Paper 422-2013

Being Continuously Discrete (or Discretely Continuous): Understanding Models with Continuous and Discrete Predictors and Testing Associated Hypotheses

David J. Pasta, ICON Late Phase & Outcomes Research, San Francisco, CA

ABSTRACT

Often a general (or generalized) linear model has both discrete predictors (included in the CLASS statement) and continuous predictors. Binary variables can be treated either as continuous or discrete; the resulting models are equivalent but the interpretation of the parameters differs. In many cases, interactions between discrete and continuous variables are of interest. This paper provides practical suggestions for building and interpreting models with both continuous and discrete predictors. It includes some examples of the use of the STORE statement and PROC PLM to understand models and test hypotheses without repeating the estimation step.

INTRODUCTION

First let us be clear on what we mean by discrete and continuous predictors. A discrete (or categorical) predictor is one which is included in the CLASS statement. I use the terms discrete and categorical interchangeably in this context. The individual values are not assumed to have any particular relationship to each other: they are treated as just "names" for the categories and are not to be interpreted quantitatively even if they are numbers. Variables for which the categories are considered to be names without even a partial ordering are referred to as being nominal variables; I prefer to use the terms discrete or categorical to emphasize the way they are being used in the model rather than the nature of the variable itself. It is possible for the underlying variable to be continuous, but what is important for our purposes is that we want to estimate the effect of each value separately and not to assume specific spacing between values.

A continuous predictor is one for which the numeric values are treated as meaningful and the estimated coefficient is interpreted as the effect of a one-unit change. In practice, ordinal variables can be treated as discrete or as continuous (and sometimes profitably as both discrete and continuous in the same analysis). In addition, continuous variables can be grouped into categories and converted into discrete variables. This issue is discussed at length in Pasta (2009), but it is worthwhile to summarize a few of the points made there. Treating an ordinal variable as continuous allows you to estimate the linear component of the relationship, as recommended by Moses et al. (1984). On the other hand, treating an ordinal variable as discrete allows you to capture much more complicated relationships. It seems worthwhile to consider both aspects of the variable.

A WORD ABOUT BINARY VARIABLES

Binary variables take on exactly two values, such as 0 and 1 or True and False or Male and Female. For analysis purposes, they can be treated as continuous or discrete. Because you generally get an equivalent model whether or not the binary variables is in the CLASS statement, it is easy to get lazy about considering the implications. In fact whether a binary variable is treated as continuous or discrete affects the parameterization of models and therefore the interpretation of results and computational algorithms. It is especially important to remember with binary variables are continuous or discrete when interpreting least squares means (LSMEANS). ***Generally***, my recommendation is to treat binary variables as discrete (include them in the CLASS) statement, but sometimes you should treat them as continuous.

PARAMETERIZATIONS

Before getting into models that include both discrete and continuous variables and, more interestingly, their interactions, it is important to understand the way that models are parameterized in SAS®. This includes an understanding of the CLASS statement and both the default and the alternative parameterizations available. This material is covered in numerous places, including several of my papers from previous conferences (Pritchard and Pasta 2004; Pasta 2005; Pasta 2009; Pasta 2010).

One parameterization for discrete variables is the “less than full rank” approach in which dummy variables (indicator variables) are created for each category. This parameterization, also called the GLM parameterization, includes all the dummy variables but recognizes that there are redundancies and uses appropriate computational methods such as generalized inverses to obtain parameter estimates. The last category (as ordered using the formatted value) ends up as the reference category. To change the reference category it is necessary to reorder the categories of the variable.

It is now possible to specify the parameterization you want to use on the CLASS statement (but be aware that which procedures support this approach depends on which version of SAS you are using). You can specify REFERENCE coding, which allows you to specify a reference category which is omitted from the design matrix in various convenient ways. Alternatively you can specify EFFECT coding, which effectively compares each category to the overall average rather than to a single category, although there is still an omitted category that you can specify. My experience is that people find EFFECT coding rather confusing at first, so I recommend the use of REFERENCE coding. Note that LOGISTIC now uses EFFECT coding by default. You can specify different coding for different variables and different reference categories (the default is LAST), making it much easier to manipulate the parameterization of discrete variables.

LEAST SQUARES MEANS AND THE OBSMARGINS OPTION

When a model includes discrete variables, the parameter estimates are often difficult to interpret and the test that they are zero may not be of interest. The LSMEANS statement allows the calculation of least squares means, also called adjusted means, for the values of a variable (or of interactions among discrete variables). There are also options to compare least squares means with or without adjustments for multiple comparisons. One of the things to pay attention to when a model includes more than one predictor variable is whether to specify the OBSMARGINS option (abbreviated OM) on LSMEANS. This option causes the LSMEANS to use the observed marginal distributions of the variable rather than using equal coefficients across classification effects (thereby assuming balance among the levels). Sometimes you want one version and sometimes you want the other, but in my work I generally find that OBSMARGINS more often gives me the LSMEANS I want. The issue of estimability also arises (assuming the model is less than full rank). It is quite possible for the LSMEANS to be nonestimable with the OM option but estimable without, or vice versa. Some time spent understanding the model, together with some tools that SAS provides, make the determination of estimability less mysterious. See Pasta (2010) for additional details.

AN EXAMPLE: AN ORDINAL VARIABLE AS DISCRETE OR CONTINUOUS OR BOTH

An ordinal variable might be treated as discrete or continuous. It can also be profitably treated as both discrete and continuous in the same model. This approach can be used to test deviations from linearity. Let’s start by considering an ordinal variable, EDUCAT, which measures years of education in categories, as a categorical (discrete) predictor of our dependent variable Y:

Code for GLM

```
proc glm data=anal;
   class educat;
   model y = educat / solution;
   title1 'EDUCAT categorical with typical labels';
run;
```

EDUCAT categorical with typical labels

Dependent Variable: y

Source	DF	Sum of 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

Parameter		Estimate	Standard 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

Source	DF	Sum of 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

Parameter		Estimate	Standard 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 statistically significant. What happens if we treat education as a continuous variable? All we need to do is omit it from the CLASS statement. But let's make a copy of EDUCAT called L_EDUCAT (for "linear education") in anticipation of the next step:

EDUCAT continuous

Dependent Variable: y

Source	DF	Sum of 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

Parameter	Estimate	Standard 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!

How do you go about testing for deviations from linearity? It's actually pretty easy, but it leads to output that people find a little odd-looking at first. For any ordinal variable, (1) put 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. Here, we have EDUCAT with K=5 categories. We created L_EDUCAT (L for Linear) as an exact copy, 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=3 degrees of freedom; 1 df is lost to the overall constant, and 1 df is lost to the linear effect L_EDUCAT. There are some details to watch out for, best expressed by looking at some SAS output.

EDUCAT continuous and categorical

Dependent Variable: y

Source	DF	Sum of 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). 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.

BINARY VARIABLES – DISCRETE OR CONTINUOUS?

Binary variables might be treated as discrete or continuous. What difference does it make? Let's take a look at a fairly simple example. We have a dependent measure Y3 that we want to predict using RACE (which takes on values BLACK, HISPANIC, and WHITE) and SEX (which takes on values FEMALE and MALE). First look at a main-effects model with both RACE and SEX discrete. We look at the LSMEANS both with the default weighting (equal weighting) and with OBSMARGINS specified (abbreviated OM).

Code for GLM with discrete SEX

```
proc glm data=analsub;
   class sex race;
   model y3 = race sex
      / solution;
   lsmeans race sex / stderr;
   lsmeans race sex / stderr om;
   title3 'y3 = race sex (discrete) without and with OM';
run;
```

```
Dependent Variable: y3
                                        Sum of
Source                        DF        Squares     Mean Square    F Value    Pr > F
Model                          3     228506.315       76168.772      30.42    <.0001
Error                        835    2090548.362        2503.651
Corrected Total              838    2319054.676

R-Square     Coeff Var      Root MSE       y3 Mean
0.098534      29.44697      50.03649      169.9207

Source                        DF       Type I SS     Mean Square    F Value    Pr > F
race                           2     72961.0062      36480.5031      14.57    <.0001
sex                            1    155545.3085     155545.3085      62.13    <.0001

Source                        DF     Type III SS     Mean Square    F Value    Pr > F
race                           2     81271.8507      40635.9253      16.23    <.0001
sex                            1    155545.3085     155545.3085      62.13    <.0001
```

```
                                                    Standard
Parameter                    Estimate                  Error    t Value    Pr > |t|

Intercept                173.0371863 B          2.34513442      73.79      <.0001
race       Black          26.7820211 B          4.72268870       5.67      <.0001
race       Hispanic        7.9571591 B          4.88460275       1.63      0.1037
race       White           0.0000000 B                   .          .           .
sex        Female        -29.9219029 B          3.79618722      -7.88      <.0001
sex        Male            0.0000000 B                   .          .           .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
the normal equations.  Terms whose estimates are followed by the
      letter 'B' are not uniquely estimable.

Least Squares Means [without OM]
                                 Standard
race           y3 LSMEAN            Error    Pr > |t|

Black          184.858256         4.282844      <.0001
Hispanic       166.033394         4.446581      <.0001
White          158.076235         2.260039      <.0001

                               Standard
sex          y3 LSMEAN            Error    Pr > |t|

Female      154.695010         3.380595      <.0001
Male        184.616913         2.458757      <.0001

Least Squares Means [with OM]
                                 Standard
race           y3 LSMEAN            Error    Pr > |t|

Black          191.010256         4.230100      <.0001
Hispanic       172.185394         4.409031      <.0001
White          164.228235         2.096816      <.0001

                               Standard
sex          y3 LSMEAN            Error    Pr > |t|

Female      148.807723         3.187314      <.0001
Male        178.729625         2.057449      <.0001
```

Now consider the output with RACE discrete but SEX continuous (omitted from the CLASS statement). Now that SEX is not in the CLASS statement, we cannot use LSMEANS to get adjusted means for the different levels of SEX.

Code for GLM with continuous SEX

```
proc glm data=analsub;
   class race;
   model y3 = race sex
      / solution;
   lsmeans race / stderr;
   lsmeans race / stderr om;
   title3 'y3 = race sex (continuous) without and with OM';
run;
Dependent Variable: y3
                                       Sum of
Source                          DF         Squares     Mean Square    F Value    Pr > F
Model                            3      228506.315       76168.772      30.42    <.0001
Error                          835     2090548.362        2503.651
Corrected Total                838     2319054.676
```

```
R-Square     Coeff Var      Root MSE       y3 Mean
0.098534      29.44697      50.03649      169.9207

Source                     DF       Type I SS     Mean Square    F Value    Pr > F
race                        2      72961.0062      36480.5031      14.57    <.0001
sex                         1     155545.3085     155545.3085      62.13    <.0001

Source                     DF     Type III SS     Mean Square    F Value    Pr > F
race                        2      81271.8507      40635.9253      16.23    <.0001
sex                         1     155545.3085     155545.3085      62.13    <.0001

                                                    Standard
Parameter                    Estimate                  Error    t Value    Pr > |t|

Intercept                113.1933804 B          6.86769622      16.48      <.0001
race       Black          26.7820211 B          4.72268870       5.67      <.0001
race       Hispanic        7.9571591 B          4.88460275       1.63      0.1037
race       White           0.0000000 B           .                .         .
sex                       29.9219029            3.79618722       7.88      <.0001

Least Squares Means [without OM]
                                 Standard
race            y3 LSMEAN           Error    Pr > |t|

Black          191.010256        4.230100      <.0001
Hispanic       172.185394        4.409031      <.0001
White          164.228235        2.096816      <.0001

Least Squares Means [with OM]
                                 Standard
race            y3 LSMEAN           Error    Pr > |t|

Black          191.010256        4.230100      <.0001
Hispanic       172.185394        4.409031      <.0001
White          164.228235        2.096816      <.0001
```

It appears that SAS has ignored our OBSMARGINS (OM) specification. Actually, because SAS sets continuous variables at their mean when calculating the least squares means for discrete variables, binary variables treated as continuous act similarly to having specified the OBSMARGINS option. That is, instead of assuming the binary variable is balanced (half the cases at one value and half at the other value), SAS uses the observed mean value. If we want to imitate the effect of omitting OM, we could calculate least squares means with the binary variables set to the average of the two values (0.5 if it is coded 0 and 1).

INTERACTING DISCRETE AND BINARY VARIABLES

When we interact binary variables that are treated as continuous with discrete variables, we lose some of the power of the LSMEANS statement. For example, if we add in the RACE*SEX interaction with SEX continuous, we cannot get the least squares means for the various combinations of RACE and SEX.

Code for GLM

```
proc glm data=analsub;
   class race;
   model y3 = race sex race*sex
      / solution;
   lsmeans race / stderr;
   lsmeans race / stderr om;
   title3 'y3 = race sex (continuous) race*sex without and with OM';
run;
```

```
Dependent Variable: y3
                                        Sum of
Source                          DF         Squares     Mean Square    F Value    Pr > F
Model                            5      273774.590       54754.918      22.30    <.0001
Error                          833     2045280.087        2455.318
Corrected Total                838     2319054.676

R-Square     Coeff Var      Root MSE       y3 Mean
0.118054      29.16135      49.55117      169.9207

Source                          DF       Type I SS     Mean Square    F Value    Pr > F
race                             2      72961.0062      36480.5031      14.86    <.0001
sex                              1     155545.3085     155545.3085      63.35    <.0001
sex*race                         2      45268.2749      22634.1375       9.22    0.0001

Source                          DF     Type III SS     Mean Square    F Value    Pr > F
race                             2      17870.4231       8935.2116       3.64    0.0267
sex                              1     161037.6917     161037.6917      65.59    <.0001
sex*race                         2      45268.2749      22634.1375       9.22    0.0001

                                                    Standard
Parameter                      Estimate                Error    t Value    Pr > |t|

Intercept                 126.5119270 B      8.25345173      15.33      <.0001
race      Black           -45.1835486 B     17.66961884      -2.56      0.0107
race      Hispanic          5.1670020 B     17.89458887       0.29      0.7728
race      White             0.0000000 B       .                .         .
sex                        22.1911783 B      4.63675523       4.79      <.0001
sex*race  Black            42.6693767 B     10.09492850       4.23      <.0001
sex*race  Hispanic          1.3841721 B     10.30476992       0.13      0.8932
sex*race  White             0.0000000 B       .                .         .

Least Squares Means [without OM]
                                  Standard
race           y3 LSMEAN             Error    Pr > |t|

Black         191.954665        4.194847      <.0001
Hispanic      171.889092        4.383842      <.0001
White         164.361243        2.077003      <.0001

Least Squares Means [with OM]
                                  Standard
race           y3 LSMEAN             Error    Pr > |t|

Black         191.954665        4.194847      <.0001
Hispanic      171.889092        4.383842      <.0001
White         164.361243        2.077003      <.0001
```

When we treat SEX as discrete, we can get more least squares means and the OBSMARGINS option has an effect. The model is equivalent but is parameterized rather differently.

Code for GLM

```
proc glm data=analsub;
   class sex race;
   model y3 = race sex race*sex
      / solution;
   lsmeans race sex race*sex / stderr;
   lsmeans race sex race*sex / stderr om;
   title3 'y3 = race sex (discrete) race*sex without and with OM';
run;
```

Dependent Variable: y3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	273774.590	54754.918	22.30	<.0001
Error	833	2045280.087	2455.318		
Corrected Total	838	2319054.676			

R-Square	Coeff Var	Root MSE	y3 Mean
0.118054	29.16135	49.55117	169.9207

Source	DF	Type I SS	Mean Square	F Value	Pr > F
race	2	72961.0062	36480.5031	14.86	<.0001
sex	1	155545.3085	155545.3085	63.35	<.0001
sex*race	2	45268.2749	22634.1375	9.22	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
race	2	35171.7257	17585.8629	7.16	0.0008
sex	1	161037.6917	161037.6917	65.59	<.0001
sex*race	2	45268.2749	22634.1375	9.22	0.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		170.8942836	B	2.44121082	70.00	<.0001
race	Black	40.1552047	B	5.63958911	7.12	<.0001
race	Hispanic	7.9353461	B	5.90301678	1.34	0.1792
race	White	0.0000000	B	.	.	.
sex	Female	-22.1911783	B	4.63675523	-4.79	<.0001
sex	Male	0.0000000	B	.	.	.
sex*race	Female Black	-42.6693767	B	10.09492850	-4.23	<.0001
sex*race	Female Hispanic	-1.3841721	B	10.30476992	-0.13	0.8932
sex*race	Female White	0.0000000	B	.	.	.
sex*race	Male Black	0.0000000	B	.	.	.
sex*race	Male Hispanic	0.0000000	B	.	.	.
sex*race	Male White	0.0000000	B	.	.	.

Least Squares Means **[without OM]**

race	y3 LSMEAN	Standard Error	Pr > |t|
Black	178.619211	4.483528	<.0001
Hispanic	167.041955	4.601325	<.0001
White	159.798694	2.318378	<.0001

sex	y3 LSMEAN	Standard Error	Pr > |t|
Female	150.048773	3.740251	<.0001
Male	186.924467	2.596816	<.0001

sex	race	y3 LSMEAN	Standard Error	Pr > |t|
Female	Black	146.188933	7.386652	<.0001
Female	Hispanic	155.254279	7.470120	<.0001
Female	White	148.703105	3.942079	<.0001
Male	Black	211.049488	5.083843	<.0001
Male	Hispanic	178.829630	5.374579	<.0001
Male	White	170.894284	2.441211	<.0001

Least Squares Means **[with OM]**

race	y3 LSMEAN
Black	Non-est
Hispanic	Non-est
White	Non-est

sex	y3 LSMEAN
Female	Non-est
Male	Non-est

sex	race	y3 LSMEAN	Standard Error	Pr > |t|
Female	Black	146.188933	7.386652	<.0001
Female	Hispanic	155.254279	7.470120	<.0001
Female	White	148.703105	3.942079	<.0001
Male	Black	211.049488	5.083843	<.0001
Male	Hispanic	178.829630	5.374579	<.0001
Male	White	170.894284	2.441211	<.0001

However, when we specify OM the least squares means for RACE and SEX are non-estimable. It turns out the marginal means for RACE and SEX are non-estimable because the distribution of RACE is different for males and females (or, equivalently, the proportion of males and females varies by race). For more on this, see Pasta (2010). But how did we get RACE least squares means when we treated SEX as continuous (which is similar to using the OM option)? LSMEANS sets all the continuous variables at their grand mean (not their subgroup mean), so the overall mean of the continuous variable SEX is used. One could do the equivalent thing when treating SEX as discrete by using an ESTIMATE statement and specifying the coefficients for SEX as the overall proportions for male and female.

INTERACTIONS BETWEEN DISCRETE AND CONTINUOUS VARIABLES

When you have a mix of discrete and continuous variables, it is sometimes quite handy to include interactions without main effects. Consider a discrete variable A and a continuous variable X. If you include A, X, and A*X you get a test of the interaction between A and X which essentially asks whether the effect of X is parallel for the different levels of A. If you find the interaction is statistically significant, it may be easier to interpret the SOLUTION if you model A and A*X. Then the parameters estimated for A*X are the slopes for each of the individual levels of A rather than deviations from the slope of the reference category of A (which is what you get when you model A, X, and A*X).

With a more complicated high-order interaction between two discrete variables and one continuous variable, such as A*B*X, it is almost always a good idea to include A*B in the model (this allows the effect of X to vary across the various levels of A*B without any imposed structure). However it may not be especially useful to include A*X and B*X as separate terms once you have established that A*B*X is significant in the presence of those terms. It's easier to see the results if you have just A, B, A*B, and A*B*X as the terms in the model.

Here's an extended example, as usual using simulated data. First we run a model with just the discrete variables. We find we have a sex*race interaction.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
race	2	35171.7257	17585.8629	7.16	0.0008
sex	1	161037.6917	161037.6917	65.59	<.0001
sex*race	2	45268.2749	22634.1375	9.22	0.0001

The SOLUTION is as usual a bit hard to read:

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		170.8942836 B	2.44121082	70.00	<.0001
race	Black	40.1552047 B	5.63958911	7.12	<.0001
race	Hispanic	7.9353461 B	5.90301678	1.34	0.1792
race	White	0.0000000 B	.	.	.
sex	Female	-22.1911783 B	4.63675523	-4.79	<.0001
sex	Male	0.0000000 B	.	.	.
sex*race	Female Black	-42.6693767 B	10.09492850	-4.23	<.0001
sex*race	Female Hispanic	-1.3841721 B	10.30476992	-0.13	0.8932
sex*race	Female White	0.0000000 B	.	.	.
sex*race	Male Black	0.0000000 B	.	.	.
sex*race	Male Hispanic	0.0000000 B	.	.	.
sex*race	Male White	0.0000000 B	.	.	.

However, the LSMEANS make it easier to see what is going on. These are the same whether you specify OM or not because these are the fully interacted values so we're just fitting the marginal means. The LSMEANS for SEX and RACE are also available without the OM option but if you specify OM they are non-estimable for the reasons given above.

sex	race	y3 LSMEAN	Standard Error	Pr > |t|
Female	Black	146.188933	7.386652	<.0001
Female	Hispanic	155.254279	7.470120	<.0001
Female	White	148.703105	3.942079	<.0001
Male	Black	211.049488	5.083843	<.0001
Male	Hispanic	178.829630	5.374579	<.0001
Male	White	170.894284	2.441211	<.0001

Now let's introduce the continuous variable, EDUYRS.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
race	2	35558.6216	17779.3108	7.35	0.0007
sex	1	161148.5447	161148.5447	66.65	<.0001
sex*race	2	49930.0508	24965.0254	10.33	<.0001
eduyrs	1	33688.3103	33688.3103	13.93	0.0002

It definitely has an effect – what about interactions? Start with a full model.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
race	2	3847.52395	1923.76198	0.81	0.4462
sex	1	7063.70798	7063.70798	2.97	0.0854
sex*race	2	3756.39302	1878.19651	0.79	0.4548
eduyrs	1	11307.44124	11307.44124	4.75	0.0296
eduyrs*race	2	4617.62450	2308.81225	0.97	0.3797
eduyrs*sex	1	29523.41710	29523.41710	12.40	0.0005
eduyrs*sex*race	2	10014.48954	5007.24477	2.10	0.1228

It looks like we can eliminate the EDUYRS*SEX*RACE interaction.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
race	2	7222.18410	3611.09205	1.51	0.2210
sex	1	1591.31036	1591.31036	0.67	0.4146
sex*race	2	44542.78631	22271.39315	9.33	<.0001
eduyrs	1	15655.34469	15655.34469	6.56	0.0106
eduyrs*race	2	13573.55566	6786.77783	2.84	0.0589
eduyrs*sex	1	20718.49220	20718.49220	8.68	0.0033

It is borderline and under some circumstances we might want to keep it, but for this example let's eliminate the EDUYRS*RACE interaction.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
race	2	39430.31041	19715.15520	8.22	0.0003
sex	1	1058.00228	1058.00228	0.44	0.5068
sex*race	2	44888.94205	22444.47103	9.36	<.0001
eduyrs	1	6599.23905	6599.23905	2.75	0.0976
eduyrs*sex	1	18373.47865	18373.47865	7.66	0.0058

This looks like it might be a reasonable model. What about the nonsignificant F test for EDUYRS? As a Type III test, it is testing the EDUYRS effect in the presence of the EDUYRS*SEX interaction and is generally not of importance. Here is the associated SOLUTION.

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		129.9580466	B	9.19099945	14.14	<.0001
race	Black	41.3190203	B	5.57974484	7.41	<.0001
race	Hispanic	12.0077010	B	5.90073985	2.03	0.0422
race	White	0.0000000	B	.	.	.
sex	Female	28.5285226	B	18.42262692	1.55	0.1219
sex	Male	0.0000000	B	.	.	.
sex*race	Female Black	-43.2573867	B	10.02230453	-4.32	<.0001
sex*race	Female Hispanic	-5.9127486	B	10.24861249	-0.58	0.5641
sex*race	Female White	0.0000000	B	.	.	.
sex*race	Male Black	0.0000000	B	.	.	.
sex*race	Male Hispanic	0.0000000	B	.	.	.
sex*race	Male White	0.0000000	B	.	.	.
eduyrs		2.8537614	B	0.61825338	4.62	<.0001
eduyrs*sex	Female	-3.5687416	B	1.28942573	-2.77	0.0058
eduyrs*sex	Male	0.0000000	B	.	.	.

It is easier to interpret this if we remove the EDUYRS main effect. The ANOVA table and the SOLUTION become the following:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
race	2	39430.31041	19715.15520	8.22	0.0003
sex	1	1058.00228	1058.00228	0.44	0.5068
sex*race	2	44888.94205	22444.47103	9.36	<.0001
eduyrs*sex	2	52061.78894	26030.89447	10.85	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		129.9580466	B	9.19099945	14.14	<.0001
race	Black	41.3190203	B	5.57974484	7.41	<.0001
race	Hispanic	12.0077010	B	5.90073985	2.03	0.0422
race	White	0.0000000	B	.	.	.
sex	Female	28.5285226	B	18.42262692	1.55	0.1219
sex	Male	0.0000000	B	.	.	.
sex*race	Female Black	-43.2573867	B	10.02230453	-4.32	<.0001
sex*race	Female Hispanic	-5.9127486	B	10.24861249	-0.58	0.5641
sex*race	Female White	0.0000000	B	.	.	.
sex*race	Male Black	0.0000000	B	.	.	.
sex*race	Male Hispanic	0.0000000	B	.	.	.
sex*race	Male White	0.0000000	B	.	.	.
eduyrs*sex	Female	-0.7149802		1.13153942	-0.63	0.5276
eduyrs*sex	Male	2.8537614		0.61825338	4.62	<.0001

We have combined the tests for EDUYRS into a single test with 2 degrees of freedom. More importantly, we now can easily read the coefficients for the slope of EDUYRS for Females and Males. Compare the two sets of estimates associated with EDUYRS:

Parameter		Estimate		Standard Error	t Value	Pr > |t|
eduyrs		2.8537614	B	0.61825338	4.62	<.0001
eduyrs*sex	Female	-3.5687416	B	1.28942573	-2.77	0.0058
eduyrs*sex	Male	0.0000000	B	.	.	.

Parameter		Estimate		Standard Error	t Value	Pr > |t|
eduyrs*sex	Female	-0.7149802		1.13153942	-0.63	0.5276
eduyrs*sex	Male	2.8537614		0.61825338	4.62	<.0001

For the first set of values the coefficient for EDUYRS is the estimate for Males and the second one is the difference Females minus Males. That provides a convenient test of the difference in slopes, which has t=-2.77 and P=0.0058. This corresponds to the F test in the ANOVA table with F=7.66 and P=0.0058. It is in fact the equivalent test (the square of a t is an F). The second set of values gives us the actual estimates for the slope of EDUYRS for Female and for Male and tests each against zero. The test for Female is nonsignificant, so under some circumstances it might be worth treating it as zero and including the EDUYRS effect only for males. This would produce a slightly different overall model with 6 instead of 7 parameters. It turns out this matches the model used to create the data.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
race	2	38797.68438	19398.84219	8.09	0.0003
sex	1	139.22315	139.22315	0.06	0.8096
sex*race	2	46624.30159	23312.15080	9.73	<.0001
male*eduyrs	1	51104.14833	51104.14833	21.32	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		129.9580466	B	9.18768067	14.14	<.0001
race	Black	41.3190203	B	5.57773004	7.41	<.0001
race	Hispanic	12.0077010	B	5.89860914	2.04	0.0421
race	White	0.0000000	B	.	.	.
sex	Female	18.7450587	B	9.97914766	1.88	0.0607
sex	Male	0.0000000	B	.	.	.

```
sex*race    Female Black         -43.8331922 B      9.97718550      -4.39      <.0001
sex*race    Female Hispanic       -5.4565270 B     10.21945577      -0.53      0.5935
sex*race    Female White           0.0000000 B       .               .         .
sex*race    Male Black             0.0000000 B       .               .         .
sex*race    Male Hispanic          0.0000000 B       .               .         .
sex*race    Male White             0.0000000 B       .               .         .
male*eduyrs                        2.8537614        0.61803013      4.62      <.0001
```

One note of caution. Be careful about using two names for related variables – you can confuse SAS if you're not careful (and get wrong answers). Here we use MALE as a zero-one dummy variable that is essentially the same as the SEX variable. It turns out everything is fine here, but if we were to mix the two variables in constructing interactions we could get into trouble.

USING THE STORE STATEMENT AND PROC PLM TO EXPLORE YOUR MODELS

Recently SAS added the STORE statement to a variety of procedures and the associated PROC PLM. STORE allows you to specify a location to save an "item store," which is a special SAS file containing the context and results of the statistical analysis. You can then use PROC PLM to manipulate the results from that analysis without having to repeat the calculations necessary to fit the model. For example, suppose you run a complex analysis and are happy with the results (no doubt after several tries) and save the results. You can then retrieve the item store and examine the model in various ways, for example specifying LSMEANS statements and a variety of ESTIMATE or CONTRAST statements. You can use ODS Graphics to look at graphical displays of the fitted model. They are perfectly easy to use and I believe that using STORE should become standard practice.

Here is a very simple example from the models we were just fitting. First here is the code where we use STORE to save the item store:

Code for GLM with STORE

```
proc glm data=analsub;
   class sex race;
   model y3 = race sex race*sex eduyrs sex*eduyrs / solution;
   store s_mod01;
   title3 'mod01: y3 = ... eduyrs sex*eduyrs';
run;

proc glm data=analsub;
   class sex race;
   model y3 = race sex race*sex sex*eduyrs / solution ;
   store s_mod02;
   title3 'mod02: y3 = ... sex*eduyrs';
run;

proc glm data=analsub;
   class sex race;
   model y3 = race sex race*sex male*eduyrs sex*eduyrs / solution;
   store s_mod03;
   title3 'mod03: y3 = ... male*eduyrs sex*eduyrs';
run;

proc glm data=analsub;
   class sex race;
   model y3 = race sex race*sex male*eduyrs / solution;
   store s_mod04;
   title3 'mod04: y3 = ... eduyrs*male';
run;
```

Now we can use PROC PLM to look at the models. First consider the LSMEANS statement. If we want to use the OBSMARGINS option, we need to point SAS to the original dataset from which the "observed" margins should be obtained. That is not part of the item store. We use the E option on LSMEANS to find out what the weights are that SAS is using for the LSMEANS.

Code for GLM with STORE and associated output

```
proc plm source=s_mod01;
   lsmeans race sex race*sex / obsmargins=analsub e;
   title3 'lsmeans from s_mod01 with /e';
run;
```

Coefficients for race Least Squares Means Using WORK.ANALSUB Margins

Effect	sex	race	Row1	Row2	Row3
Intercept			1	1	1
race		Black	1		
race		Hispanic		1	
race		White			1
sex	Female		0.2944	0.2944	0.2944
sex	Male		0.7056	0.7056	0.7056
sex*race	Female	Black	0.3214		
sex*race	Female	Hispanic		0.3411	
sex*race	Female	White			0.2772
sex*race	Male	Black	0.6786		
sex*race	Male	Hispanic		0.6589	
sex*race	Male	White			0.7228
eduyrs			13.969	13.969	13.969
eduyrs*sex	Female		4.1125	4.1125	4.1125
eduyrs*sex	Male		9.8566	9.8566	9.8566

race Least Squares Means

race	Margins	Estimate	Standard Error	DF	t Value	Pr > |t|
Black	WORK.ANALSUB	Non-est	.	.	.	.
Hispanic	WORK.ANALSUB	Non-est	.	.	.	.
White	WORK.ANALSUB	Non-est	.	.	.	.

Coefficients for sex Least Squares Means Using WORK.ANALSUB Margins

Effect	sex	race	Row1	Row2
Intercept			1	1
race		Black	0.1669	0.1669
race		Hispanic	0.1538	0.1538
race		White	0.6794	0.6794
sex	Female		1	
sex	Male			1
sex*race	Female	Black	0.1822	
sex*race	Female	Hispanic	0.1781	
sex*race	Female	White	0.6397	
sex*race	Male	Black		0.1605
sex*race	Male	Hispanic		0.1436
sex*race	Male	White		0.6959
eduyrs			13.969	13.969
eduyrs*sex	Female		13.969	
eduyrs*sex	Male			13.969

sex Least Squares Means

sex	Margins	Estimate	Standard Error	DF	t Value	Pr > |t|
Female	WORK.ANALSUB	Non-est	.	.	.	.
Male	WORK.ANALSUB	Non-est	.	.	.	.

Coefficients for sex*race Least Squares Means Using WORK.ANALSUB Margins

Effect	sex	race	Row1	Row2	Row3	Row4	Row5	Row6
Intercept			1	1	1	1	1	1
race		Black	1			1		
race		Hispanic		1			1	
race		White			1		1	
sex	Female		1	1	1			
sex	Male					1	1	1
sex*race	Female	Black	1					
sex*race	Female	Hispanic		1				
sex*race	Female	White			1			
sex*race	Male	Black				1		
sex*race	Male	Hispanic					1	
sex*race	Male	White					1	
eduyrs			13.969	13.969	13.969	13.969	13.969	13.969
eduyrs*sex	Female		13.969	13.969	13.969			
eduyrs*sex	Male					13.969	13.969	13.969

sex*race Least Squares Means

sex	race	Margins	Estimate	Standard Error	DF	t Value	Pr > |t|
Female	Black	WORK.ANALSUB	146.56	7.3245	831	20.01	<.0001
Female	Hispanic	WORK.ANALSUB	154.59	7.4569	831	20.73	<.0001
Female	White	WORK.ANALSUB	148.50	3.9096	831	37.98	<.0001
Male	Black	WORK.ANALSUB	211.14	5.0248	831	42.02	<.0001
Male	Hispanic	WORK.ANALSUB	181.83	5.3517	831	33.98	<.0001
Male	White	WORK.ANALSUB	169.82	2.4240	831	70.06	<.0001

Let's say we're interested in the estimates for white males and females. We can mimic the LSMEANS values in an ESTIMATE statement by using the mean of EDUYRS, which we can see from the LSMEANS output is 13.969. (We could also calculate this and obtain it to greater accuracy, but 13.969 is close enough.) It is worth doing so just to make sure your code is right. For more on coding ESTIMATE statements, see my earlier papers. But suppose we're actually interested in the estimates for a high school graduate, with EDUYRS=12. Here's how we can do that with PROC PLM:

Code for PLM

```
proc plm source=s_mod01;
   estimate 'mod01 at 13.969 white female' intercept 1 race 0 0 1 sex 1 0 sex*race 0 0 1  0 0 0
      eduyrs 13.969 sex*eduyrs 13.969 0;
   estimate 'mod01 at 13.969 white male'   intercept 1 race 0 0 1 sex 0 1 sex*race 0 0 0  0 0 1
      eduyrs 13.969 sex*eduyrs 0 13.969;
   estimate 'mod01 at 12 white female' intercept 1 race 0 0 1 sex 1 0 sex*race 0 0 1  0 0 0
      eduyrs 12 sex*eduyrs 12 0;
   estimate 'mod01 at 12 white male'   intercept 1 race 0 0 1 sex 0 1 sex*race 0 0 0  0 0 1
      eduyrs 12 sex*eduyrs 0 12;
   title3 'plm using s_mod01';
run;
```

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
mod01 at 13.969 white female	148.50	3.9096	831	37.98	<.0001

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
mod01 at 13.969 white male	169.82	2.4240	831	70.06	<.0001

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
mod01 at 12 white female	149.91	4.3370	831	34.56	<.0001

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
mod01 at 12 white male	164.20	2.8148	831	58.34	<.0001

Now we are in a position to do similar calculations for the other three models:

Code for PLM

```
proc plm source=s_mod02;
   estimate 'mod02 at 12 white female' intercept 1 race 0 0 1 sex 1 0 sex*race 0 0 1  0 0 0
      sex*eduyrs 12 0;
   estimate 'mod02 at 12 while male'   intercept 1 race 0 0 1 sex 0 1 sex*race 0 0 0  0 0 1
      sex*eduyrs 0 12;
   title3 'plm using s_mod02';
run;

proc plm source=s_mod03;
   estimate 'mod03 at 12 white female' intercept 1 race 0 0 1 sex 1 0 sex*race 0 0 1  0 0 0
      male*eduyrs  0 sex*eduyrs 12 0;
   estimate 'mod03 at 12 while male'   intercept 1 race 0 0 1 sex 0 1 sex*race 0 0 0  0 0 1
      male*eduyrs 12 sex*eduyrs 0 12;
   title3 'plm using s_mod03';
run;

proc plm source=s_mod04;
   estimate 'mod04 at 12 white female' intercept 1 race 0 0 1 sex 1 0 sex*race 0 0 1  0 0 0
      male*eduyrs  0;
   estimate 'mod04 at 12 while male'   intercept 1 race 0 0 1 sex 0 1 sex*race 0 0 0  0 0 1
      male*eduyrs 12;
   title3 'plm using s_mod04';
run;
```

And here are the summarized results for all four models:

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
mod01 at 12 white female	149.91	4.3370	831	34.56	<.0001
mod01 at 12 white male	164.20	2.8148	831	58.34	<.0001
mod02 at 12 white female	149.91	4.3370	831	34.56	<.0001
mod02 at 12 while male	164.20	2.8148	831	58.34	<.0001
mod03 at 12 white female	149.91	4.3370	831	34.56	<.0001
mod03 at 12 white male	164.20	2.8148	831	58.34	<.0001
mod04 at 12 white female	148.70	3.8949	832	38.18	<.0001
mod04 at 12 white male	164.20	2.8138	832	58.36	<.0001

The first three models give the same values; they are equivalent models parameterized differently, so the ESTIMATE statements look different but the results are the same. The last model is slightly different (having eliminated a nonsignificant term) but gives similar results for the WHITE group at EDUYRS=12.

CONCLUSION

When you have both discrete (categorical) and continuous variables in your model, it is likely that you will want to explore interactions among them. Ordinal and binary variables might be treated as discrete or continuous, depending on exactly what hypotheses you want to test. There are many tools in SAS for building and understanding models that include both discrete and continuous variables. The STORE statement and PROC PLM provide a convenient method for estimating combinations of parameters and testing hypotheses without having to re-estimate a model. I believe that using STORE whenever a complex model is estimated should become a habit.

REFERENCES

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, 212-30 http://www2.sas.com/proceedings/sugi30/212-30.pdf

Pasta David J. (2009), “Learning when to be discrete: continuous vs. categorical predictors,” Proceedings of the SAS Global Forum 2009, 248-2009 http://support.sas.com/resources/papers/proceedings09/248-2009.pdf

Pasta, David J. (2010), “Practicalities of using ESTIMATE and CONTRAST statements,” Proceedings of the SAS Global Forum 2010, 269-2010 http://support.sas.com/resources/papers/proceedings10/269-2010.pdf

Pasta, David J. (2011), “Those confounded interactions: Building and interpreting a model with many potential confounders and interactions,” Proceedings of the SAS Global Forum 2011, 347-2011 http://support.sas.com/resources/papers/proceedings11/347-2011.pdf

Potter, Lori and Pasta, David J (1997), "The sum of squares are all the same—how can the LSMEANS be so different?", Proceedings of the Fifth Annual Western Users of SAS Software Regional Users Group Conference, San Francisco: Western Users of SAS Software

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, 194-29 http://www2.sas.com/proceedings/sugi29/194-29.pdf

ACKNOWLEDGEMENT

Some of the material in this paper previously appeared in Pasta (2009), Pasta (2010), and Pasta (2011). My thanks to my coauthors on previous papers , Stefanie Silva Millar, Lori Potter, and Michelle Pritchard Turner, for their help.

CONTACT INFORMATION

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

David J. Pasta
Vice President, Statistical & Strategic Analysis
ICON Late Phase & Outcomes Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
(415) 371-2111
david.pasta@iconplc.com
www.iconplc.com