0.831       0.177       0.988       1.000
              <.0001      0.0784      <.0001

eduval       0.937       0.261       0.943       0.889       1.000
              <.0001      0.0088      <.0001      <.0001
```

The EDUVAL variable is the "right" answer and therefore has the highest correlation with the outcomes YEDUC1 and YEDUC2. For YEDUC1, treating education as equally-spaced, EDUCAT, does a little less well and treating education according to the years represented, EDUYRS, does a little less well than that. But there is not much practical difference among the values. For YEDUC2, the strength of the correlation is in the same relationship. However, we might now draw a different conclusion if we were hypothesis-testing, as the "right" coding has P<0.01, the equally-spaced alternative has P<0.05, and the coding according to years has P>0.05. The moral of this story? For me, it's that borderline P-values might move around a bit if you change the spacing, but mostly one coding is about as good as another with ordinal variable.

USING TWO OR MORE VARIABLES TO CAPTURE COMPLEX PATTERNS

When you find variables have a partial ordering, you can generally capture the ordering(s) by using two or more variables. Consider a predictor that can be characterized by an ordered string of 4 binary digits (0 or 1). It may be that the number of "1" values provides a partial ordering, but there is no *a priori* ordering of the placement of the 1s. You can create a variable that is the sum of the binary digits (i.e. the number of 1s) to include in the model along with the original variable as a categorical variable. You will be able to assess the linear effect of the partial ordering along with the details of the individual categories.

Another situation where two or more variables are useful is when working with a nearly-linear effect with distortion in some part of the range, usual at one extreme or another. You could use a continuous variable together with a binary (indicator) variable for cases that are zero. Similarly, you could use a continuous variable together with a set of binary variables to pull out outliers. Another common use would be to allow a top category – "9 or more" for example – to be a different distance from the next lower category than the spacing for the other part of the scale.

Piecewise constant and piecewise linear models can be constructed, too, and possibly combined with categorical variables for individual values. When a threshold effect is expected – either at the low or at the high end – it can be fit with a simple piecewise model. Generally it is desirable for the model to be continuous, but it can be worthwhile to test for discontinuities by allowing the values below (or above) a certain threshold to be fit separately. See Pasta (2005) for more on piecewise models.

One special case that comes up frequently is age. In some situations, age is given in years ("age at last birthday"). In other situations, age is calculated using the date of birth and the date of an event (a clinic visit, for example, or a standardized test). Without getting into the wisdom of using the inaccurate (but easy-to-code) approach of dividing the later SAS date minus the earlier SAS date by 365.25, at the least you will be faced with the decision of whether to use the original age variable or put age in categories and treat the categories as approximately linear. Certainly there are cases where an exact age is the way to go (in whole years, either age at least birthday or nearest birthday, or in fractional years). But I usually find myself dividing the ages into categories – often 5- or 10-year age groups, but sometimes narrower or wider intervals. I am then in a position to evaluate the (approximately) linear effect and the deviations from linearity in a reasonably straightforward way. It's hard to examine nonlinearities using exact age.

CONCLUSION

Before you treat your continuous measures as continuous variables (with linear effects), consider whether you should "discretize" them and treat them as categorical to better understand the relationships. Before you treat your discrete variables as categorical, consider whether you should at least evaluate the linear component by treating it as continuous. For variables with partial orderings, or with both linear and nonlinear components, consider combining continuous and categorical variables together. Soon you, too, will see predictor variables in shades of gray.

REFERENCES

Blalock, Hubert M. Jr. (1979), *Social Statistics*, Revised Second Edition, New York: McGraw-Hill

Cochran, W. G. (1968), "The effectiveness of adjustment by subclassification in removing bias in observational studies," *Biometrics,* 24:295-313.

Cochran, William G. (1983), *Planning and Analysis of Observational Studies*, Eds. Lincoln E. Moses and Frederick Mosteller, New York: John Wiley & Sons

Cox, D. R. (1957), "Note on grouping," *J. American Statistical Association,* 52:543-7.

Moses, Lincoln E., Emerson John D., and Hosseini, Hossein (1984), "Analyzing data from ordered categories," *New England Journal of Medicine*, 311:442-8. Reprinted as Chapter 13 in Bailar, John C. III and Mosteller, Frederick (1992) Medical Uses of Statistics, 2nd Ed., Boston, MA: NEJM Books

Pasta, David J. (2005), "Parameterizing models to test the hypotheses you want: coding indicator variables and modified continuous variables," Proceedings of the Thirtieth Annual SAS Users Group International Conference, Paper 212-30. http://www2.sas.com/proceedings/sugi30/212-30.pdf

Pritchard, Michelle L. and Pasta, David J. (2004), "Head of the CLASS: Impress your colleagues with a superior understanding of the CLASS statement in PROC LOGISTIC," Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference, Paper 194-29.
http://www2.sas.com/proceedings/sugi29/194-29.pdf

Velleman, Paul F. and Wilkinson, Leland (1993), "Nominal, ordinal, interval, and ratio typologies are misleading," *The American Statistician*, 47:1, 65-72.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David J. Pasta
Vice President, Statistics & Data Operations
ICON Clinical Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
(415) 371-2111
david.pasta@iconplc.com
www.iconplc.com