Modeling and Interpreting Interactions in Multiple Regression

Donald F. Burrill
The Ontario Institute for Studies in Education
Toronto, Ontario Canada

A method of constructing interactions in multiple regression models is described which produces interaction variables that are uncorrelated with their component variables and with any lower-order interaction variables. The method is, in essence, a partial Gram-Schmidt orthogonalization that makes use of standard regression procedures, requiring neither special programming nor the use of special-purpose programs before proceeding with the analysis. Advantages of the method include clarity of tests of regression coefficients, and efficiency of winnowing out uninformative predictors (in the form of interactions) in reducing a full model to a satisfactory reduced model. The method is illustrated by applying it to a convenient data set.

PRELIMINARIES

In a linear model representing the variation in a dependent variable Y as a linear function of several explanatory variables, interaction between two explanatory variables X and W can be represented by their product: that is, by the variable created by multiplying them together. Algebraically such a model is represented by Equation [1]:

$Y = a + b1X + b2\ W + b3\ XW + e$. [1]

When X and W are category systems, Eq. [1] describes a two-way analysis of variance (AOV) model; when X and W are (quasi-)continuous variables, Eq. [1] describes a multiple linear regression (MLR) model.

In AOV contexts, the existence of an interaction can be described as a difference between differences: the difference in means between two levels of X at one value of W is not the same as the difference in the corresponding means at another value of W, and this not-the-same-ness constitutes the interaction between X and W; it is quantified by the value of b3.

In MLR contexts, an interaction implies a change in the slope (of the regression of Y on X) from one value of W to another value of W (or, equivalently, a change in the slope of the regression of Y on W for different values of X): in a two-predictor regression with interaction, the response surface is not a plane but a twisted surface (like "a bent cookie tin", in Darlington's (1990) phrase). The change of slope is quantified by the value of b 3.

INTRODUCTION

In attempting to fit a model (like Eq. [1]) to a set of data, we may proceed in either of two basic ways:

1. Start with a model that contains all available candidates as predictors, then simplify the model by discarding candidates that do not contribute to explaining the variability in the dependent variable; or
2. Start with a simple model and elaborate on it by adding additional candidates.

In either case we will wish (at any stage in the analysis) to compare a "full model" to a "reduced model", following the usage introduced by Bottenberg & Ward, 1963 (or an "augmented model" to a "compact model", in Judd & McClelland's (1989) usage). If the difference in variance explained is negligible, we will prefer the reduced model and may consider simplifying it further. If the difference is large enough to be interesting, we suspect the reduced model to be oversimplified and will prefer the full model; we may then wish to consider an intermediate model, or a model even more elaborate than the present full model.

In our context, the "full model" will initially contain as predictors all the original variables of interest and all possible interactions among them.

Traditionally, all possible interactions are routinely represented in AOV designs (one may of course hope that many of them do not exist!), and in computer programs designed to produce AOV output; while interactions of any kind are routinely not represented in MLR designs, and in general have to be explicitly constructed (or at least explicitly represented) in computer programs designed to produce multiple regression analyses. This may be due in part to the fact that values of the explanatory variables (commonly called "factors") in AOV are constrained to a small number of nicely spaced values, so that (for balanced AOV designs) the factors themselves are mutually orthogonal, and their products (interaction effects) are orthogonal to them.

Explanatory variables (commonly called "predictors") in MLR, on the other hand, are usually not much constrained, and are seldom orthogonal to each other, let alone to their products. One consequence of this is that product variables (like XW) tend to be correlated rather strongly with the simple variables that define them: Darlington (1990, Sec. 13.5.6) points out that the products and squares of raw predictors in a multiple regression analysis are often highly correlated with each other, and with the original predictors (also called "linear effects"). This is seldom a difficult problem with simple models like Eq. [1], but as the number of raw predictors increases the potential number of product variables (to represent three-way interactions like VWX, four-way interactions like UVWX, and so on) increases exponentially; and the intercorrelations of raw product variables with other variables tend to increase as the number of simple variables in the product increases.

As a result, more complex models tend to exhibit multicollinearity, even though the idea of an interaction is logically independent of the simple variables (and lower-order interactions) to which it is related. This phenomenon may reasonably be called spurious multicollinearity . The point of this paper is that spurious multicollinearity can be made to vanish, permitting the investigator to detect interaction effects (if they exist) uncontaminated by such artifacts.

These high intercorrelations lead to several difficulties:

1. The set of predictors and all their implied interactions (in a "full model") may explain an impressive amount of the variance of the dependent variable Y, while none of the regression coefficients are significantly different from zero.
2. The regression solution may be unstable, due to extremely low tolerances (or extremely high variance inflation factors (VIFs)) for some or all of the predictors.
3. As a corollary of (2.), the computing package used may refuse to fit the full model.

An example illustrating all of these characteristics is displayed in Exhibit 1.

EXHIBIT 1

In this example four raw variables (P1, G, K, S) and their interactions (calculated as the raw products of the corresponding variables) are used to predict the dependent variable (P2). P1 and P2 are continuous variables (pulse rates before and after a treatment); G, K, and S are dichotomies coded [1,2]: G indicates treatment (1 = experimental, 2 = control); K indicates smoking habits (1 = smoker, 2 = non-smoker); S indicates sex (1 = male, 2 = female).

These computations were carried out in Minitab. (Similar results occur in other statistical computing packages.) The first output from the regression command (calling for 15 predictors) was * P1.G.S.K is highly correlated with other X variables * P1.G.S.K has been removed from the equation followed by * NOTE * P1 is highly correlated with other predictor variables and a similar message for each of the other predictors remaining in the equation. The values of the regression coefficients, their standard errors, t-ratios, p-values, and variance inflation factors (VIF) are displayed in the table below, followed by the analysis of variance table.

```
              Standard
Predictor   Coefficient       error        t        p       VIF

Constant         131.7        111.7        1.18    0.242

P1               -1.345        1.537     -0.87    0.385     440.6
G               -38.51        51.12      -0.75    0.454     957.6
K               -79.86        65.49      -1.22    0.226    1412.1
S                19.63        63.00       0.31    0.756    1454.4
G.S             -26.29        38.88      -0.68    0.501    2906.0
G.K              21.96        24.36       0.90    0.370    1230.8
S.K              22.49        31.64       0.71    0.479    1953.4
G.S.K             1.542        9.798      0.16    0.875     842.6
P1.G              0.8671       0.6807     1.27    0.207    1101.9
P1.K              1.2570       0.9498     1.32    0.190    1845.6
P1.S              0.3258       0.7536     0.43    0.667    1673.4
P1.G.S            0.0113       0.3787     0.03    0.976    1784.3
P1.G.K           -0.4236       0.3788    -1.12    0.267    1663.1
P1.S.K           -0.2912       0.4085    -0.71    0.478    2078.3


         Source     DF      SS        MS       F       p

       Regression    14    22034.0   1573.9   26.60   0.000
       Error         77     4556.0     59.2
       Total         91    26590.0

                          2                 2
         s = 7.692       R  = 82.9%       R  (adj) = 79.8%/
```

ORTHOGONALIZED PREDICTORS

These difficulties can be avoided entirely by orthogonalizing the product and power terms with respect to the linear effects from which they are constructed. This point is discussed in some detail (with respect to predictors in general) in Chapter 5 of Draper and Smith (1966, 1981), and the Gram-Schmidt orthogonalizing procedure is described in their Sec. 5.7. Because that discussion is couched in matrix algebra, it is largely inaccessible to anyone who lacks a strong mathematical background. Also, they write in terms of orthogonalizing the whole X matrix; but in fact a partial orthogonalization will often suffice.

In presenting the Gram-Schmidt procedure Draper and Smith (ibid.) observe that the predictors can be ordered in importance, as least in principle -- that is, the investigator may be interested first in the effect attributable to X1 , then to the additional variance that can be explained by X2 , then to whatever increment is due to X3, and so on. For the example with which they illustrate the procedure (generating orthogonal polynomials), this assumption is reasonable.

However, the investigator may not always have (or be willing to impose) a strict a priori ordering on all the predictors. Suppose that we have four predictors U, V, W, X, which are moderately intercorrelated; and that we are interested in a model that includes all the two-way interactions between them, all the three-way interactions, and the four-way interaction. Now a natural ordering begins to emerge, but only a partial one: we will wish to see what effects are attributable to the linear terms alone, then what additional effects are due to the two-way interactions terms, then the three-way terms, and so on. In general, we are unlikely to be interested in retaining (e.g.) two-way interactions in the final model unless they provide an improvement over a model containing the original variables alone.

A sequence of non-orthogonalized models.

One way of proceeding in such a case is to fit several models in a hierarchical sequence of formal models, using as predictors:

1. The original variables only.
2. The original variables and the two-way interactions.
3. The original variables and the two- and three-way interactions.
4. The original variables and all interactions.

Then make the usual comparison between models (change in sum of squares divided by change in degrees of freedom, and that ratio divided by the residual mean square for the more elaborate model, for an F-test of the hypothesis that the additional terms did not explain any more of the variance in the dependent variable).

One drawback to proceeding thus is that not all statistical packages will perform this F-test automatically, leaving the investigator to work it out on his own. Another drawback is that, if the three-way interaction terms (for example) do add significantly to the variance explained, it is then necessary to remove them (or add them, depending on the starting model used) one at a time in successive models, to find out precisely which of them is (or are) producing the effect.

This procedure, which is recommended by many authors (e.g., Aiken & West 1991), requires a series of regression analyses. If, as may well be expected, the interactions are strongly correlated with the linear effects (the original variables) or with each other, there still may be some lurking ambiguity in interpreting the regression coefficients.

A single orthogonalized model.

However, if all of the interactions have been orthogonalized with respect to the lower-order terms, one need only fit the full model. Then the standard t-tests of the regression coefficients will indicate directly which predictors (original variables and interactions) contribute significantly to explaining variance in Y, and which do not, and which (if any) are borderline cases for which some further investigation may be useful.

By "orthogonalized with respect to the lower-order terms" I mean that each interaction variable (originally the raw product of the corresponding original variables) is represented by the residual part of that product, after the original variables and any lower-order interaction variables have been partialed out of it. Consequently every such variable correlates zero with all the lower-order variables, and may be thought of as a "pure interaction" effect at its own level.

A procedure for orthogonalizing interactions.

We may proceed as follows:

1. Two-way interactions (UV, UW, ..., WX): Form the simple product (U.V = U*V, e.g.), regress it on the two original variables (fit the regression model U.V = a + b1 U + b2 V + residual), and save the residual as a new variable UV. Use UV to represent the interaction between U and V; and proceed similarly for each pair of original variables. (Notice that UV has mean zero, and correlates zero with both U and V, because it is a residual.)
2. Three-way interactions (UVW, UVX, ..., VWX):
after the two-way interactions have been constructed, form either the simple product (e.g., U.V.W="U*V*W)" of the three original variables, or multiply the interaction term for two of them by the third (e.g., U.V.W="UV*W," or U.V.W="U*VW," etc.). Regress this product on the three original variables and the three two-way interactions: that is, fit the model U.V.W="a" + b1 U + b 2V + b3 W + b4 UV + b5 UW + b 6VW + residual, and save the residual as a new variable UVW. Use this to represent the three-way interaction between U, V, and W; and proceed similarly for all other three-way interactions.
3. Four-way interactions (UVWX):
After the two- and three-way interactions have been constructed, form a suitable four-way product (e.g., U.V.W.X="U*V*W*X" or U.V.W.X="[UV]*[WX]" or U.V.W.X="U*[VWX]" or ...). Regress this product on 14 variables the four original variables, the six two-way interactions, and the four three-way interactions and use the residual UVWX from this regression to represent the four-way interaction between U, V, W, and X.
4. Higher-way interactions:

   If higher-way interactions are to be modeled, the extension of this procedure is straightforward. For five-way interactions the initial product is regressed on 30 variables (five original, ten two-way, ten three-way, and five four-way). For six-way interactions 62 variables are involved (six original, 15 two-way, 20 three-way, 15 four-way, and six five-way).

5. Curvilinear terms:

   If curvilinear functions of one or more of the original variables are to be modeled, these too can be orthogonalized. For quadratic and cubic terms in X, for example, we would construct X(2), regress it on X(3), and use the residual XSQ to represent the quadratic component of X; then construct X , regress it on X and XSQ, and use its residual XCUB to represent the cubic component. Interactions between these components and any other original predictors (U, V, W) would then be constructed and orthogonalized in the same general manner.

Now a single regression analysis of the full model (the original variables and all their interactions) will produce unambiguous results, in the sense that any interaction that explains a significant amount of the variance of the dependent variable Y will have a significant regression coefficient, and any interaction whose regression coefficient is not significant can be discarded in arriving at a suitable reduced model. (For the relevant output from Minitab for the example presented in Exhibit 1, see Exhibit 2, Method D.)

Orthogonalizing may produce some useful side effects, as well. As Darlington (1990, Sec. 13.3.2; and cited in Aiken & West, 1991) also points out, it is sometimes possible for interaction and curvilinear effects to be confused for each other. But if all lower-order predictors have been partialed out of the interaction and curvilinear terms, it becomes possible to tell whether the distribution of the data permits them to be distinguished from each other, and if so whether one or the other, or both, belong in the model.

A COMPARISON OF METHODS

We continue by illustrating the results of multiple regression analyses carried out by four parallel procedures. The data set used for illustration is the PULSE data supplied with Minitab (the data set used in Exhibit 1):

"Students in an introductory statistics course participated in a simple experiment. The students took their own pulse rate .... They then were asked to flip a coin. If their coin came up heads, they were to run in place for one minute. Then everyone took their own pulse again. The pulse rates and some other data [were recorded.] (Ryan, Joiner, & Ryan, 1985, 318-319)"

The variables used in these analyses are P1 (first pulse rate), P2 (second pulse rate; the dependent variable), G (group: 1 = ran in place, 2 = did not run in place), K (smoker: 1 = smokes regularly, 2 = does not smoke regularly), S (sex: 1 = male, 2 = female). The sample size is 92.

Defining the four methods.

- A. In Method A, the four original variables are used in their raw form, and the interactions are constructed by multiplying together the relevant original variables.
- B. The three dichotomies are recoded from [1,2] to [0,1]. (For G and K, 0 = those who did not; for S, 0 = male.) The fourth (continuous) variable is left in its original form. Interaction variables are constructed by multiplying together the relevant variables.
- C. In this method, all four variables are centered: represented as deviations from their own (sample) means. Interaction variables are constructed by multiplying together the relevant variables.
- D. Here the four variables are used in their raw form. Interaction variables are initially constructed by multiplying together the relevant variables, and are then orthogonalized as described above.

In each analysis, the fifteen predictors were specified in the same (hierarchical) order: linear terms, then two-way interactions, three-way interactions only after the relevant two-way interactions, and the four-way interaction last.

EXHIBIT 2

Values of the regression coefficients, their standard errors, t-ratios, p-values, variance inflation factors (VIF), and sequential sums of squares (SEQ SS) for a full model (15 predictors), for each of four different methods for constructing interaction variables.

The analysis of variance, displayed at the end of the table, is identical for all four methods.

```
   Method A   (Raw variables and products)

                        Standard
Predictor  Coefficient    error      t       p       VIF      SEQ SS

Constant    149.9        228.5       0.66   0.514

P1           -1.577        2.976    -0.53   0.598    1629.6   10096.1
G           -51.6        152.3      -0.34   0.735    8389.8    7908.0
K           -91.0        138.3      -0.66   0.513    6219.3     116.7
S             5.3        168.6       0.03   0.975   10278.6    1087.0
G.S         -15.5        124.3      -0.12   0.901   29317.3    2129.0
G.K          29.74        88.41      0.34   0.738   16003.6     295.6
S.K          30.99        98.14      0.32   0.753   18549.5      62.2
G.S.K        -4.67        68.55     -0.07   0.946   40712.3      11.0
P1.G          1.032        1.926     0.54   0.594    8712.7     122.4
P1.K          1.402        1.846     0.76   0.450    6878.1      51.9
P1.S          0.504        2.085     0.24   0.810   12649.2      61.8
P1.G.S       -0.121        1.494    -0.08   0.936   27402.2      12.6
P1.G.K       -0.523        1.150    -0.45   0.651   15125.2      49.6
P1.S.K       -0.399        1.242    -0.32   0.749   18958.3      30.1
P1.G.S.K      0.0772       0.8429    0.09   0.927   35106.2       0.5
```

```
             Method B   ([0,1] dichotomies and products)

                         Standard
 Predictor  Coefficient    error      t       p      VIF     SEQ SS

 Constant       1.25       13.14      0.10   0.925

 P1             0.9716      0.1839    5.28   0.000    6.2   10096.1
 G             16.99       22.91      0.74   0.461  189.9    7908.0
 K              9.88       18.13      0.54   0.587  106.8     116.7
 S             17.65       19.06      0.93   0.357  131.5    1087.0
 G.S           24.83       33.40      0.74   0.460  180.2    2129.0
 G.K           25.06       32.23      0.78   0.439  180.8     295.6
 S.K          -21.65       55.27     -0.39   0.696  372.3      62.2
 G.S.K         -4.67       68.55     -0.07   0.946  300.0      11.0
 P1.G          -0.0196      0.3276   -0.06   0.952  218.2     122.4
 P1.K          -0.1115      0.2543   -0.44   0.662  123.5      51.9
 P1.S          -0.2264      0.2635   -0.86   0.393  153.7      61.8
 P1.G.S        -0.0335      0.4512   -0.07   0.941  221.4      12.6
 P1.G.K        -0.4458      0.4515   -0.99   0.327  211.5      49.6
 P1.S.K         0.2441      0.6582    0.37   0.712  384.3      30.1
 P1.G.S.K       0.0772      0.8429    0.09   0.927  314.1       0.5
```

```
       Method C   (Centered original variables and products)

Predictor  Coefficient  Std. error     t       p      VIF    SEQ SS

Constant     80.919        1.057      76.53   0.000

P1            0.81929      0.09264     8.84   0.000    1.6   10096.1
G           -21.935        2.085     -10.52   0.000    1.6    7908.0
K             2.400        2.701       0.89   0.377    2.4     116.7
S             8.604        2.400       3.58   0.001    2.1    1087.0
G.S         -22.677        4.664      -4.86   0.000    1.8    2129.0
G.K          -7.055        4.998      -1.41   0.162    2.0     295.6
S.K           3.492        6.497       0.54   0.592    3.0      62.2
G.S.K         0.95        11.77        0.08   0.936    2.4      11.0
P1.G          0.1590       0.1887      0.84   0.402    1.6     122.4
P1.K          0.1771       0.2016      0.88   0.382    2.0      51.9
P1.S         -0.1559       0.1927     -0.81   0.421    1.6      61.8
P1.G.S        0.0100       0.3814      0.03   0.979    1.7      12.6
P1.G.K       -0.4164       0.3893     -1.07   0.288    1.8      49.6
P1.S.K       -0.2735       0.4543     -0.60   0.549    2.5      30.1
P1.G.S.K      0.0772       0.8429      0.09   0.927    2.2       0.5
```

```
         Method D   (Raw variables and orthogonal products)

 Predictor  Coefficient  Std. error     t       p     VIF    SEQ SS

 Constant     43.134       7.516        5.74   0.000

 P1            0.76401     0.08419      9.08   0.000   1.3   10096.1
 G           -20.846       1.732      -12.04   0.000   1.1    7908.0
 K             2.015       1.902        1.06   0.293   1.2     116.7
 S             8.358       1.853        4.51   0.000   1.2    1087.0
 G.S         -22.819       4.108       -5.55   0.000   1.4    2129.0
 G.K          -7.308       3.797       -1.92   0.058   1.1     295.6
 S.K           2.208       4.852        0.46   0.650   1.6      62.2
 G.S.K         1.542       9.862        0.16   0.876   1.6      11.0
 P1.G          0.2262      0.1794       1.26   0.211   1.4     122.4
 P1.K          0.1848      0.1777       1.04   0.302   1.4      51.9
 P1.S         -0.1366      0.1826      -0.75   0.457   1.4      61.8
 P1.G.S        0.0113      0.3811       0.03   0.976   1.4      12.6
 P1.G.K       -0.4236      0.3813      -1.11   0.270   1.6      49.6
 P1.S.K       -0.2912      0.4111      -0.71   0.481   1.4      30.1
 P1.G.S.K      0.0772      0.8429       0.09   0.927   1.0       0.5
```

```
                        Analysis of variance.

           Source    DF      SS       MS       F       p

         Regression   15   22034.5   1469.0   24.51   0.000
         Error        76    4555.5     59.9
         Total        91   26590.0

                        2                2
          s = 7.742       R  = 82.9%       R  (adj) = 79.5%
```

Results from the full model.

Exhibit 2 displays the regression coefficients and related statistics for each analysis. The sequential sum of squares ("SEQ SS") accounted for by each variable remains the same across all four methods: this is partly due to specifying the variables in the same order for each method, and partly to specifying them in hierarchical order: linear effects first, followed by successively higher-order interactions.

Notice the instability of the coefficients reported for Method A: the coefficients themselves differ, sometimes wildly, from the corresponding coefficients reported for the other methods, and the variance inflation factors (VIF) are all quite high. (Since VIF =

1/tolerance, all the tolerance values are quite low. Forcing in the 15th predictor has increased all the VIF values by an order of magnitude: compare Exhibit 1.) None of the coefficients, in the presence of the other 14, differs significantly from zero.

Merely recoding the dichotomies from [1,2] to [0,1] for Method B has reduced the VIF values by two orders of magnitude. But apart from P1, whose coefficient is now significant, it is hard to see one's way to a reasonable reduced model by examining the coefficients and their t values. (Good indirect information is available from the sequential sums of squares, but not all regression packages produce these or, equivalently, the sequential changes in 2 R .)

Centering the original variables (Method C) has had the effect of reducing all VIFs to values less than 4. Five regression coefficients are now reported as significant.

In partialling out lower-order terms (Method D) all VIF values are less than 2. They would all be 1.0 if P1 had been partialled out of G, P1 and G out of K, and so on, for a complete orthogonalization. Five regression coefficients are reported with p<0.0005 (these are the same five as those reported in Method C), and a sixth one has p < 0.06.

The results that are invariant across all methods -- the analysis of variance summary table, s (the standard error of estimate, equal to the square root of the error mean square), and R**2 -- are displayed at the bottom of Exhibit 2.

Exhibit 3 shows the correlation matrices for each method. Notice particularly the differences between Method A (for which five correlations exceed .9) and Method D (for which no correlations exceed .4 among the predictors).

EXHIBIT 3. CORRELATION MATRIX FOR EACH METHOD

Correlation Matrix for Method A : Raw products.

```
    P2     P1      G      K      S    G.S    G.K    S.K  G.S.K   P1.G   P1.K   P1.S  P1.G.S
P1.G.K  P1.S.K
P1           616
G           -577   -052
K           -046   -129    066
S            309    285    107    129
G.S         -189    147    667    155    783
G.K         -453   -114    749    681    182    590
S.K          180    112    139    640    821    691    522
G.S.K       -198    041    574    533    707    897    785    862
P1.G        -255    399    884    005    211    672    627    165    536
P1.K         259    347    038    874    248    206    582    656    518    192
P1.S         488    601    055    045    929    688    086    704    576    306    321
P1.G.S      -003    415    591    093    789    953    491    644    815    733    276    802
P1.G.K      -247    200    713    631    250    614    940    540    777    749    697    258
611
P1.S.K       354    382    100    554    830    660    436    951    791    250    716    824
702
550
P1.G.S.K    -053    259    540    487    735    891    727    851    967    611    589    689
884
803     857
```

```
Correlation Matrix  for  Method B :    [0,1] dichotomies.

               P2     P1      G      K      S    G.S    G.K    S.K  G.S.K   P1.G   P1.K   P1.S
P1.G.S
P1.G.K  P1.S.K
P1            616
G             577    052
K             046    129    066
S             309    285   -107   -129
G.S           711    271    470    047    470
G.K           234    111    494    586   -038    255
S.K           275    328    076    467    394    362    339
G.S.K         335    178    272    322    272    579    550    691
P1.G          645    173    982    083   -055    536    511    110    316
P1.K          116    264    072    978   -078    073    584    540    358    105
P1.S          382    400   -079   -089    983    501   -019    449    296   -012   -027
P1.G.S        726    333    464    051    464    986    257    363    579    548    085    514
P1.G.K        262    182    487    577   -016    282    985    369    591    525    597    011
296
P1.S.K        279    366    066    462    390    344    322    990    662    109    547    457
359
365
P1.G.S.K      344    221    268    318    268    570    542    681    985    325    366    305
589
602     673


Correlation Matrix  for  Method C :    Centered variables.

               P2     P1      G      K      S    G.S    G.K    S.K  G.S.K   P1.G   P1.K   P1.S
P1.G.S
P1.G.K  P1.S.K
P1            616
G            -577   -052
K            -046   -129    066
S             309    285    107    129
G.S          -290   -110   -054    117    054
G.K          -065    031   -032   -055    114    066
S.K          -110   -128    119   -114    066    115    000
G.S.K        -078   -115    074    009    125   -009   -123   -070
P1.G         -093   -050    025    031   -107    347   -138   -109   -063
P1.K         -091   -295    028    097   -110   -090   -062    376    003    002
P1.S          184    177   -108   -124    140   -138   -130   -234   -058   -145    075
P1.G.S       -346   -152    317   -106   -099    195   -057   -010   -158    233    091   -093
P1.G.K       -017    007   -126   -069   -099   -064    126   -014    351   -289   -042    100
-007
P1.S.K        070    042   -098    338   -164   -010   -003    285   -063    100    092   -236
045
-199
P1.G.S.K      078    103   -057   -010   -020   -141    334   -061    305   -015   -209    085
-278
198    -147
```

```
Correlation Matrix  for  Method D :    Orthogonal interactions.

               P2     P1      G      K      S    G.S    G.K    S.K  G.S.K   P1.G   P1.K   P1.S
P1.G.S
P1.G.K  P1.S.K
P1            616
G            -577   -052
K            -046   -129    066
S             309    285    107    129
G.S          -344   -131    000    114   -000
G.K          -084    023   -000    000    124    063
S.K          -142   -170    120   -000    000    134   -013
G.S.K        -104   -167    000   -000    000    000    000    000
P1.G         -050   -000   -000    023   -095    350   -135   -108   -091
P1.K          093   -000    009    000   -038   -132   -052    369   -000   -016
P1.S          063   -000   -112   -119   -000   -137   -158   -243   -093   -128    147
P1.G.S       -017    000   -000   -162    000   -000   -029   -077   -204   -000    078   -000
P1.G.K       -094    000   -000    000   -120    028   -000   -018    393    000   -000    077
089
P1.S.K        115   -000   -166   -000   -000   -109    038   -000    046    121   -000   -000
193
-235
P1.G.S.K      004    000    000   -000    000   -000   -000   -000   -000   -000   -000    000
000
000    -000
```

Selecting a reduced model.

From Exhibit 2 we can see from any of the methods that the four-way interaction can be discarded (b = .077, t = .09, p > .9). (From Exhibit 1, where the four-way interaction was omitted, we can see that all the three-way interactions can also be discarded.) Method A provides no insight into further reducing the model: all p values are > .4, all but one are > .5.

In Method B the full model shows that P1 should be retained, but doesn't help identify other useful predictors (except by using the sequential sums of squares, which some statistical packages do not report).

When the original variables are centered (Method C), four coefficients are significant in the full model: P1, G, S, G.S. (This may not be the best reduced model, because one or more of the other predictors might be significant in the absence of the rest.)

When the interaction terms have been orthogonalized (Method D), the same four predictors should clearly be retained; in addition, the G.K interaction (p <.06) may be significant when the other predictors are absent, and this possibility should be examined directly.

If the G.K interaction is included as a predictor, so should be the K linear effect, for hierarchical completeness; particularly for Methods A, B, and C. Results for this hierarchical six-predictor model (P1, G, K, S, G.S, G.K) are displayed in Exhibit 4, for each method.

- Using the original variables and their raw products for interactions (Method A), all six predictors are significant.

- When the dichotomies are recoded to [0,1] (Method B), the two linear terms S and K are not significant. (When they are discarded, the four-predictor reduced model yields the results displayed in Exhibit 5a.)
- For centered original variables, and for orthogonalized interactions (Methods C and D), the linear term K is not significant.

Exhibit 4

Reduced (hierarchical) model fitted by each method: predicting P2 from P1, G, S, G.S, G.K, and K.

```
Method        Predictor      Coef        St dev       t       p      VIF    Seq SS

METHOD  A          Constant      -24.62       13.67       -1.80   0.075

(Raw)              P1        0.76297     0.07786     9.80   0.000    1.1   10096.1

                          G       21.108       7.556       2.79   0.006   21.2    7908.0
                          S       42.334       6.074       6.97   0.000   13.7    1177.8
                          G.S    -20.649       3.510      -5.88   0.000   24.0    2066.5
                          G.K     -8.030       3.567      -2.25   0.027   26.8      14.4
                          K       14.891       5.914       2.52   0.014   11.7     369.7


METHOD  B          Constant       16.296       5.610       2.90   0.005

( [0, 1] coding)  P1        0.76297     0.07786     9.80   0.000    1.1   10096.1

                          G       15.600       2.392       6.52   0.000    2.1    7908.0
                          S        1.036       2.137       0.48   0.629    1.7    1177.8
                          G.S     20.649       3.510       5.88   0.000    2.0    2066.5
                          G.K     -8.030       3.567      -2.25   0.027    2.3     369.3
                          K        1.169       2.325       0.50   0.616    1.8      14.8


METHOD  C          Constant       80.6373      0.8024    100.49   0.000

(Centered)               P1        0.76297     0.07786     9.80   0.000    1.1   10096.1
                          G      -21.012       1.663     -12.64   0.000    1.0    7908.0
                          S        8.892       1.769       5.03   0.000    1.2    1177.8
                          G.S    -20.649       3.510      -5.88   0.000    1.0    2066.5
                          G.K     -8.030       3.567      -2.25   0.027    1.0     319.1
                          K        1.886       1.786       1.06   0.294    1.1      65.0


METHOD  D          Constant       42.292       7.045       6.00   0.000

(Orthogonal)             P1        0.76297     0.07786     9.80   0.000    1.1   10096.1
                          G      -20.316       1.657     -12.26   0.000    1.0    7908.0
                          S        8.302       1.764       4.71   0.000    1.2    1177.8
                          G.S    -20.649       3.510      -5.88   0.000    1.0    2066.5
                          G.K     -8.030       3.567      -2.25   0.027    1.0     303.4
                          K        2.096       1.782       1.18   0.243    1.1      80.7


                      Common results for all methods in Exhibit 4:

                                   2              2
                     s = 7.637     R  = 81.4%     R  (adj) = 80.0%

                                Analysis of Variance

                     Source    df      SS       MS       F       p

                   Regression    6   21632.5   3605.4   61.82   0.000
                   Error        85    4957.5     58.3
                   Total        91   26590.0
```

Exhibit 5

(A) OPTIMAL REDUCED MODEL, METHOD B

```
                 Coefficients

                Predictor     Coef      St dev      t       p      Seq SS

                Constant     16.352      5.556      2.94   0.004

                P1            0.77275    0.07546   10.24   0.000   10096.1
                G            14.868      2.064      7.20   0.000    7908.0
                G.S          21.582      2.869      7.52   0.000    3234.2
                G.K          -6.893      2.707     -2.55   0.013     371.1



                      s = 7.566     R**2  = 81.3%     R**2  (adj) = 80.4%

                                   Analysis of Variance

                      Source    df      SS       MS       F       p

                    Regression    4   21609.5   5402.4   94.37   0.000
                    Error        87    4980.5     57.2
                    Total        91   26590.0
        (b)  Intercepts for
                                   six-predictor model

                          Value         Description          N

                                  Those who ran in place :

                          53.58   Female nonsmokers            7
                          46.72   Female smokers               4
                          31.90   Male nonsmokers            16
                          25.04   Male smokers                 8

                                  Those who did not run :

                          18.50   Female smokers               4
                          17.47   Male smokers               12
                          17.33   Female nonsmokers          20
                          16.30   Male nonsmokers            21

For those who ran in place :


                        P2  = 0.763 P1  + 31.90 + 21.68 S - 6.86 K

For those who did not run :


                        P2  = 0.763 P1  + 16.30 + 1.036 S + 1.169 K



                                    (c)  Intercepts for
                                  four-predictor model

                          Value         Description          N

                                  Those who ran in place :

                          52.80   Female nonsmokers            7
                          45.91   Female smokers               4
                          31.22   Male nonsmokers             16
                          24.33   Male smokers                 8

                                  Those who did not run :

                          16.35   All persons                57
```

For those who ran in place :

$$P2 = 0.773 P1 + 31.22 + 21.58 S - 6.89 K$$

For those who did not run :

$$P 2 = 0.773 P1 + 16.35$$

INTERPRETATION

We have seen that if we partial out lower-order terms in constructing interaction variables, we can immediately decide on an appropriate reduced model after running only one regression, fitting the full model containing all interactions. That model may not be quite the optimal or preferred model for describing and interpreting the behavior of the variables in the data set, as we shall see; but we will have managed to discard most of the variables -- the interactions in particular -- that do not contribute usefully to accounting for variance in the dependent variable.

We have also seen that (at least sometimes) centering the original variables before constructing the product terms can do almost as well.

Neither of these methods, however, may be useful for interpreting the results associated with the reduced model that eventually emerges.

Having discovered that the two-way interactions G.S and G.K ought to be included in an initial reduced model, and that the linear terms P1, G, and S appear to be necessary as well, we also include K in the first reduced model (because G.K may not be interpretable in the absence of K -- in other words, we retain a model that is hierarchical). This model, predicting P2 from (P1, G, S, K, G.S, G.K), we fit using each of the four methods, for comparison.

For all four methods, there is only one continuous predictor, P1. All the other predictors are categorical variables, whose regression coefficients thus imply a difference in the intercept -- that is, a difference in the height between parallel lines -- on a plot of P2 vs. P1. (Had P1 interacted with any of the other predictors, some of the lines would have different slopes from the others.) Exhibit 6a displays such a plot, for the six-predictor models whose coefficients are given in Exhibit 4. Intercepts of the eight regression lines are reported in Exhibit 5b, in order from the highest line to the lowest. For the four-predictor model of Exhibit 5a, the intercepts are reported in Exhibit 5c and the plot appears in Exhibit 6b.

Exhibit 6a and 6b can be found at the end of the document.

Method B: [0,1] dichotomies.

In the data used as an example throughout this paper, interpretation is easiest when the dichotomies have been scored [0,1] and the interaction terms are the simple products of the linear variables (Method B): G.S = 1 when both G = 1 and S = 1, that is for females who ran, and is zero otherwise; similarly, G.K = 1 for smokers who ran, and is zero otherwise; S.K = 1 for female smokers, and is zero otherwise; and G.S.K = 1 for female smokers who ran, and is zero otherwise. For data constructed by this method, the most parsimonious model is the one in which the predictors are P1, G, G.S, and G.K, reported in Exhibit 5a.

- For the group that did not run, in this method G = 0, and thus G.K = G.S = 0 : the only significant predictor of the final pulse rate is the initial pulse rate, not surprisingly.
- For male non-smokers who ran, G = 1, bG = 14.9 -- that is, the pulse rate in this group is 15 beats per minute faster on average than for those who did not run.
- For females who ran, G = 1 and G.S = 1, bG.S = 21.6 -- pulse rates for females who ran averaged 22 beats per minute faster than for males who ran.
- For smokers who ran, G = 1 and G.K = 1, bG.K = -6.9 -- pulse rates for smokers who ran averaged 7 beats per minute slower than for non-smokers who ran.

Method A: [1,2] dichotomies.

For Method A, the most parsimonious model includes all six predictors. Interpretation is more difficult than for Method B for several reasons:

- The two interaction variables each have three values (1, 2, 4) instead of two.
- While for Method B the [0,1] coding and the absence of terms for S and K made it clear that for those who did not run there is only one predictor of the second pulse rate (namely the initial pulse rate), in Method A this can only be observed by computing the effective intercept for each of the four groups who did not run and observing that these four intercepts are virtually identical.

Method C: centered variables.

Here the most parsimonious model includes five predictors, but we should retain all six for hierarchical completeness. Interpretation is even more difficult than for Method A:

- Initial pulse rate is now expressed as a deviation from the sample mean, rather than in straightforward number of beats per minute. It thus depends on characteristics of the sample which may not be generalized to other situations: in particular, on the sex ratio in the sample.
- The dichotomies now have the values [-p,+q], where p + q = 1 and p is the proportion of cases falling in the category originally scored "2".
- The two interactions now have four values each; unless p for one dichotomy is equal to either p or q for the other, in which case there are three values.

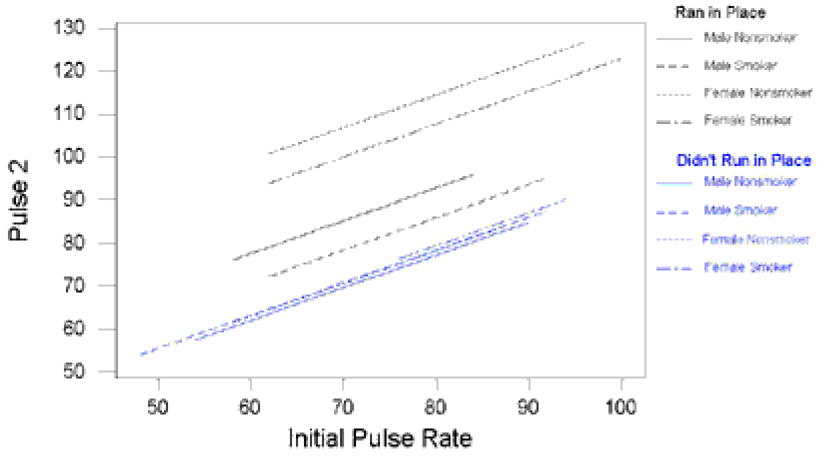
Method D: orthogonalized interactions.

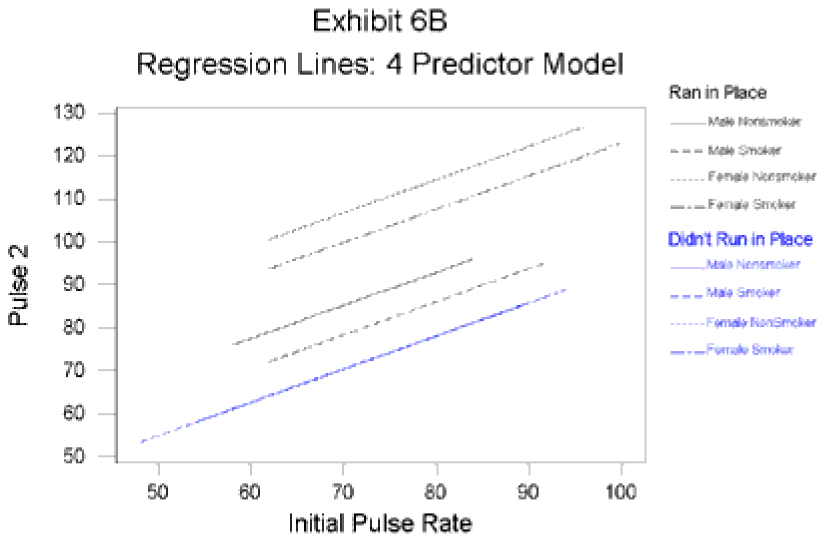
Here the most parsimonious model includes five predictors, omitting K; but unlike Method C, the non-significant K can be omitted, because the orthogonality of G.K implies that K is not useful in its own right, regardless of whether the model includes G.K or not. The significant G.K interaction implies that the effect of smoking for those who ran differs from the effect for those who did not run; the nonsignificant K effect implies that there is no overall effect of smoking.

Since P1, G, and S are in their original form, the interpretation of their coefficients is similar to the interpretation given for Method A. However, the interaction terms are more opaque: we need first to find out what their values are for the four cases represented by each interaction, before we can interpret the coefficients. These values in turn depend on the results of the regression analyses used to orthogonalize the interactions. Interpreting the regression coefficients for Method D is therefore more difficult than for the other methods. While Method D helps us to arrive rapidly at a reduced model, once we have such a model we may prefer to apply one of the other methods if a goal of the enterprise is to interpret the regression coefficients obtained.

Exhibit 6A

Regression Lines: 6 Predictor Model





REFERENCES

Aiken, L. S., and S. G. West (1991). Multiple regression: Testing and interpreting interactions. Newbury Park: Sage Publications.

Bottenberg, R. A. and J. H. Ward, Jr. (1963). Applied multiple linear regression. Technical Documentary Report PRL-TDR-63-6. Lackland AFB, Texas: U. S. Department of Commerce, Clearinghouse for Federal Scientific and Technical Information.

Cohen, J. (1978). Partialed products are interactions; partialed vectors are curve components. Psychological Bulletin, 85, 858-866.

Darlington, R. B. (1990). Regression and linear models. New York: McGraw-Hill.

Draper, N. and H. Smith (1966). Applied regression analysis. (Revised ed., 1981). New York: John Wiley & Sons.

Judd, C. M. and G. H. McClelland (1989). Data analysis: A model-comparison approach. New York: Harcourt Brace Jovanovich.

Ryan, B. F., B. L. Joiner and T. A. Ryan Jr. (1985). Minitab Student Handbook (2nd ed.). Boston: Duxbury Press.