TN 20: ANALYSIS OF VARIANCE MODELS WITH INTERACTION EFFECTS AND THEIR POTENTIAL ROLE IN UNDERSTANDING AND PREDICTING RECREATION BEHAVIOUR

By J. Arseneault, A. Dionne, J. Beaman, M. Renoux

ABSTRACT

This paper focuses on certain issues that are important in understanding the value of analysis of variance, ANOVA*, models in recreational research, particularly whether a simple analysis of variance model is structurally sound and whether its use may lead to errors in estimating behaviour.

The results of the Michigan Automatic Interaction Detector (AID) program and of regression analyses of the 1969 and 1972 Canadian Outdoor Recreation Demand Study, CORDS, National Survey Data on Canadian's Participation in Outdoor Activities are included to illustrate how interaction effects affect analysis when trying to explain participation in an activity using socio-economic variables as the independent variables. The results of the analysis of simulated data are presented to show the degree to which AID explains data as compared to using a correct analysis of variance model. Conclusions are that:

1. significant interaction effects exist when one tries to explain Canadian residents' participation in particular outdoor activities in terms of socio-economic characteristics;
2. a simple main effect analysis of variance model is not adequate to explain most recreational behaviour;
3. the use of the AID analysis program gives one an idea of the magnitude of the sum of squares associated with interaction effects but its use does not provide a systematic way of identifying effects; and
4. repeated application of traditional regression methods to identify interactions does result in "finding" interaction terms that improve a simple linear model but the improvement achieved can be limited and, what is more, the possibility of type II errors in using regression repeatedly to determine interaction effects raises serious questions about using this approach to improve (define appropriate) models.

NOTE: In CORDS using analysis of variance, ANOVA, does not refer to running a program that "partitions" variance based on the assumption that data were collected using a designed experiment. In the terminology of 2006, one is referring to using multiple regression to *analyze the variance in a dependent variable based on the values that independent variables happened to take – in the general case based on the "general linear model" presented in Scheffe 1959 pp. 13-22).

INTRODUCTION AND PURPOSE

As of the 1970s, the best known use of analysis of variance modelling techniques for predicting participation in recreation was Mueller and Gurin (1961). More recently other examples of applying this technique have become frequent. In the Canadian Outdoor Recreation Demand (CORD) Study, Hendry (1970) suggested using analysis of variance (in the form of dummy variable analysis); this technique was actually pursued using a 1969 National Survey to develop differentials related to age, sex, family status, etc. Subsequently, TN 12 of the CORDS showed how the socio-economic differentials for 26 activities could be used in making projections. Renoux (1973, 1975) used this methodology to develop a hunting model and other models for the Province of Quebec.

An important result of previous CORD Study work and Renoux's research has been the identification of a number of problems associated with the use of analysis of variance techniques. The purpose of this paper is to focus on the need to include interaction effects in modelling behaviour. The structure of the note reflects the history of research on interaction effects and regression models. It is convenient to present some definitions and then describe some early research from which no results are presented. Early research led to successful research on interaction effects and finally led to some findings that are presented in this paper. Through this strategy of showing the background of research the authors believe that the reader will get the best "feel" possible for the multitude of problems involved in developing and improving the kind of models of concern here.

The data used in the various analyses presented are not described in any detail but are documented in CORDS Volume III (consult the CORDS web posting to find out about availability of the data as well as of documentation).

DEFINITIONS

Here the term main effects model is used to refer to a model in which participation or non-participation in an activity is expressed as the cumulative sum of socio-economic effects. It is the kind of model defined and described in detail in TN 12 (see also TN 15). In equation form such a model is:

(1) $Y(i,J,,L,M,Q) = U + B(1,J) + B(2,K) + B(3,L) + B(4, M) + B(5,Q) + \varepsilon(i)$

WHERE $Y(i,J,K,L,M,Q)$ is 0 or 1 depending on whether individual i participated in the activity being considered;

J,K,L,M,Q refer to the socio-economic categories that person i is in where J refers to a category of the socio-economic variable 1, K refers to a category of the socio-economic variable 2, say Age, L refers to a category of the socio-economic variable 3, etc.

U is a general level/probability of participation in the activity under consideration;

$B(1,J)$ is the effect of being in category J of the socioeconomic varIable 1, $B(2,K)$ Is the effect of being in category K of the socioeconomic variable 2, and $B(3,L)$ is the effect of being in category L of the socio-economic variable 3, etc.

$\varepsilon(i)$ is an error term.

It is possible that the kind of model just defined oversimplifies the interrelationships between the variables used to predict behaviour. For example, it is likely that relationships between gender, age and education exist that help one understand hunting participation. For example, having a high level of education and being old may mean something quite different from having a high level of education and being young. A person with a high level of education who is old may tend to come from an urban non-hunting background and thus have a very low probability of hunting, particularly if female. The main effects model does not allow for the interaction effect between education and age affecting hunting just described except in as much as one calculates one model for males and one for females as is done in TN 12. (TN 27 pursues the effects of interactions in another context.)

Consider a very simple model involving interactions. It is assumed that the participation in an activity can be explained by means of two socio-economic effects, A and B and by the Interaction between them, AB. Stated mathematically, individual behaviour is represented by Equation 2a and the somewhat more general form Equation 2b.

(2a) $Y(i,j,k) = U + A(i) + B(j) + AB(i,j) + \varepsilon(k)$

(2b) $Y(i,j,k) = \mu + \beta(i,1) + \beta(j,2) + \beta(i,j,3) + \varepsilon(i)$

WHERE $Y(i,j,k)$ is the dependent variable, μ is an average effect for all individuals to which other effects are a “correction”;

β(i,1) is the differential (offset from μ) effect of factor/variable at level i;

β(j,2) is differential effect of factor two being at level j;

β(i,j,3) is the adjustment to β(i,1)+β(j,2) that is necessary because these occur together;

ε(i) is a random error term applying to a particular individual.

There are conditions which the β( ) coefficients must satisfy which the reader may review in a number of sources (e.g. Wasserman, 1974). However, the important point here is that to have a great deal of freedom in how interaction effects are defined, the kind of formulation introduced here can be used. The formulation is such that many unknowns are usually computed to explain behaviour. μ and (m-1), (n-1) and (m-1)(n-i) values of β(,1)'s, β(,2)'s and β(, ,3)'s respectively must be computed to define the model when all of the usual constraints on the model parameters are considered. Now if in Equation 2, instead of just having three types of β( )'s one wants to have the 10 socio-economic variables for which effects were calculated in TN 12, then to have a model like the one just introduced with interaction terms between every pair of variables there must be (10)(9)/2 = 45 such terms. For N variables the number of terms is (N)(N-1)/2.

Furthermore, just as for one interaction term for variables with m and n levels there are (m-1) times (n-1) unknown β( )'s to be calculated. For a "complete" 2-way interaction model with 10 variables this means the number of parameters to computed is Σ (n-1)(m-1) over 45 second order interaction terms. Even if one estimates all the 2-way interaction effects they have not necessarily dealt adequately with interaction. One can consider 3-way interactions. These depend on the value of three variables. If all three-way interactions between variables are to be considered when there are N variables in a model, the number of 3-way interactions is N(N-1)(N-2)/(3)(2)(1) which equals 120 for N = 10. As one might guess, for each of these there are (n-1) (m-1) (p-1) interaction β(,)'s to be estimated (n, m and p refer to the number of levels of the three variables of a particular third order interaction). So it should be clear that one cannot simply insert all possible first, second, third etc. order interaction terms into a model and proceed to estimate parameters.

The discussion thus far suggests that when very many socio-economic variables are considered, the number of coefficients is so large that estimates could not be made on any 1970s computer. A more practical consideration, and one relevant any time, is that when large numbers of parameters are to be estimated and these involve complicated interactions, the situation readily arises where, even when a very large data set is available, there are only one or two people who fit into certain classes/categories on which there must be information to estimate interaction parameters. The problem is compounded when there are numerous categories with nobody. Without information coefficients cannot be estimated.

Beaman and Renoux recognized the kind of problem just described when they began work on the research which eventually led to this TN. They examined the differences between the model defined by Equation 1 and a much more adequate model “including” interaction affects. The more adequate model that they chose to use was a model defined by the Michigan AID Computer Program. The reader can learn how this program works from material in TN 4 or TN 27 where there are examples of its use. One can also refer to the original writings about the AID Program by Sonquist and Morgan (1964).

What was done by Renoux and Beaman was to try to determine which interactions should be considered if a model was to fit a given set of data. Using the model defined by Equation 1 and using an AID model, predictions for individuals were written out on magnetic tape (one available storage medium to use in the 70s) when a “large” amount of information was to be stored for further computer processing. The idea was that if cases could be identified where there were large differences between AID predictions and the analysis of variance predictions, these would provide a clue as to what interaction terms should be incorporated into the main effect regression model. It was planned that the differences between these two predicted values would (for example) be examined by looking at the average value of it for various cross classifications of socio-economic variables.

Unfortunately, much work led to few results. As one might guess from the variance-explained values reported in Table 1, the differences between AID predictions and analysis of variance predictions were highly variable. Little was learned about interactions to add but what was evident was that AID analyses explained much more variance than the main effects regression model.

In Table 1 you find quite a comprehensive set of percentage of variance-explained values in which AID results on the 1969 CORD Study National Survey Data on Canadian Residents' Participation in Outdoor Activities are compared with main effect model results. Regression results for 1972 are also given for reference purposes. The R^2 values for 1972 ANOVA are comparable with the 1969 values because they were produced for persons 18 years of age which was the sampling universe in the 1969 study. Also they were produced for participation-non participation as the dependent variable and with the same independent variables. What one should notice from Table 1 is that, as a rule of thumb, use of AID resulted in explaining about twice as much variance as the main effects regression model. If one looks at the list of activities in column one of Table 1 they first see swimming participation in a city and they may note that for males 17% of the variance was explained using AID. The use of ANOVA resulted in 13% of the variance being explained. This is obviously not a 2 to 1 ratio but, then, for bird-watching for males, one notices the balance shifting as with ANOVA 2.5% of the variance is explained compared to 14% for AID. For outdoor photography for males, there is 12% for AID and 6% for ANOVA, which is very close to the 2 to 1 ratio suggested earlier. A similar ratio holds for male use of Historic Sites and when one examines R^2 for the female use model for Historic Sites one sees that there is a ratio of .08 to .04 or 2 to l. The results continue in a similar manner. In the odd case the ANOVA model is not too much "poorer" than the AID model while most results show that there was a great deal of variance to be explained which the ANOVA model does not explain.

When this situation was recognized the decision was made to involve other researchers in the attempt to find interaction effects of the magnitude that (it appeared clear from the difference between AID and ANOVA analyses) it should be possible to find. The researchers who took on this task were confronted with two problems. One was becoming familiar with CORD data study, and the other was developing a strategy for estimating interaction effects that would explain something like the amount of variance that it seems clear was possible to explain by interaction effects. This thrust of the research effort began with an exploratory analysis of the CORD Study 1972 National Survey data on Canadian Residents' Participation in Outdoor Activities, which data had become available since Renoux and Beaman had begun their work. The variables used from these data and their coding are shown in Table 2.

TABLE 1: COEFFICIENTS OF DETERMINATION, R^2'S, FOR PARTICIPATION IN EACH ACTIVITY IN 1969 AND 1972, FOR MALES AND FEMALES, OBTAINED THROUGH ANOVA AND AID

		Males			Females	
	1969	1969	1972	1969	1969	1972
	AID		ANOVA	AID		ANOVA
PARTICIPATION IN CITY						
1. Swimming	(.176)	.132	---	(.194)	.139	-–
2. Nature/Bird Watching	(.142)	.024	---	(.075)	.016	-–
3. Outdoor Photography -	(.119)	.062	---	(.106)	.053	-–
4. Visit Historic Sites	(.117)	.062	.051	(.084)	.042	.036
5. Visit Other Parks	(.108)	.060	---	(.108)	.049	-–
6. Drive for Pleasure	(.086)	.050	.043	(.077)	.042	.045
7. Sightseeing Urban	(.105)	.066	.030	(.090)	.057	.025
8. Toboggan/Sledding	(.191)	.0b6	---	(.165)	.039	-–
9. Picnicking	(.111)	.035	.018	(.102)	.049	.019
10. Walk/Hiking	(.146)	.091	.064	(.125)	.079	.059
11. Golfing	(.117)	.069	---	(.151)	.034	-–
12. Ice Skating	(.208)	.159	.094	(.203)	.146	.059
13. Bicycling	(.171)	.081	.085	(.194)	.096	.102
PARTICIPATION IN COUNTRY						
14. Swimming	(.202)	.155	---	(.168)	.144	-–
1b. Nature/Bird Watching	(.075)	.014	---	(.088)	.032	-–
16. Visit Historic Sites	(.096)	.056	.075	(.101)	.058	.065
17. Visit Other Parks	(.092)	.052	---	(.094)	.063	-–
18. Drive for Pleasure	(.081)	.057	.056	(.095)	.060	.071
19. Sightseeing	(.101)	.073	.058	(.128)	.088	.063
20. Toboggan/Sledding	(.180)	.088	---	(.164)	.087	---
21. Picnicking	(.114)	.082	.092	(.133)	.096	.100
22. Walk/Hiking	(.121)	.065	.069	(.099)	.061	.063
23. Golfing	(.128)	.077	---	(.141)	.025	---
24. Ice Skating	(.139)	.056	.098	(.151)	.055	.072
25. Bicycling	(.148)	.046	.063	(.142)	.076	.066
OTHER PARTICIPATION						
26. Swimming	(.280)	.244	---	(.260)	.217	---
27. Tent Camping	(.120)	.069	.179	(.088)	.043	.100
28. Trailer Camping	(.061)	.022	.038	(.084)	.082	.036
20. Pickup Camping	(.156)	.019	.024	(.125)	.022	.024
30. Hunting	(.118)	.084	.091	(.150)	.035	.030
31. Power Boating	(.127)	.088	.071	(.112)	.064	.065
32. Canoeing	(.146)	.080	.094	(.094)	.036	.066
33. Sailing	(.131)	.053	.058	(.217)	.032	.061
34. Water Skiing	(.186)	.155	---	(.215)	.067	-–
35. Nature/Bird Watching	(.068)	.017	---	(.081)	.029	-–
36. Outdoor Photography	(.122)	.079	---	(.091)	.055	-–
37. Visit Historic Sites	(.112)	.078	.092	(.101)	.073	.074
38. Visit Other Parks	(.101)	.076	---	(.105)	.077	-–
39. Drive for Pleasure	(.102)	.066	.069	(.104)	.067	.095
40. Sightseeing	(.112)	.088	.062	(.130)	.099	.060
41. Climbing	(.095)	.037	---	(.130)	.029	-–
42. Snow Skiing	(.208)	.125	.090	(.223)	.102	.087
43. Snowmobiling	(.160)	.103	.131	(.127)	.074	.103
44. Toboggan/Sledding	(.173)	.107	---	(.158)	.102	-–
45. Picnicking	(.133)	.090	.101	(.148)	.115	.104
46. Walk/Hiking	(.142)	.086	.091	(.114)	.078	.089
47. Golfing	(.174)	.121	---	(.141)	.043	-–
48. Ice Skating	(.233)	.193	.176	(.226)	.171	.128
49. Horseback Riding	(.198)	.102	.108	(.213)	.109	.114
50. Bicycling	(.156)	.083	.135	(.182)	.118	.147
51. Tennis	(.235)	.139	---	(.266)	.117	-–
52. Fishing	---	---	.084	---	---	.053
53. Hunting/Fishing	---	---	.104	--	---	.058
54. Small Game Hunting	--	---	.077	--	---	.032

TABLE 2: 1972 CANADIAN'S PARTICIPATION IN OUTDOOR ACTIVITIES VARIABLES USED IN ANALYSES REPORTED IN THIS PAPER

Variable Description	Original Value	Recoded for Table 3
I. AGE		
10 to 11 years	1	1
12 14	2	1
15	3	1
16 17	4	1
18 19	5	1
20	6	2
21 24	7	2
25 29	8	2
30 34	5	3
3S 39	10	3
40 44	11	4
49	12	4
50 55	13	4
56 b4	14	4
65 and over	15	4
II. EDUCATION		
No formal	0	1
Some public school	1	1
Finished public	2	1
Some High School	3	2
Finished High School	4	3
Some tech-Senior College	5	3
Graduate of tech-Senior College	6	3
Some university	7	3
Graduate of university	8	3
III. FAMILY SIZE		
One	1	1
Two	2	2
Three	3	3
Four	4	4
Five	5	4
Six	6	5
Seven	7	5
Eight	8	5
Nine	9	5
Ten and over	10	5
IV. INCOME		5
0 - $ 2,999	1	1
3,000 - 4,499	2	1
4,500 - 5,999	3	1
6,000 - 7,499	4	2
7,500 - 8,999	5	2
9,000 - 10,499	6	2
10,500 - 11,999	7	3
12,000 - 13,999	8	3
14,000 and over	9	3
V. CITY SIZE		
500,000 and over	1	1
100,000 - 500,000	2	1
30,000 - 100,000	3	2
10,000 - 30,000	4	3
1,000 - 10,000 -	5	4
Rural	6	5
VI. GENDER		
Male	1	0
Female	2	1

Initially, a number of equations were derived to give the researchers a feel for what second order interaction effects were relatively important (Arsenault, Dionne, & Ritchie 1975). A methodology adopted for doing this was as follows:

a) Each of several control variables, sex, age and education, was chosen in turn (see right-hand column of Table 3).

b) Regressions were carried out to determine how the form of an equation to explain participation in hunting depended on the value of the control variable. For example, with age as the control variable, participation in hunting was predicted for persons 10 to 19 with a first independent variable/ sex = x(1), then for each of the other independent variables X(2) to X(5), resulting in the equations which follow and the others that would be written if one followed across the first line under the heading age:

P(of person 10 - 19) = C - .222*(1)sex relation
" = C - .088*(3)education relation
" = C - .000*(4)household size relation
" = C - .003*(5)income relation

WHERE C is a constant.

c) Similar results were derived for other levels of the control variables. Specifically, the age group 20 to 29 regressions were made giving equations like the ones above with the first three such equations havIng coefficients of X(1) x (3) and x(4) of -.0222, .021 and 007 respectively.

To comment further, as shown in Table 3 when age is the independent variable the equations obtained to estimate the probability of hunting for females is:

(3) $\hat{p}$ = 0.064 - 0.012 x(2)

and for males,

(4) $\hat{p}$ = 0.341 - 0.052 x(2)

Because the lines defined by Equations 3 and 4 are not parallel and have slopes that are significantly different at the .05 level, (the t-test was applied) it may be concluded that there is an interaction effect between sex and age. If there were no interaction effect then the difference between the sexes could be accounted for by a sex effect as in the two equations following:

(5) $\hat{p}$ (for example) = constant + male effect + B x(2)

(6) $\hat{p}$ (for example) = constant + female effect + 3 x(2)

WHERE B is a regression coefficient of age that applies to both sexes

Since, as one can see from Table 2, variable x(2) has more than two values (e.g. 1, 2, 3, 4, 5, etc.), one can write the following based on Equations 3 and 4:

$\hat{p}$ = .J64 + .012 = .076 for x(2) = 1 for a female

$\hat{p}$ = .064 + (.012)2 = .088 for x(2) = 2 for a female etc. ·· for all levels of x(2)

$\hat{p}$ = .341 - .052 = .J93 for x(2) = 1 for a male etc. ·· for all levels of x(2)

The reader can readily confirm that the system of equations given above cannot be solved so that parameters are determined which make Equation 4 compatible with Equations 5 and 6. Having the two coefficients of .012 and .072 in Equations and 4 makes it possible to reflect the fact that age has a much more pronounced effect on male participation in hunting than it does for females in the sense that young males may have a very high probability of hunting while older males have a very Low probability similar to the general level of hunting for females. For females, what is necessary to reflect behaviour is that there be a quite drastic peak in probability of participating from almost nothing to maybe a .10 probability of participating. However, this peak in relative terms is not drastic in absolute terms since compared to males one sees that old males have almost no probability of participating whereas a young male has a 50% chance of participating.

TABLE 3: METHOD l: RESULTS OF PREDICTING HUNTING PARTICIPATION* USING 1972 CORD NATIONAL SURVEY DATA

		Independent Variables Used in Regression with Selected Control Variables		
		Sex	Age	Education
"Control" Variables		X(1)	X(2)	X(3)
Sex				
Males	(1)		-.052	.022
Females	(2)		-.012	.003
Age				
10 to 19 years	(1)	-.222		.088
20 to 29 years	(2)	-.223		.021
30 to 39 years	(3)	-.216		.021
40 years & over	(4)	-.101		.003
Education				
No Formal-Finished public school	(1)	-.135	-.026	
Some high school	(2)	-.250	-.054	
Finished high school and +	(3)	-.159	-.029	

	Household Size	Income	City Size
"Control" Variables	X(4)	X(5)	X(6)
Sex			
Males	.031	.020	.343
Females	.006	.012	.009
Age			
10 to 19 years	.000	.003	.045
20 to 29 years	.007	.004	.026
30 to 39, years	.010	.022	.027
40 years 6 over	.006	.022	.016
Education			
No Formal-Finished public school	.014	.034	.025
Some high school	.034	.027	.033
finished high school and +	.006	.005	.021

*** See the text for material on how to read the Table. Also one may note that significance test on differences between the β's were calculated but are not presented here because they play no rote in the discussion or in arriving at the conclusions reached in this paper.**

This should make it clear why the great number of drastically differing slopes in Table 3 present clear evidence that there are interactions that should be considered in developing models to explain peoples' participation in outdoor activities. One could present statistical tests for the difference in coefficients, which is what was done in an earlier report on the data presented in Table 3 (Arsenault, Dionne & Ritchie 1975). But this is not done since there are problems in comparing regression coefficients (a) because they are inter-correlated, (b) because the results of including third and fourth variables are not considered and also (c) because subsequent results presented in this paper are more important in confirming the magnitude and significance of interactions. Those who wish to look at more material on what interactions there are and on their detection may refer to Renoux (1973, 1975) for specific examples that have to do with the data of concerned here. More general discussion is found in Sonquist & Morgan (1964).

A FIRST ATTEMPT TO DERIVE A FAIRLY GENERAL MODEL WITH INTERACTIONS

The step taken after the screening procedure just described to show the value in pursuing the matter of detecting interactions was one of introducing cross product terms into a linear model. Unless theory provides a clear guide, one starts with a model which offers some chance of success yet is also relatively manageable. In econometric research, and in some other areas where concerns with interactions arise, interactions are often first introduced by defining them in terms of a cross product of variables. So, if one assumes (as is done below) that 5 socio-economic variables are needed in an equation to explain people's behaviour and outdoor activities, one may write the following equation:

(7) $Y=\mu+\beta_1X_1+\cdots+\beta_5X_5+\beta_{1,2}X_1X_2+\cdots+\beta_{4,5}X_4X_5+\varepsilon$

WHERE e.g., X_1 is the age variable, X_2 is education level, X_3 is a variable giving household size, etc. as indicated in Table 2, ($\beta_{n,n}X^n$ terms are omitted since there is one variable);

β's are the regression coefficients; and

ε is an error term.

In the preceding equation the variables listed can be visualized as interval variables. However, one might think that “education” is ordinal if not nominal. However, if variables are to be multiplied as indicated in the equation, it is important that the multiplication means something. For “gender” with 0-1 and other 0-1 variables there is a meaning but if for occupation there is not even an obvious order related to the activity being considered, occupation cannot be included in such a formulation. What one has gained by having a coefficient for e.g., the age by education interaction is a model with fewer parameters. However, if for an activity participation increase with age to a point and then declines, then neither β_1X_1 or cross product terms with β_1X_1 (e.g.$\beta_{1,2}X_1X_2$) can reflect that “curved” relation. One needs terms in X^n such as X^2. But, even when variables are interval one does not usually know whether the age-education effect can really be modeled by taking a multiple of age and education or even by including powers of variables. The multiplicative interaction terms which appear in equations like Equation 3 are generally a guess at what should appear. The alternative to using powers to allow for “curvature” of effects is to form categories. By looking at effects for the ordered categories, or even fitting the effects, one can often form a realistic picture of how effects change with a variables value (linearly or otherwise). So, without theoretical justification, there is little point in considering powers of the variables, or considering products of the variables including three, four or five variables in a product. Regardless, at the time that analysis was beginning making estimates based on Equation 7 seemed to make sense.

Returning to the main theme, the hope was that when regressions were carried out to determine the unknowns in Equation 7 enough interactions would have been considered. Table 4 shows the results that were obtained. One sees in the right-hand column that there was a nominal increase in the value of R^2 when the interaction terms were included in the model. For example, when one looks at the results for a male model for tent camping, one sees the R^2 was increased from about .16 to .18. The first R^2 is for the five “main effect” (individual variable) parameters (β) being estimated, whereas the second R^2 is for when 15 β were estimated. If the interaction effects *were not important*, the increase in variance explained would be that explained by adding 15 random error terms. In practical terms the introduction of these terms should have explained about 10/(number of cases - 15) which is about .5% of the variance that remained to be explained. But in fact the 2% explained is much in excess of the approximately .5% of the variance that would have been explained by chance. An appropriate F test for the significance of this variance explained is the F test with 10 and infinite number of degrees of freedom: F(10,infinity) approx = (2000/10)(.015/.985) = 3.04 which is significant since it exceeds the .05 level of 2.54.

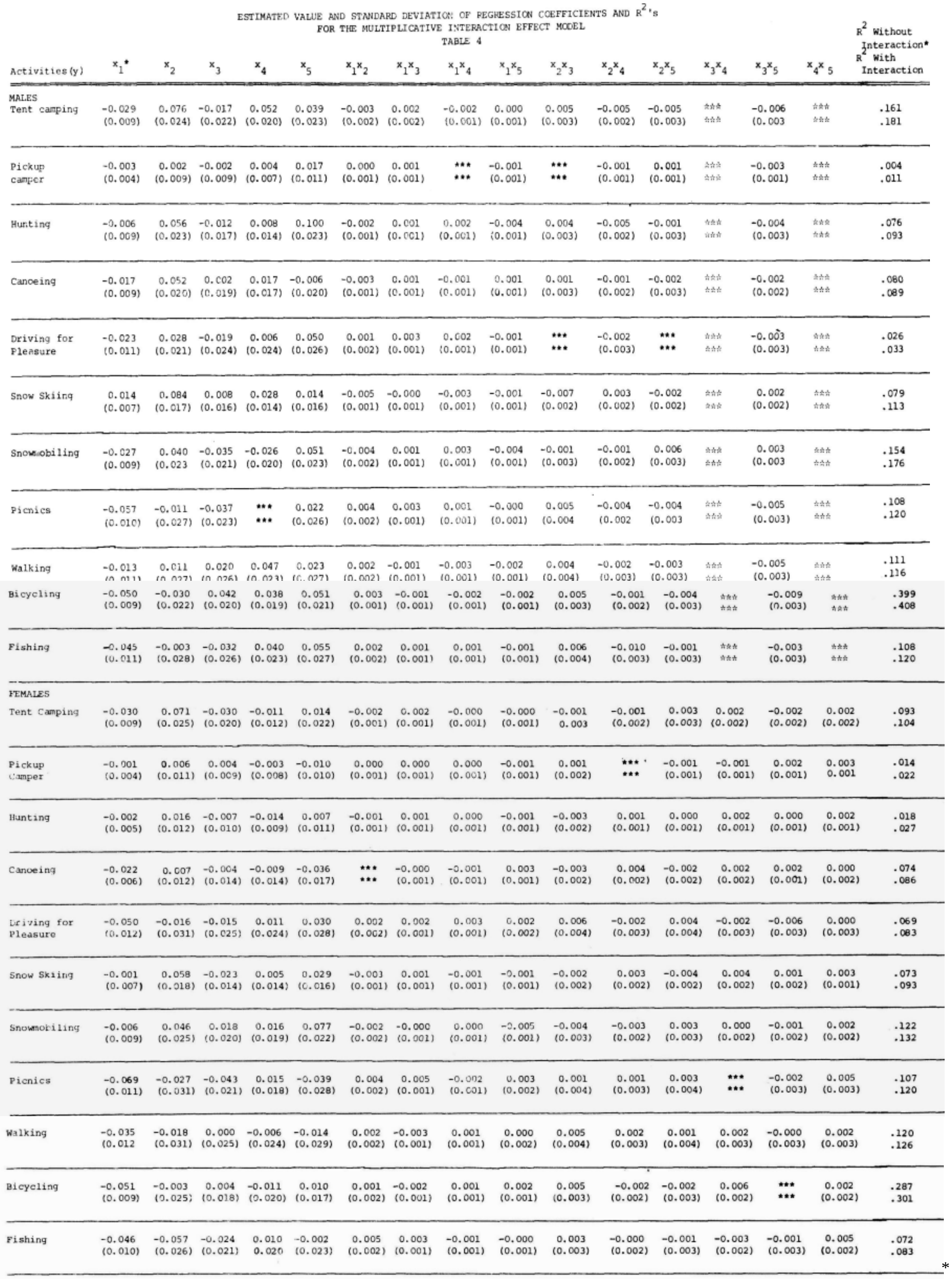
There is really no need to go into this kind of statistical test just introduced to see that the results of introducing the interactions are significant. The very fact that many of the regression coefficients, more than expected by chance, for the product terms are about twice or more of their standard deviation. In aggregate such persistent significance indicates significance at the .05 level (the standard deviations are given in brackets below coefficients). One may notice that the coefficient with value -.00291 for the tent camping model for males is almost twice its standard deviation, which is .00160. Similarly the coefficient for the $\beta_{1,4}$ term, for the $\beta_{2,4}$ term and for the $\beta_{3,5}$ term are also substantial in comparison to their standard deviations. Also, there is the odd coefficient for the interaction terms that has a magnitude more than three times its standard deviation and which will allow one to accept with more confidence. These can be considered to confirm significance based on the “conservative” two times rule that Draper and Smith (1966) have suggested be used in some tests of significance in doing regressions where the distributions are in doubt.

Table 4 shows the parameters of some models that are highly statistically significant improvements over the regression models without interaction terms. However, one may wonder how significant the improvement of an R^2 from .16 to .18 is, compared to what could be achieved. When one compares the results of using the AID program with the "simple" regression results, there may be some surprise that introducing the interaction has explained so little variance. There is certainly no basis for a feeling of elation because the interaction results are statistically significant. This is particularly important in understanding why developing the kind of equations for which coefficients are presented in Table 4 was not pursued. The researchers, who were trying to improve on the simple model, saw that the improved model, though it offered a significant improvement, did not appear to offer the improvement to be expected if the variance that 'was available to be explained by interaction %as being explained. When R^2 was (on the average) being increased by 10 to 20%, getting in the proper interaction terms would increase it by 100%. Something else needed to be done. Or did it?

VALIDATION OF AID RESULTS

If it does not appear that a predictive model is explaining the variance that it should, an obvious first step might appear to be to incorporate more terms into the model. But concerns about doing so have already been raised. Another Line of inquiry is to determine whether, in fact, the model is doing well but that the limit of the R^2 which might be attained has been assessed incorrectly (a low R^2 is fine). The possibility is that the AID program, because of the way it is set up to search for variance, finds variance even if it cannot he explained by a model that is perfectly appropriate to the data. By going to a simulation approach one knows what the true model is because one has used the model to generate observations and one can then determine by how much (if at all) AID indicates an excess of variance explained over what can be expected to be explained by a model that is appropriate to the data.

TABLE 4:



ESTIMATED VALUE AND STANDARD DEVIATION OF REGRESSION COEFFICIENTS AND R^2's FOR THE MULTIPLICATIVE INTERACTION EFFECT MODEL

TABLE 4

Activities (y)	x_1*	x_2	x_3	x_4	x_5	x_1x_2	x_1x_3	x_1x_4	x_1x_5	x_2x_3	x_2x_4	x_2x_5	x_3x_4	x_3x_5	x_4x_5	R^2 Without Interaction* R^2 With Interaction
MALES																
Tent camping	-0.029 (0.009)	0.076 (0.024)	-0.017 (0.022)	0.052 (0.020)	0.039 (0.023)	-0.003 (0.002)	0.002 (0.002)	-0.002 (0.001)	0.000 (0.001)	0.005 (0.003)	-0.005 (0.002)	-0.005 (0.003)	*** ***	-0.006 (0.003	*** ***	.161 .181
Pickup camper	-0.003 (0.004)	0.002 (0.009)	-0.002 (0.009)	0.004 (0.007)	0.017 (0.011)	0.000 (0.001)	0.001 (0.001)	*** ***	-0.001 (0.001)	*** ***	-0.001 (0.001)	0.001 (0.001)	*** ***	-0.003 (0.001)	*** ***	.004 .011
Hunting	-0.006 (0.009)	0.056 (0.023)	-0.012 (0.017)	0.008 (0.014)	0.100 (0.023)	-0.002 (0.001)	0.001 (0.001)	0.002 (0.001)	-0.004 (0.001)	0.004 (0.003)	-0.005 (0.002)	-0.001 (0.003)	*** ***	-0.004 (0.003)	*** ***	.076 .093
Canoeing	-0.017 (0.009)	0.052 (0.020)	0.002 (0.019)	0.017 (0.017)	-0.006 (0.020)	-0.003 (0.001)	0.001 (0.001)	-0.001 (0.001)	0.001 (0.001)	0.001 (0.003)	-0.001 (0.002)	-0.002 (0.003)	*** ***	-0.002 (0.002)	*** ***	.080 .089
Driving for Pleasure	-0.023 (0.011)	0.028 (0.021)	-0.019 (0.024)	0.006 (0.024)	0.050 (0.026)	0.001 (0.002)	0.003 (0.001)	0.002 (0.001)	-0.001 (0.001)	*** ***	-0.002 (0.003)	*** ***	*** ***	-0.003 (0.003)	*** ***	.026 .033
Snow Skiing	0.014 (0.007)	0.084 (0.017)	0.008 (0.016)	0.028 (0.014)	0.014 (0.016)	-0.005 (0.001)	-0.000 (0.001)	-0.003 (0.001)	-0.001 (0.001)	-0.007 (0.002)	0.003 (0.002)	-0.002 (0.002)	*** ***	0.002 (0.002)	*** ***	.079 .113
Snowmobiling	-0.027 (0.009)	0.040 (0.023	-0.035 (0.021)	-0.026 (0.020)	0.051 (0.023)	-0.004 (0.002)	0.001 (0.001)	0.003 (0.001)	-0.004 (0.001)	-0.001 (0.003)	-0.001 (0.002)	0.006 (0.003)	*** ***	0.003 (0.003	*** ***	.154 .176
Picnics	-0.057 (0.010)	-0.011 (0.027)	-0.037 (0.023)	*** ***	0.022 (0.026)	0.004 (0.002)	0.003 (0.001)	0.001 (0.001)	-0.000 (0.001)	0.005 (0.004	-0.004 (0.002	-0.004 (0.003	*** ***	-0.005 (0.003)	*** ***	.108 .120
Walking	-0.013 (0.011)	0.011 (0.027)	0.020 (0.026)	0.047 (0.023)	0.023 (0.027)	0.002 (0.002)	-0.001 (0.001)	-0.003 (0.001)	-0.002 (0.001)	0.004 (0.004)	-0.002 (0.003)	-0.003 (0.003)	*** ***	-0.005 (0.003)	*** ***	.111 .116
Bicycling	-0.050 (0.009)	-0.030 (0.022)	0.042 (0.020)	0.038 (0.019)	0.051 (0.021)	0.003 (0.001)	-0.001 (0.001)	-0.002 (0.001)	-0.002 (0.001)	0.005 (0.003)	-0.001 (0.002)	-0.004 (0.003)	*** ***	-0.009 (0.003)	*** ***	.399 .408
Fishing	-0.045 (0.011)	-0.003 (0.028)	-0.032 (0.026)	0.040 (0.023)	0.055 (0.027)	0.002 (0.002)	0.001 (0.001)	0.001 (0.001)	-0.001 (0.001)	0.006 (0.004)	-0.010 (0.003)	-0.001 (0.003)	*** ***	-0.003 (0.003)	*** ***	.108 .120
FEMALES																
Tent Camping	-0.030 (0.009)	0.071 (0.025)	-0.030 (0.020)	-0.011 (0.012)	0.014 (0.022)	-0.002 (0.001)	0.002 (0.001)	-0.000 (0.001)	-0.000 (0.001)	-0.001 0.003	-0.001 (0.002)	0.003 (0.003)	0.002 (0.002)	-0.002 (0.002)	0.002 (0.002)	.093 .104
Pickup Camper	-0.001 (0.004)	0.006 (0.011)	0.004 (0.009)	-0.003 (0.008)	-0.010 (0.010)	0.000 (0.001)	0.000 (0.001)	0.000 (0.001)	-0.001 (0.001)	0.001 (0.002)	*** ' ***	-0.001 (0.001)	-0.001 (0.001)	0.002 (0.001)	0.003 0.001	.014 .022
Hunting	-0.002 (0.005)	0.016 (0.012)	-0.007 (0.010)	-0.014 (0.009)	0.007 (0.011)	-0.001 (0.001)	0.001 (0.001)	0.000 (0.001)	-0.001 (0.001)	-0.003 (0.002)	0.001 (0.001)	0.000 (0.001)	0.002 (0.001)	0.000 (0.001)	0.002 (0.001)	.018 .027
Canoeing	-0.022 (0.006)	0.007 (0.012)	-0.004 (0.014)	-0.009 (0.014)	-0.036 (0.017)	*** ***	-0.000 (0.001)	-0.001 (0.001)	0.003 (0.001)	-0.003 (0.002)	0.004 (0.002)	-0.002 (0.002)	0.002 (0.002)	0.002 (0.001)	0.000 (0.002)	.074 .086
Driving for Pleasure	-0.050 (0.012)	-0.016 (0.031)	-0.015 (0.025)	0.011 (0.024)	0.030 (0.028)	0.002 (0.002)	0.002 (0.001)	0.003 (0.001)	0.002 (0.002)	0.006 (0.004)	-0.002 (0.003)	0.004 (0.004)	-0.002 (0.003)	-0.006 (0.003)	0.000 (0.003)	.069 .083
Snow Skiing	-0.001 (0.007)	0.058 (0.018)	-0.023 (0.014)	0.005 (0.014)	0.029 (0.016)	-0.003 (0.001)	0.001 (0.001)	-0.001 (0.001)	-0.001 (0.001)	-0.002 (0.002)	0.003 (0.002)	-0.004 (0.002)	0.004 (0.002)	0.001 (0.002)	0.003 (0.001)	.073 .093
Snowmobiling	-0.006 (0.009)	0.046 (0.025)	0.018 (0.020)	0.016 (0.019)	0.077 (0.022)	-0.002 (0.002)	-0.000 (0.001)	0.000 (0.001)	-0.005 (0.001)	-0.004 (0.003)	-0.003 (0.002)	0.003 (0.003)	0.000 (0.002)	-0.001 (0.002)	0.002 (0.002)	.122 .132
Picnics	-0.069 (0.011)	-0.027 (0.031)	-0.043 (0.021)	0.015 (0.018)	-0.039 (0.028)	0.004 (0.002)	0.005 (0.001)	-0.002 (0.001)	0.003 (0.002)	0.001 (0.004)	0.001 (0.003)	0.003 (0.004)	*** ***	-0.002 (0.003)	0.005 (0.003)	.107 .120
Walking	-0.035 (0.012	-0.018 (0.031)	0.000 (0.025)	-0.006 (0.024)	-0.014 (0.029)	0.002 (0.002)	-0.003 (0.001)	0.001 (0.001)	0.000 (0.002)	0.005 (0.004)	0.002 (0.003)	0.001 (0.004)	0.002 (0.003)	-0.000 (0.003)	0.002 (0.003)	.120 .126
Bicycling	-0.051 (0.009)	-0.003 (0.025)	0.004 (0.018)	-0.011 (0.020)	0.010 (0.017)	0.001 (0.002)	-0.002 (0.001)	0.001 (0.001)	0.002 (0.001)	0.005 (0.003)	-0.002 (0.002)	-0.002 (0.003)	0.006 (0.002)	*** ***	0.002 (0.002)	.287 .301
Fishing	-0.046 (0.010)	-0.057 (0.026)	-0.024 (0.021)	0.010 0.020	-0.002 (0.023)	0.005 (0.002)	0.003 (0.001)	-0.001 (0.001)	-0.000 (0.001)	0.003 (0.003)	-0.000 (0.002)	-0.001 (0.003)	-0.003 (0.002)	-0.001 (0.003)	0.005 (0.002)	.072 .083

*

*** Indicates that this variable was not considered in the regression because its explanatory power was too small.

It was decided to generate a dependent variable Y having 0 and 1 values indicating participation or non-participation. In the simulation, five variables with four levels of each variable were defined. Values of Y around a grand mean of one-half were generated for 1500 cases. The formula for regression coefficients was:

$\beta(i,j) = (1/4)((i - 2.5)/1.5)/2[1-j]$

WHERE j indicates the variables 1 to 5 and i indicates the level of the variable that an individual has, 1 to 4;

$\beta(i,j)$ are the coefficients in: $E(y) = U + \beta(j,i)X(j,i)$ (as noted U was taken to be ½)

A random number routine was used to independently and randomly generate the levels of the 5 independent variables that characterize an observation. For example, (1, 3, 4, 1, 2) could define a person for whom an observation was made. For this person:

$E(y) = (1/2)+(1/4)[(-1)+(1/3)(1/2)+(1)(1/4) + (-1)(1/8)+(-1/3)(1/16)]$

In the above one has (1/3)(1/2) as what could be described as the third term in E(y) because the person has level 3 of variable 2. The (1/4) which is in each $\beta(i,j)$ appears as a factor that multiplies all five $\beta(i,j)^{'}$s. An observed Y was generated using random numbers so an observation 1 had a probability of E(y) and 0 a probability of 1-E(y). In generating collections of levels for variables, e.g. (1,3,4,1,2), it was considered that people were in levels 1 to 4 of each variable in the ratios 4/3/2/1 so that 4 times as many people were assigned to level 1 of a variable as to level 4. The sum 4+3+2+1=10 so cumulatively, level 4 is associated with 0 to .1, 2 with .1+ to .3, etc. Therefore, using a random number routine, for a variable X(i), if the random number generated was under l/10 a person was assigned to level 4 of the variable X(i), if not level 4 but up to .2 the level was set to 3, etc.

The results of the simulation study are shown in Table 5 where the ANOVA figures are calculated on the basis of theory (because there was no need to estimate these results). The results for AID analysis are the average results for 100 analysis runs. As can be seen the AID model when applied to the given data to which another model is structurally appropriate explains only slightly more variation than the model which is actually appropriate to the data, the difference in explanatory power only being noticeable in the third figure of R^2. The difference in R^2 is truly minimal and certainly much less than the difference between the ANOVA model and AID results reported in Table 1.

Thus the difference in the R^2 suggested by an AID run and the R^2 found using regression models should not be large if the regression models are truly appropriate to the data. So, it can be concluded that there is good evidence that for models very similar to those developed using the CORD Study data, the AID program detects relatively large sums of squares which almost certainly do not relate to spurious interactions. It also appears safe to say that the results provide a clear indication that there is a great deal more variance to be explained in the CORD Study data than was explained by using the model with interaction terms for which results are presented in Table 4. Introducing the interaction terms only explained about 20% of the variance that should be explained if appropriate interaction terms had been considered. If the models had been good, R^2 should have gone up 100% on the average, not just by 20% as was the case. There may be many more interactions to be considered and/or the interactions may be a different type from those which are implicit in the formulation that was used.

TABLE 5: SUMMARY RESULTS OF AID AND ANOVA ANALYSES OF SIMULATED DATA TO WHICH AN ANOVA MODEL IS STRUCTURALLY APPROPRIATE

	MODEL	
	AID	ANOVA
(1) Total sum of squares	336.2792	336.2792
(2) Between sum of squares	53.3586	56.8900
(3) Within sum of squares	276.9206	279.2892
Mean of y	0.32733	0.3393
S.D. of R^2=(2)/(1)	0.170	0.168

The validation of AID results has only brought one back to the point of seeing that little was gained by cross product analysis but that much must be achieved if models are to adequately explain the relation between socio-economic variables and participation.

DISCUSSION

The commentary above in some sense presents a logical sequence which has occurred in considering how models should be developed that may be used to explain people’s participation in activities in terms of their socio-economic characteristics. However, one very important practical question remains. When the logical sequence has been built up it shows that interaction terms should be considered in developing the kinds of models of which applications have beer introduced in TN 12 and 13. Therefore, how simple should models be that are applied in the way indicated in the TN 12? For example, should estimates be made for sub-areas of Canada based on National or Provincial data using relatively simple models or is this a dangerous practice? If an area of Canada is very similar to the nation as a whole then one can see why one could use a model which is deficient in certain respects. Even if interactions are not considered, predictions could be close. However, when one recognizes the disparities in Canada in terms of what activities can be carried out, what differentials there are in terms of age, income etc. then the dangers inherent in using parameters for a National ,model in making predictions for a sub-area of the country are obvious. The area for which predictions are made can be such that the National parameters are not relevant. One must be very concerned that the National parameters are only aggregate parameters with no particular relevance to any sub-areas that deviate substantially from the national average.

In the context of this paper the crux of the concern is not whether there are disparities within Canada but whether the effects of these disparities on peoples' participations in activities can be adequately modeled only considering first order effects. Should second order effects be considered because these explain regional differences? Are higher order effects important? We do not know and need to know if models are to be used with confidence.

One is confronted with the fact that the simple analysis of variance model appears to only “tap” part the variance that should be explained by socio-economic characteristics. As of the mid 1970s it is impossible to say how much this deficiency of a model influences predictions by resulting in bias. Actually, there is one area of model deficiency on which comment can be made. In TN 20 supply factors are derived that show that, for at least some activities, regression equations should include not only socio-economic variables, but a measure of supply in the various areas in which people live for whom predictions are made. These supply factors can be visualized by:

Probability of participating	=	A function of demographics and e.g. activity group membership	+	A supply factor effect relevant to the activity group of concern	+ Error	

Now, even though supply factors may only account for 1/5 the variance that socio-economic characteristics do, one need only look at TN 25 to see that (for example for skiing) the supply factors for Alberta and B.C. are very important in making correct predictions of participation.

The problem with supply effects is that they should be considered but they may not be known or may only be known inaccurately. As indicated in TN 20 massive amounts of data are required to estimate supply effects from participation data and there is no known way, as of the 1970s, to calculate them based on inventory information on what facili‘ies there are. Developing formulae for computing supply factors for “activity groups” based on "resource inventory" data would appear to be very important if good use is to be made of ANOVA models. Of course, the implicit challenge is understanding substitutability and its relation to supply and user groups so interaction between activities and their supply is properly modeled (see specifically Ch. 5 and 6).

On another matter, comments on AID have ignored an important point. That is that there are kinds of interactions which it is convenient to consider and there are kinds which it is extremely inconvenient to consider. The reader may well ask why there is not a proposal to forget about using regression models to make predictions (Cesario gives an example of using an AID model to make predictions in TN 4). The reason not to use AID models in making predictions is that they require detailed multivariate information. Such information is may be available using Canada Census micro data files or even by special tabulation. Regardless, if models are seriously deficient, modelling can produce results that are better than those from a local survey. But if you do not have any idea if results are really bad, presumably you go with a survey if you can afford that. In this context AID predictions being slightly more accurate than ANOVA predictions in terms of their variance, is of no value if the inability to consider supply effects results in serious error in using AID models. The big problem for prediction, given the 1970 state of the art is validity, not reliability.

Even if it is not desirable to use AID to make predictions, one may not see what the problem is in considering any arbitrary interaction between 2, 3 or more variables. The problem is that if there is an age-education interaction effect for males, one must be able to specify how many males there are in a specific age-education groups to introduce this interaction into estimation. This may not be difficult in some cases. Census data may be used. However, as of the 70s, getting the information on males by education for small areas of Canada to define a trend was not necessarily trivial. When other variables are considered on which data are collected on a sample basis (collected using the long census questionnaire that is only administered to a sample of Canadian Residents even in a Census), the problem is compounded because data for small areas obtained by special requests for tabulations may be costly, variable and present other problems. If one is to make projections of the number of males in certain age-education groups for 20 years in the future, one must consider the consequence of introducing interactions into a model if they are ones for which one cannot make reasonable projections of the relevant $n_{i,j}$, the “net” subpopulation to the $\beta_{i,j}$ applies (re “net” see TN 6).

The preceding paragraph raises an issue taken up in TN 6. If the accuracy/reliability of a model's parameters is not all that is of concern in using a model and if one is concerned about both the accuracy of the $\beta_{i,j}$'s and of the $n_{i,j}$, the number of people in certain socio-economic groups, one should not concentrate on the $\beta_{i,j}$ and problems with interactions if, in relative terms, there is large inaccuracy is in the $n_{i,j}$'s which can actually be corrected.

When it comes to the matter of modelling using a small sample, one should look at the results presented in TN 6 and recognize that unless sample sizes are in the order of 4,000 or larger then predictions made using the regression results are going to be extremely inaccurate. If some kind of statement is to be made about participation by people in a small area in a certain activity, a reasonable choice may be to use a telephone survey or some other means of obtaining information quickly rather than making predictions using modelling results. Given all the objections that can be raised to telephone surveys, etc. little is gained by replacing the results of such work with results produced using a theoretical model when it can be shown that these results have errors which are probably far greater than any errors that arise in a well planned telephone survey.

Turning to a quite different and less practical matter, an analyst often wants to use regression results to draw some kind of conclusions about what is happening in the world or in the universe that he is considering. The failure to introduce interaction terms into a model when, in fact, they relate to about 50% of the variance that could be explained by the model can certainly be expected to distort the picture of reality that an analyst infers.

In closing this discussion one should note, as indicated in other TN, if one is calculating people's expected probability of participation in an activity then the very fact that probabilities are being estimated suggests that each individual observation has a unique variability associated with it. This heteroscedasticity problem, which is encountered in dealing with dependent variables which cannot be accepted as having a constant variance, is the topic of concern in the Cicchetti and Smith paper included as an appendix to this volume, and there is also useful commentary in the review of Chapter VII.

CONCLUSION

This article has presented some rather distressing findings about the structure of models commonly used for predicting participation and frequency of participation in outdoor activities. It is clear that interaction effects play an important role in explaining peoples' participation in outdoor activities. Neglecting such factors could result in errors arising which would mean that estimates made have substantial biases. However, as pointed out, there is no evidence as to whether (once supply factors are taken into account) biases tend to be very small because the people to whom interaction effects apply tend to be very homogeneously distributed among the population. There has been no research which shows whether or not there are some sectors of the population for which interaction effects are extremely important and others for which a simple model would be quite appropriate. Until such research has been carried out to clarify this matter it must be recognized that there are dangers in making predictions using regression models.