Paper 347-2011

Those Confounded Interactions: Building and Interpreting a Model with Many Potential Confounders and Interactions

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

ABSTRACT

When constructing a general (or generalized) linear model with many possible covariates, it is often important to consider very carefully the role of potential confounders and interactions. Once a model is constructed, the interpretation of the coefficients depends on which other terms are included in the model. It is especially important to interpret interactions carefully if the model does not include all lower-order interactions and main effects. Depending on the context, you may want to set a low threshold for statistical significance for interactions (including them even if they are of borderline significance) or a high threshold (including them only if they are highly significant). Some variables you may want to treat as concomitant, including them regardless of statistical significance. Important confounders may fall into this category. This paper provides practical suggestions for building and interpreting models with interactions.

INTRODUCTION

Building a statistical model often includes both art and science. When you have a carefully-designed, randomized study, often the model is established in the design phase. In that situation, the art may be more in the design of the study than in the design of the analysis. Nonetheless, there may be specific interactions between variables that need to be assessed. Some experimental designs have certain factors that are "confounded" with others. For example, in an incomplete block design it may be that it is not possible to provide a separate test of certain interactions: they may be completely confounded with other effects. This technical use of the term confounding in the context of experimental design is an important one and is discussed at length in classic texts such as Chapter 6 of Cochran and Cox (1957).

The term "confounding" takes on a slightly different meaning in the context of observational research. A confounding variable (confounder) is one which is associated with the outcome (dependent variable) and also a specific predictor of interest (independent variable). Failure to take account of (control for) a confounder can result in erroneous conclusions. It is for this reason that observational research focuses as much attention as it does on identifying and adjusting for potential confounders.

Note that if a variable is associated only with the outcome or only with the predictor of interest (but not with both), it is not a confounder. A variable which is associated with the outcome may be useful to include in the model (to reduce residual variance, for example), but omitting it does not confound the effect of the predictor of interest. Similarly, a variable which is associated with the predictor of interest but has no effect on the outcome is not a confounder.

One of the most recognizable types of confounding is "confounding by indication." This is jargon for "sick people get treatments." If you simply measured infections and antibiotic treatments, you might well be led to the conclusion that antibiotics cause infections. Or that if you avoid hospitals you're much less likely to die.

Statistical interactions can take many forms. One form is when the effect of one predictor is modified according to the value of another predictor. For two characteristics, it may be that either characteristic alone has no effect but when both characteristics are present there is a measurable effect. Or it may be that either variable alone produces an effect but having both present does not increase the effect. More commonly, the magnitude of the effect may be only somewhat different (but statistically significantly different) for combinations of two (or more) variables than one would expect from the effect of each variable alone.

In building a statistical model, you may be concerned with both confounders and interactions. You may be especially concerned with interactions involving confounders, but understanding all the interactions in a model is key to the interpretation of the model. It is sometimes difficult to understand models in the presence of interactions, especially when not all possible interactions are included in the model.

BASICS OF INTERACTIONS AND THE OBSMARGINS OPTION ON LSMEANS

As a preface to understanding interactions and their interpretation, 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).

Models with a single predictor variable (factor) are reasonably straightforward. However, once you go beyond a single factor, things get more complicated. Among other things, the question arises whether to use 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.

Let's consider an example with two categorical variables, race (with five levels) and sex (with two levels). In order to get a handle on estimability, we can ask for the general form of all estimable functions by include the E option on the MODEL statement. To illustrate the difference, we include one LSMEANS statement with the OM option and one without:

CODE FOR GLM:

```
proc glm data=anal;
class race sex;
model y1 = race sex / solution e;
lsmeans race sex / stderr tdiff e;
lsmeans race sex / stderr tdiff e om;
title3 'GLM by race sex';
run;
```

FROM PROC GLM:

```
GLM by race sex
                Class Level Information
Class         Levels    Values
race               5    Asian Black Hispanic Other White
sex                2    Female Male

Number of Observations Read        1000
Number of Observations Used        1000

General Form of Estimable Functions
Effect                 Coefficients
Intercept              L1
race      Asian        L2
race      Black        L3
race      Hispanic     L4
race      Other        L5
race      White        L1-L2-L3-L4-L5
sex       Female       L7
sex       Male         L1-L7
```

Notice that the levels of RACE are in alphabetic order by formatted value. The interpretation of the "General Form of Estimable Functions," obtained by specifying the E option on the MODEL statement, is that the coefficients given as L followed by a number can be assigned any numerical value, but that the coefficient for some effects are derivable from the others. For example, if L2 is assigned a 1 and all the other coefficients are assigned 0, then for the function to be estimable the coefficient for White would

need to be -1. This would then estimate the difference between Asian and White. If you wanted just the value for Asian, you could assign L1 and L2 a value of 1, and the estimate would be of the Intercept plus Asian (and the coefficient for White would be 1-1=0).

```
Dependent Variable: y1
                                      Sum of
Source                        DF         Squares     Mean Square    F Value    Pr > F
Model                          5      82152.751       16430.550      12.94    <.0001
Error                        994    1262544.332        1270.165
Corrected Total              999    1344697.083

Source                        DF     Type III SS     Mean Square    F Value    Pr > F
race                           4     80028.47498     20007.11875      15.75    <.0001
sex                            1      2590.81573      2590.81573       2.04    0.1535


                                                 Standard
Parameter                      Estimate              Error    t Value    Pr > |t|
Intercept                63.39550656 B       1.65404269      38.33      <.0001
race       Asian          0.96269902 B       3.79755107       0.25      0.7999
race       Black         15.72610471 B       3.33992898       4.71      <.0001
race       Hispanic      23.67301601 B       3.48135125       6.80      <.0001
race       Other         -2.94225607 B       5.41062521      -0.54      0.5867
race       White          0.00000000 B        .                .         .
sex        Female        -3.43990976 B       2.40856801      -1.43      0.1535
sex        Male           0.00000000 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.
```

The SOLUTION, given above, includes tests of the difference between each race and the White category (the last category). These are more attractively displayed using the LSMEANS and the TDIFF option. For more information about SOLUTION, see Usage Note *38384:* How to interpret the results of the SOLUTION option in the MODEL statement of PROC GLM? at //support.sas.com/notes/index.html.

The first LSMEANS are *without* the OM option:

```
Least Squares Means
                    Coefficients for race Least Square Means
                                        race Level
Effect                                  Asian    Black    Hispanic    Other    White
Intercept                                   1        1           1        1        1
race       Asian                            1        0           0        0        0
race       Black                            0        1           0        0        0
race       Hispanic                         0        0           1        0        0
race       Other                            0        0           0        1        0
race       White                            0        0           0        0        1
sex        Female                         0.5      0.5         0.5      0.5      0.5
sex        Male                           0.5      0.5         0.5      0.5      0.5

                               Standard                    LSMEAN
race           y1 LSMEAN          Error    Pr > |t|       Number
Asian         62.6382507      3.5119521      <.0001            1
Black         77.4016564      3.0099679      <.0001            2
Hispanic      85.3485677      3.1771906      <.0001            3
Other         58.7332956      5.2036501      <.0001            4
White         61.6755517      1.5532976      <.0001            5
```

Least Squares Means for Effect race
t for H0: LSMean(i)=LSMean(j) / Pr > |t|
Dependent Variable: y1

i/j	1	2	3	4	5
1		-3.20959	-4.82644	0.623286	0.253505
		0.0014	<.0001	0.5332	0.7999
2	3.209586		-1.82924	3.112197	4.708515
	0.0014		0.0677	0.0019	<.0001
3	4.826444	1.829242		4.376613	6.79995
	<.0001	0.0677		<.0001	<.0001
4	-0.62329	-3.1122	-4.37661		-0.54379
	0.5332	0.0019	<.0001		0.5867
5	-0.25351	-4.70851	-6.79995	0.543792	
	0.7999	<.0001	<.0001	0.5867	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

Least Squares Means
Coefficients for sex Least Square Means

Effect		sex Level Female	Male
Intercept		1	1
race	Asian	0.2	0.2
race	Black	0.2	0.2
race	Hispanic	0.2	0.2
race	Other	0.2	0.2
race	White	0.2	0.2
sex	Female	1	0
sex	Male	0	1

sex	y1 LSMEAN	Standard Error	H0:LSMEAN=0 Pr > |t|	H0:LSMean1=LSMean2 t Value	Pr > |t|
Female	67.4395095	2.2041923	<.0001	-1.43	0.1535
Male	70.8794193	1.7677880	<.0001		

The second LSMEANS are *with* the OM option:

Least Squares Means
Coefficients for race Least Square Means

Effect		race Level Asian	Black	Hispanic	Other	White
Intercept		1	1	1	1	1
race	Asian	1	0	0	0	0
race	Black	0	1	0	0	0
race	Hispanic	0	0	1	0	0
race	Other	0	0	0	1	0
race	White	0	0	0	0	1
sex	Female	0.326	0.326	0.326	0.326	0.326
sex	Male	0.674	0.674	0.674	0.674	0.674

race	y1 LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Asian	63.2367950	3.4954641	<.0001	1
Black	78.0002007	2.9918561	<.0001	2
Hispanic	85.9471120	3.1501100	<.0001	3
Other	59.3318399	5.2019535	<.0001	4
White	62.2740960	1.4819289	<.0001	5

Least Squares Means for Effect race
t for H0: LSMean(i)=LSMean(j) / Pr > |t|
Dependent Variable: y1

i/j	1	2	3	4	5
1		-3.20959	-4.82644	0.623286	0.253505
		0.0014	<.0001	0.5332	0.7999
2	3.209586		-1.82924	3.112197	4.708515
	0.0014		0.0677	0.0019	<.0001
3	4.826444	1.829242		4.376613	6.79995
	<.0001	0.0677		<.0001	<.0001
4	-0.62329	-3.1122	-4.37661		-0.54379
	0.5332	0.0019	<.0001		0.5867
5	-0.25351	-4.70851	-6.79995	0.543792	
	0.7999	<.0001	<.0001	0.5867	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

Least Squares Means

Coefficients for sex Least Square Means

Effect		sex Level Female	Male
Intercept		1	1
race	Asian	0.104	0.104
race	Black	0.142	0.142
race	Hispanic	0.128	0.128
race	Other	0.047	0.047
race	White	0.579	0.579
sex	Female	1	0
sex	Male	0	1

sex	y1 LSMEAN	Standard Error	H0:LSMEAN=0 Pr > |t|	H0:LSMean1=LSMean2 t Value	Pr > |t|
Female	65.1806844	1.9762366	<.0001	-1.43	0.1535
Male	68.6205941	1.3735697	<.0001		

The difference between the two LSMEANS are that the first assumes an equal distribution across the categories of the variables, whereas the second uses the observed marginal distribution. In this context, when comparing males and females we can either assume the five levels of the RACE variable each have 20% of the population (the default) or we can use the actual distribution of the RACE values (the OM option). In this constructed example, we get the same answer for the difference 'female-male' but rather different values for the least squares means. In general, both the least squares means and their differences will change when OM is specified.

TWO FACTORS AND INTERACTION WITH NON-ESTIMABLE LSMEANS

Here is an example with a two-way interaction:

CODE FOR GLM:

```
proc glm data=anal;
class sex race;
model y1 = race sex race*sex / solution;
lsmeans race sex / stderr e;
lsmeans race sex / stderr e om;
estimate 'male'   intercept 1 sex 0 1 race .2 .2 .2 .2 .2  race*sex 0 0 0 0 0  .2 .2 .2 .2 .2;
estimate 'female' intercept 1 sex 1 0 race .2 .2 .2 .2 .2  race*sex .2 .2 .2 .2 .2  0 0 0 0 0;
estimate 'female-male' sex 1 -1 race*sex .2 .2 .2 .2 .2  -.2 -.2 -.2 -.2 -.2;
title3 'GLM by race sex race*sex';
run;
```

It turns out in this example, the OM version of the least squares means are nonestimable. This is because the observed proportion of males varies by race (or, to put it another way, the distribution across race is different for the two sexes). This makes the coefficients inconsistent. The ESTIMATE statements mimic the least squares means for the sex variable without the OM option:

sex	y1 LSMEAN	Standard Error	Pr > |t|
Female	68.1117839	2.5611683	<.0001
Male	70.5239982	1.9646266	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
male	70.5239982	1.96462660	35.90	<.0001
female	68.1117839	2.56116830	26.59	<.0001
female-male	-2.4122143	3.22790036	-0.75	0.4551

Any consistent distribution for the other variables can be used to calculate least squares means – you are not limited to the uniform distribution or the observed distribution. For example, we might want to use the distribution over the RACE variable from some reference population. In the example presented here, it might make sense to use the distribution used to generate the simulated data. Or, we might want to use the observed overall distribution of RACE (not separately by sex) instead. Here's how we might code the ESTIMATE statements and the associated output from GLM:

CODE FOR GLM:

```
proc glm data=anal;
class sex race;
model y1 = race sex race*sex / solution e;
estimate 'male (h)' intercept 1  sex 0 1  race  .10 .15 .15 .05 .55
          race*sex 0 0 0 0 0  .10 .15 .15 .05 .55;
estimate 'female (h)' intercept 1  sex 1 0 race  .10 .15 .15 .05 .55
          race*sex .10 .15 .15 .05 .55  0 0 0 0 0;
estimate 'female-male (h)' sex 1 -1 race*sex .10 .15 .15 .05 .55
          -.10 -.15 -.15 -.05 -.55;
estimate 'male (o)' intercept 1  sex 0 1  race  .104 .142 .128 .047 .579
          race*sex 0 0 0 0 0  .104 .142 .128 .047 .579;
estimate 'female (o)' intercept 1  sex 1 0 race  .104 .142 .128 .047 .579
          race*sex .104 .142 .128 .047 .579  0 0 0 0 0;
estimate 'female-male (o)' sex 1 -1 race*sex .104 .142 .128 .047 .579
          -.104 -.142 -.128 -.047 -.579;
estimate 'female-male (off)' sex 1 -1 race*sex .105 .142 .128 .047 .579
          -.105 -.142 -.128 -.047 -.579;
title3 'GLM by race sex race*sex with hypothetical and observed race weights';
run;
```

Notice the last ESTIMATE statement, with the difference marked "(off)" in the label. That ESTIMATE statement is nonestimable because the coefficients add up to 1.001 instead of 1.000 within each sex. It is easy to end up with values, even with many decimal places specified, that are off just a bit and which are therefore reported as nonestimable. One solution is to adjust the values slightly to make sure they add up to 1. Another solution is to use the DIVISOR option to ensure the values add up. See Pasta (2010) for details. Here are the results from GLM:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
male (h)	69.2219208	1.38231738	50.08	<.0001
female (h)	65.7862376	1.98334675	33.17	<.0001
female-male (h)	-3.4356832	2.41753297	-1.42	0.1556
male (o)	68.6049088	1.37605701	49.86	<.0001
female (o)	65.1115576	1.98344075	32.83	<.0001
female-male (o)	-3.4933513	2.41403606	-1.45	0.1482

Notice that we get somewhat different estimates of the female-male difference depending on how we weight the levels of RACE (and both differ from the default values in the previous output).

WHICH INTERACTIONS TO INCLUDE?

One of the key questions you need to answer is which interactions to include. Consider a situation with four discrete predictor variables (included on the CLASS statement), for example A B C D. To include all the possible interactions means 6 possible two-way interactions, 4 possible three-way interactions, and 1 four-way interaction. The total number of terms in the model is 16, including the constant. Each of the four variables can be included or not in each term and there are therefore 2**4 or 16 possible terms including the constant and the main effects. If some of the four factors have many levels, the number of degrees of freedom can get large very quickly.

One systematic approach is to start with all possible interactions and systematically remove higher-order interactions that are not statistically significant, but never removing an interaction if it is contained in a higher-order interaction that is retained. For example you might drop the A*B*C*D term and then the A*B*C and then the A*B*D terms. But if A*C*D and B*C*D are statistically significant, you would need to retain A*C, A*D, C*D, B*C, and B*D, regardless of statistical significance, according to this rule. You could drop A*B if it is not statistically significant, as this term is not part of either A*C*D or B*C*D.

This is a perfectly reasonable approach and leads to results that are reasonably straightforward to interpret. However, it tends to produce models that have “too many” interactions in them. A more parsimonious model may fit the data nearly as well and be much easier to understand. Whether the simpler model or more complex model is preferred depends on the context of the model and falls into the “art” part of model building.

There are some approaches that tend to result in fewer interactions. One is to retain interactions only if the effect is “material” as well as statistically significant. By setting a criterion that the F statistic needs to be at least 3 (or even 4), you can eliminate interactions that are “not large” even though they are statistically significant. This stricter criterion can produce a simpler model that fits nearly as well.

Another approach is to test nested groups of interactions. If the A*B*C interaction is significant, make sure that the simultaneous test of A*B*C, A*B, A*C, and B*C is also significant. In other words, include a three-way interaction only if it is significant and the test of it with all the next lower-order interactions is also significant. This idea can be extended to testing each two-way interaction simultaneously with the three-way interaction and requiring the test be statistically significant. If any are not significant (e.g. the combined test of A*B*C plus A*C is nonsignificant), then those two interactions would be dropped.

This sort of cavalier model-building with repeated significance tests should not, of course, be considered to be a formal test of the associated hypotheses. In addition to the issue of multiple comparisons, there are many aspects of this activity that capitalizes on chance and can lead to suboptimal models. However, it remains a potentially useful exploratory approach.

A related approach to model building with many potential interactions would be to use “all subsets regression” modeling to evaluate many models. This approach, previously accessed through PROC RSQUARE, is now available directly in PROC REG. It takes some effort to make sure all “contained” effects are present when higher-order interactions are present. Which naturally leads to the question, “Do I have to always include all ‘contained’ effects?”

WHAT IF I OMIT SOME OF THE TERMS CONTAINED WITHIN AN INTERACTION?

What happens if you include the interaction A*B, say, and omit the main effect A or B or both? Or what if you include A*B*C and A*B and the main effects A and B but don’t include A*C or B*C? You’ve omitted some of the “contained” interactions, which violates one of the rules stated in the previous section. Is the resulting model still valid?

Yes, the resulting model is valid. However, the interpretation of the parameter estimates changes and even the interpretation of the statistical tests for the various terms. Consider a model that includes a term for A and A*B but not for B. The test for A*B now includes the main effect of B and so it is no longer a

test of the A*B interaction only; it is a broader test of the effect of B. If you omit the A main effect as well and include only the A*B interaction, you are in essence "putting all the degrees of freedom in one basket." You are asking about the impact of A and B without dividing the effect between the variables and without separating main effects from interactions.

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.

EXAMPLE

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 look 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 in trouble.

INTERACTIONS AMONG CONTINUOUS VARIABLES

When multiple continuous variables are involved and you are concerned about interactions, there is a temptation simply to include cross-product terms such as X*Y, X*Z, and Y*Z. Inclusion of these cross-product terms certainly tests for interactions among the continuous variables, but it tests only a very specific form of interaction, unlike the situation with discrete variables. Furthermore, the statistical significance of cross-product terms is dependent on the form of the underlying variable, especially when some but not all cross-product terms are included.

Of perhaps even more practical importance is that the effect of interactions for continuous variables is sensitive to the linearity assumption. In general any ordinal variable can be profitably treated as approximately linear with equal spacing without much loss of statistical power. (See Pasta 2009 for additional discussion on this point.) However, the product of two ordinal variables is much more sensitive to the assumption of equal spacing between levels. Special attention needs to be paid if either continuous variable can take on zero values, as the cross-product term degenerates at that point.

INTERACTIONS AMONG CONFOUNDERS

Remember that a variable cannot be a confounder unless it is related both to the variable of interest and the outcome. Interactions between confounders must meet the same test to themselves be confounders – an interaction among confounders is not a difficulty if the interaction is not related to the variable of interest or is not related to the outcome. In practice, if there is a statistically significant interaction between confounders it is usually also a confounder. In general it does not hurt to include those interactions in the model – or indeed any possible confounders.

Deserving special attention are interactions with the variable of interest, whether the interaction is with a confounder or with another variable that is not a confounder (because, for example, it is not related to the variable of interest). For example, in a model where sex is unrelated to treatment (the variable of interest), it may be that the effect of the treatment is different for the different sexes. That does not make sex a confounder; it merely has an interaction with treatment. Suppose men have a greater response to a treatment than women. As long as there is no relationship between treatment and sex (e.g. because this is a randomized study), sex is not a confounder. But there is a sex-by-treatment interaction that is important and should be evaluated.

CONCLUSION

Confounding and interactions are important to consider when building statistical models. Confounding has the potential to lead to incorrect conclusions if not addressed through modeling or other means such as stratification or matching. The inclusion of interactions (whether of confounding variables or other variables) can lead to complicated models that are difficult to interpret. Sometimes it is worthwhile to simplify models by eliminating all but the largest interactions. After testing for the statistical significance of the marginal effect of interactions in models that include all lower-order interactions, it is sometimes useful to parameterize models without the lower-order interactions in order to get coefficients that are easier to interpret. Special care needs to be taken when evaluating interactions involving multiple continuous variables.

REFERENCES

Cochran, W. G., and Cox, G. M. (1957), Experimental Designs, 2nd Ed., John Wiley & Sons, New York

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

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 (2010). 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 Analysis
ICON Late Phase & Outcomes Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
+1.415.371.2111
david.pasta@iconplc.com
www.iconplc.com