Karnaugh Maps, Interaction Effects, and Creating Composite Dummy Variables for Regression Analysis in SAS® Software

Lawrence C. Marsh and Karin L. Wells
Department of Economics
University of Notre Dame
Notre Dame, Indiana 46556

ABSTRACT

Using multiple sets of dummy (binary or indicator) variables in a single regression may lead to substantial multicollinearity problems especially when providing for interaction effects. For example, agricultural output may be affected by type of soil, method of soil preparation, type of seed, type of fertilizer, method of weed and insect control, and other discrete factors that may involve significant interaction effects. This paper demonstrates how to use SAS®[1] programming to completely eliminate these multicollinearity problems while at the same time fully providing for all possible interaction effects among the characteristics represented by all of the original sets of dummy variables. In addition, we apply this approach to the dependent variable in a regression model explaining the joint decision of multiple decision makers.

The strategy is to replace all of the original sets of dummy variables with one new set of dummy variables that represents all characteristics simultaneously. This new single set of dummy variables may be referred to as a set of *composite* dummy variables. Karnaugh maps[2] are generated in SAS® to produce a single multidimensional contingency table representing all of the discrete sets of characteristics simultaneously. A new dummy variable is automatically created by SAS® for each cell in the multidimensional contingency table that contains one or more observations. The perfect multicollinearities among the original sets of dummy variables are found in the cells with zero observations.

INTRODUCTION

This paper is written for statistical consultants, teachers of statistics, and applied statisticians working with binary, indicator, or dummy explanatory variables in regression analysis, analysis of variance and covariance, and similar statistical methods. In particular, it is for those concerned with the limitations inherent in Venn diagrams and characteristic dummy variables,[3] especially in the presence of substantial interaction effects and multicollinearity problems.

As the number of multiple sets of characteristics increases, the Venn diagram quickly fails as a useful device. Instead of the Venn diagram, we suggest the use of Karnaugh maps, especially when interaction effects are widely evident in the data.

Moreover, standard characteristic dummy variables with interaction terms often result in extensive multicollinearity problems when several sets of such characteristics dummy variables are used in

---

[1] SAS® is a registered trademark of SAS Institute Inc., Cary, NC in the USA and other countries. The symbol ® indicates USA registration.

[2] See Karnaugh (1953), Rybak and Rybak (1973), and Rybak (1975) for definition and discussion of Karnaugh maps.

[3] Characteristic dummy variables by definition represent only a single concept or characteristic (e.g. one-way analysis of variance).

the same model. By replacing the traditional characteristic dummy variables with a single set of composite dummy variables,[4] the multicollinearity problem is completely eliminated, at least among all of the dummy variables, while fully taking into account all of the interaction effects.

VENN DIAGRAMS VERSUS KARNAUGH MAPS

The obvious problem with Venn diagrams, as usually represented, is that they are inadequate for keeping track of all possible interaction effects when there are several overlapping sets. Karnaugh maps serve as a simple device for keeping track of all of the possible interactions. The example below explains Karnaugh maps and how they can be extended to account for the interactions between multiple sets of characteristic dummy variables.

Start with a contingency table with four rows corresponding to four different religions (e.g. Buddhist, Christian, Jewish and Muslim) and two columns, one for those with a high school diploma and the other for those without a high school diploma.

Table 1.

	HS diploma	no HS diploma
Buddhist	18	23
Christian	21	20
Jewish	25	16
Muslim	20	22

This contingency table is in the form of a four by two matrix. The cells represent interaction between religion and high school diploma status.

[4] Composite dummy variables by definition represent multiple sets of characteristics simultaneously.

If we now wish to introduce three racial groups into the analysis, we can do so by copying the original four by two matrix twice and placing these two copies alongside the original four by two matrix to form a four by six matrix.

The original cell entries are then redistributed within this larger matrix to reflect the interactions (joint distribution) among these three sets of characteristic dummy variables (religion, high school diploma status, and race).

Table 2.

	Black		White		Hispanic	
	HS	nHS	HS	nHS	HS	nHS
B	5	7	10	9	3	7
C	5	6	11	6	5	8
J	3	9	18	7	4	0
M	12	14	0	6	8	2

In a similar manner the extended matrix may be duplicated as new sets of characteristic dummy variables are incorporated into the analysis. The new copies may be added to the side or bottom of the latest extended matrix in such a way as to keep the newly formed matrix as square as possible (for convenience). Thus, gender, occupation, industry of employment, nationality and other characteristics may be added.

Note, these Karnaugh maps serve as contingency tables for keeping track of relative frequencies (sample data) or discrete multivariate probability distributions (population or census data). Karnaugh maps are particularly useful for relative frequencies or ad hoc discrete probability distributions that cannot be adequately represented in compact mathematical formulas. In any case, Karnaugh maps form the basis for the creation of composite dummy variables discussed in the next section.

CHARACTERISTIC VERSUS COMPOSITE DUMMY VARIABLES

A single set of characteristic dummy variables represents a set of mutually exclusive and exhaustive outcomes. Perhaps the set of variables in a regression includes an intercept or constant term as a default category or characteristic. Generally as additional sets of characteristic dummy variables are added, the possibility of finding perfect linear combinations across the sets increases rapidly. This problem is well known to applied researchers. Often, in practice, important characteristics may be left out of the analysis altogether in order to avoid such perfect multicollinearity or SAS® may delete them from the model automatically as in PROC REG models.

Furthermore, in addition to being prone to producing perfect multicollinearities, using simple sets of characteristic dummy variables ignores possible interaction effects among these dummies. For example, the pay gap between men and women may be different for blacks than for whites. Including two separate sets of characteristic dummy variables ignores such a possibility and, in fact, forces the analysis to implicitly impose the restriction of an equal pay gap between men and women regardless of any confounding characteristic such as race. Of course, incorporating the necessary interaction terms, along with the two sets of dummies, to account for interaction effects just adds to the multicollinearity problem.

With all this in mind we introduce the concept of composite dummy variables. The idea here is to replace the original multiple sets of characteristic dummy variables with a single set of (composite) dummy variables. A single, mutually exclusive, set of composite dummy variables is created to correspond to the set of cells in the multivariate contingency table described above. A composite dummy variable is created for each cell that contains at least one observation. A composite dummy variable is not created for any cell with zero observations since these cells correspond to the perfect multicollinearities among the original sets of characteristic dummy variables. For example in Table 2 above, there are 24 cells but only 22 composite dummy variables are created. Thus, the composite dummy variables consist of a set of orthogonal, mutually exclusive and exhaustive variables. This one set represents all combinations of the original sets of characteristic dummy variables, as they exist in the data.

CREATING COMPOSITE DUMMY VARIABLES

If the original variables RACE, GENDER and OCCUPATN are coded with single digits and INDUSTRY is coded with a two digit code, then the following SAS® code can be used to create a single variable that has a different value for each of the different possible combinations of RACE, GENDER, INDUSTRY and OCCUPATN. Notice the multiplier 100 is not used in this example because INDUSTRY needs two spaces for its two digits.

```
01 DATA COMPOSIT; SET INDD.DEMO;
02 COMBO=10000*RACE+1000*GEN-
     DER+10*INDUSTRY+OCCUPATN;
```

To determine the number of composite dummy variables that must be created in this case, the newly created variable COMBO must now be reduced to a unique set of values. The following code not only reduces COMBO to a unique set, but also keeps track of how many duplicates, M, there are in the originally created COMBO variable.

After the set of composite dummy variables is created, the variable M is used to create an indicator variable IN. The variable IN is needed to reintegrate the newly created dummy variables back into the original data set.

```
03 PROC SORT;  BY COMBO;
04 DATA REDUCED;
05 SET COMPOSIT; BY COMBO; M+1;
06 IF LAST.COMBO THEN OUTPUT;
07 IF LAST.COMBO THEN M=0;
```

In the data set REDUCED, the variable M now represents the number of observations in each of the cells, with at least one observation, of the Karnaugh map's multidimensional contingency table.

For each cell with more than zero observations a unique dummy variable representing that cell can be created. On the other hand, if there are a large number of cells and possibly more dummy variables being created than the investigator really wants, then a variable LOWER can be used to control the number of dummy variables created. The value of LOWER represents the number of observations that must be exceeded in a cell in order for a dummy variable to be created for that cell. This is useful for goodness-of-fit tests where a minimum of five observations per cell is generally required in order to produce a test statistic that reasonably approximates a variable with a chi-square distribution.

Now, when the variable IN equals one, the creation of a dummy variable is authorized for the observations in that cell. When the variable IN equals zero, no dummy variable is created to represent the observations in that cell.

```
08 DATA EXPANDED; SET REDUCED;
09 LOWER=0;
10 DO I = 1 TO M;
11  IF M GT LOWER THEN IN = 1;
                  ELSE IN = 0;
12 IF COMBO = .  THEN IN = 0;
13 OUTPUT;
14 END;
```

If LOWER is greater than zero, then separate data sets are created for those observations in cells which are not represented in the new set of composite dummy variables (HOLD) and those observations in cells which are represented by composite dummy variables (CREATE).

```
15 DATA  HOLD;
      SET EXPANDED;  IF IN = 0;
16 DATA  CREATE;
      SET EXPANDED;  IF IN = 1;
```

In order to produce the composite dummy variables for all observations in the data set CREATE, we need to first find out how many unique values of COMBO there are in the CREATE data set. This task could be done more easily if it were not for the fact that we also need to produce a copy of that unique set of values which are stored in the new data set UNIQUE.

```
17 DATA  UNIQUE;
      SET CREATE;  BY COMBO;
18 IF LAST.COMBO THEN OUTPUT;
19 KEEP COMBO;
```

The key in SAS® to easily creating the composite set of dummy variables is PROC TRANSPOSE[5]. For each of the unique set of values of the variable COMBO, PROC TRANSPOSE produces a

[5] For an earlier example of using PROC TRANSPOSE to create dummy variables for searching for the number and location of spline knots see Marsh (1983). For a more recent example of automatic dummy variable creation using PROC TRANSPOSE see Marsh and Wells (1994).

variable with the prefix D followed by integers from one to as many dummy variables as are needed to represent all of the different possible values of COMBO in the data set.

In this situation PROC TRANSPOSE creates just one observation with as many D variables as are needed. These variables are pulled into the new data set DUM. Note, one additional variable is added, DEND, which is set to a value of one. Actually DEND could be set to any value since the only reason for creating it is to be able to refer to the sequence D1--DEND without having to know how many variables have been created. In other words, we can refer to this set of variables without actually knowing how many of them there are. This is convenient especially if you change either the variable of interest or the observations associated with that variable in such a way as to alter the number of values of COMBO .

```
20 PROC TRANSPOSE DATA=UNIQUE
                  PREFIX=D;
21 DATA DUM; SET; DEND=1;
```

If you print out the variables D1--DEND at this point you will see that they are not yet dummy variables. In fact, all of them except DEND have the corresponding value of COMBO. The value of COMBO will be replaced shortly with single digit (0 or 1) dummy variable values. Next, we need to "pull the window shade down" by duplicating the one observation in DUM as many times as needed to match the number of observations in the data set EXPANDED.

```
22 DATA MATCH;
     IF _N_=1 THEN SET DUM;
     SET EXPANDED;
```

Each cell in the Karnaugh map has its own unique value of COMBO and is now going to be assigned its own integer Z, which has a value of one for all observations in the first cell, two for all observations in the second cell, et cetera.

For each cell the do loop DO OVER D; assigns a value of 0 or 1 to each of the D1--DEND variables which have their own integer Y which is one for D1, two for D2, et cetera.

```
23 PROC SORT DATA=MATCH
     OUT=MATCH; BY COMBO;
24 DATA DUMMIES;
     SET MATCH; BY COMBO;
25 IF FIRST.COMBO THEN Z+1;
26 ARRAY D D1--DEND;
27 DO OVER D; Y+1; D=0;
28  IF Y=Z THEN D=1; END; Y=0;
```

Thus, for each cell only the dummy variable corresponding to that cell is assigned a value of one while the other dummy variables are all assigned a value of zero. Once this is accomplished for a particular cell then the counter Y which keeps track of which dummy variable is being operated on is reset to zero.

Now the cells for which dummy variables have been created can be recombined with the cells which did not have enough observations to warrant their own dummy variables (if LOWER is set higher than zero).

```
29 DATA RESTORE;
     SET DUMMIES HOLD;
30 ARRAY DUM D1--DEND;
31 DO OVER DUM;
32 IF DUM=. THEN DUM=0;
33 END;
```

Finally, as a check, it is useful to compare the frequency counts on the values of COMBO with those of the corresponding dummy variables to make sure they are the same.

```
33 PROC FREQ;
       TABLES COMBO D1--DEND;
```

DISCRETE CHOICE DEPENDENT VARIABLES

Up until now we have considered only independent variables as dummy variables. Now we look at the issue of dependent dummy variables and, in particular, discrete choice variables where several decision makers produce a joint decision. A few examples include the Supreme Court, a Board of Directors, a married couple, and a trial jury.

The problem is one of determining how individual decision makers contribute to the final group decision. One approach is to treat the joint decision as simply the product of the individual decisions. This is just an application of the characteristic dummy variable approach and assumes no interaction effects. It implies that the decision makers do not influence one another in any way and, therefore, make independent decisions.

Alternatively, we propose the composite dummy variable approach to this problem. This allows for interaction effects and, therefore, interaction among the decision makers in influencing one another in either a positive or negative sense. A positive correlation between the decisions of two decision makers implies that they tend to influence one another in a positive way, whereas, a negative correlation implies a tendency to disagree with one another.

We start with a binomial setup which extends to a Poisson process when the binomial probability becomes a function of some explanatory variables. Assume that a "no" decision means that random variable $y_1 = 0$ with probability P_1 and that a "yes" decision means that $y_1 = 1$ with probability P_2 such that $P_1 + P_2 = 1$. This single decision maker setup is portrayed in Table 3.

For two decision makers we introduce a second binary random variable y_2 and redefine the probabilities as implied by Table 4. Note that one copy of Table 3 has been placed along side the original to produce Table 4.

Table 3.

first no $y_1 = 0$	first yes $y_1 = 1$
P_1 0	P_2 1

Table 4.

first no $y_1 = 0$	first yes $y_1 = 1$	first no $y_1 = 0$	first yes $y_1 = 1$
P_1 0 0	P_2 1 0	P_3 0 1	P_4 1 1
$y_2 = 0$ second no		$y_2 = 1$ second yes	

To introduce a third decision maker, we simply make one copy of Table 4 and place it along side of the original to generate Table 5.

Table 5.

first no $y_1 = 0$	first yes $y_1 = 1$	first no $y_1 = 0$	first yes $y_1 = 1$	first no $y_1 = 0$	first yes $y_1 = 1$	first no $y_1 = 0$	first yes $y_1 = 1$
P_1 0 0 0	P_2 1 0 0	P_3 0 1 0	P_4 1 1 0	P_5 0 0 1	P_6 1 0 1	P_7 0 1 1	P_8 1 1 1
$y_2 = 0$ second no		$y_2 = 1$ second yes		$y_2 = 0$ second no		$y_2 = 1$ second yes	
$y_3 = 0$ third no				$y_3 = 1$ third yes			

Table 6.

1 no y_1=0	1 yes y_1=1	1 no y_1=0	1 yes y_1=1	1 no y_1=0	1 yes y_1=1	1 no y_1=0	1 yes y_1=1	1 no y_1=0	1 yes y_1=1	1 no y_1=0	1 yes y_1=1	1 no y_1=0	1 yes y_1=1	1 no y_1=0	1 yes y_1=1
P_1 0000	P_2 1000	P_3 0100	P_4 1100	P_5 0010	P_6 1010	P_7 0110	P_8 1110	P_9 0001	P_{10} 1001	P_{11} 0101	P_{12} 1101	P_{13} 0011	P_{14} 1011	P_{15} 0111	P_{16} 1111
$y_2 = 0$ second no		$y_2 = 1$ second yes		$y_2 = 0$ second no		$y_2 = 1$ second yes		$y_2 = 0$ second no		$y_2 = 1$ second yes		$y_2 = 0$ second no		$y_2 = 1$ second yes	
third no $y_3 = 0$				third yes $y_3 = 1$				third no $y_3 = 0$				third yes $y_3 = 1$			
fourth no $y_4 = 0$								fourth yes $y_4 = 1$							

In a similar manner, the corresponding joint probabilities for four decision makers are given in Table 6.

Returning to Table 3, representing the single decision maker, it is clear that there are many alternative ways of providing an appropriate and useful structure for the probabilities, P_1 and P_2. We suggest the following approach which can easily be expanded to accommodate many decision makers and to allow for choice among many alternative, mutually exclusive and exhaustive choices. In particular, for decisions made by a single decision maker, we define P_1 and P_2 as follows:

$$P_1 = \frac{1}{1+e^{X_i\beta_{12}}} \quad \text{and} \quad P_2 = \frac{e^{X_i\beta_{12}}}{1+e^{X_i\beta_{12}}}$$

With only a single decision maker, these are the marginal probabilities, since there are no

joint or conditional distributions in the single decision maker case.

In the case of two decision makers, the probabilities get more interesting. For this case, we can transform Table 4 into the equivalent, but more convenient and recognizable, Table 7 as follows:

Table 7.

joint (y_1, y_2)	lst no $y_1 = 0$	lst yes $y_1 = 1$	marginal for y_2
2nd no $y_2 = 0$	P_1 0 0	P_2 1 0	$y_2 = 0$ $P_1 + P_2$
2nd yes $y_2 = 1$	P_3 0 1	P_4 1 1	$y_2 = 1$ $P_3 + P_4$
marginal for y_1	$y_1 = 0$ $P_1 + P_3$	$y_1 = 1$ $P_2 + P_4$	

The corresponding probabilities are given as:

$$P_1 = \frac{1}{1 + e^{X_1'\beta_{12}} + e^{X_2'\beta_{23}} + e^{X_1'\beta_{14} + X_2'\beta_{24}}}$$

$$P_2 = \frac{e^{X_1'\beta_{12}}}{1 + e^{X_1'\beta_{12}} + e^{X_2'\beta_{23}} + e^{X_1'\beta_{14} + X_2'\beta_{24}}}$$

$$P_3 = \frac{e^{X_2'\beta_{23}}}{1 + e^{X_1'\beta_{12}} + e^{X_2'\beta_{23}} + e^{X_1'\beta_{14} + X_2'\beta_{24}}}$$

$$P_4 = \frac{e^{X_1'\beta_{14} + X_2'\beta_{24}}}{1 + e^{X_1'\beta_{12}} + e^{X_2'\beta_{23}} + e^{X_1'\beta_{14} + X_2'\beta_{24}}}$$

In general, when there are G decision makers, the number of probabilities, K, is:

$$K = 1 + \sum_{g=1}^{G} 2^{(g-1)}$$

For example, for a single decision maker G=1 and $K = 1 + 2^0 = 1 + 1 = 2$ probabilities, which are shown as P_1 and P_2 in Table 3. For two decision makers G=2 and $K = 1 + 2^0 + 2^1 = 4$ as in Table 4. More generally, for G decision makers, a set of K joint probabilities are generated for the ordered cells $k = 1, \ldots, K$ as follows[6]:

$$P_k = \frac{e^{\sum_{g=1}^{G} X_g'\beta_{gk} y_{gk}}}{\sum_{k=1}^{K} e^{\sum_{g=1}^{G} X_g'\beta_{gk} y_{gk}}}$$

where $y_{gk} = 0$ if the g^{th} decision maker in the k^{th} cell says "no", while $y_{gk} = 1$ if the g^{th} decision maker in the k^{th} cell says

[6] Readers may note some similarity between these probability expressions and the sigmoidal phantom units used in hidden layers in typical stochastic neural network systems as in Brummett and Marsh (1993).

"yes". Furthermore, X_g^t is a row vector of explanatory variable values specific to the g^{th} decision maker, and β_{gk} is the vector of coefficients for the gth decision maker's explanatory variables in contributing to the k^{th} joint probability.

The correlation between the decisions of decision makers i and j is defined as usual:

$$\rho_{y_i,y_j} = \frac{\mathbf{Cov}(y_i,y_j)}{\sqrt{\mathbf{Var}(y_i)}\sqrt{\mathbf{Var}(y_j)}}$$

In the case of the two decision makers depicted in Table 7, we have:

$$\mathbf{Cov}(y_1,y_2) = P_4 - (P_2+P_4)(P_3+P_4)$$

$$\mathbf{Var}(y_1) = (P_2+P_4) - (P_2+P_4)^2$$

$$\mathbf{Var}(y_2) = (P_3+P_4) - (P_3+P_4)^2$$

and, therefore, the correlation between the decisions of decision makers one and two is:

$$\rho_{y_1,y_2} = \frac{P_4 - (P_2+P_4)(P_3+P_4)}{\sqrt{(P_2+P_4)-(P_2+P_4)^2}\sqrt{(P_3+P_4)-(P_3+P_4)^2}}$$

MAXIMUM LIKELIHOOD ESTIMATES

When the decision of each of the two decision makers is revealed separately, the likelihood function is as follows:

$$L = \prod_{i=1}^{n} P_1^{(1-y_{i1})(1-y_{i2})} P_2^{y_{i1}(1-y_{i2})} P_3^{(1-y_{i1})y_{i2}} P_4^{y_{i1}y_{i2}}$$

In the case of partial observability when only the joint decision is revealed, it is still possible to infer the individual decisions using the following likelihood function:

$$L = \prod_{i=1}^{n} (P_1+P_2+P_3)^{(1-y_{i12})} P_4^{y_{i12}}$$

where $y_{i12} = 1$ if the joint decision is "yes" and $y_{i12} = 0$ if the joint decision is "no".

Maximizing the likelihood function is achieved with any of the usual algorithms such as Newton-Raphson[7] or Method of Scoring to estimate the above model of joint decision making either in the full observability case or the partial observability case depending on available data.

SUMMARY AND CONCLUSIONS

The objective of this paper has been to demonstrate the usefulness of the composite dummy variable approach for both independent and dependent variables in regression analysis. The key to this approach is to recognize the importance of separately identifying each cell in the appropriate Karnaugh map. By representing each of the cells that contain at least one observation, we are able to take into account interaction effects and, therefore, correlations in the context of both independent and dependent variables in regression analysis. This approach, in some sense, simplifies the analysis while at the same time it deals effectively with such problems as multicollinearity and partial observability.

[7] See Brunson and Marsh (1991) or Marsh, Maudgal and Raman (1994) for a discussion of and application of Newton-Raphson and other nonlinear regression algorithms in SAS® Software.

REFERENCES AND RELATED PAPERS

Brummett, Daric L. and Lawrence C. Marsh, "Estimating Neural Network Flexible Form Production Functions," *Proceedings of SAS® Users Group International*, vol. 18, 1993, pages 344-350.

Brunson, Kevin D. and Lawrence C. Marsh, "Using SAS/IML® Software in a New Approach to Dealing with Multicollinearity," *Proceedings of SAS® Users Group International*, vol. 16, 1991, pages 637-641.

Karnaugh, M., "The Map Method of Synthesis of Combinational Logic Circuits," *American Institute of Electrical Engineers, Transactions, part I, Communications and Electronics,* vol. 72, November 1953.

Marsh, Lawrence C., "On Estimating Spline Regressions," *Proceedings of SAS® Users Group International*, vol. 8, 1983, pages 723-728.

Marsh, Lawrence C., Manjula Maudgal, and Jaishankar Raman, "Alternative Methods of Estimating Piecewise Linear and Higher Order Regression Models Using SAS® Software," *Proceedings of SAS® Users Group International*, vol. 15, 1990, pages 523-527.

Marsh, Lawrence C. and Karin L. Wells, "Transforming Data, Restructuring Data Sets, Creating Look-Up Tables, and Forming Person-Year Records for Event History Analysis," *Proceedings of the Midwest SAS® Users Group,* vol. 5, 1994, pages 260-266.

Rybak, Janet, "Diagrams for Set Theory and Probability Problems of Four or More Variables," *The American Statistician,* Vol. 29, no. 2, May 1975.

Rybak, John, and Rybak, Janet, *Map Logic and Other Extensions of Traditional Logic,* December 1973.