TN 12: ANALYSIS OF VARIANCE AS A TOOL FOR ESTIMATING PARTICIPATION IN OUTDOOR RECREATION ACTIVITIES

By S. Rousseau, J. Beaman, M. Renoux and J. Hendry

ABSTRACT

The use of analysis of variance, ANOVA*, in arriving at an understanding of participation in outdoor recreation is probably best known from its use in the Mueller and Gurin volume of the ORRRC report (1961). In that volume there is discussion of how the estimated effects of socio-economic characteristics on people's participation provide insight into recreational behaviour that could not have been obtained by merely tabulating data.

This paper goes beyond that and deals with applying analysis of variance to data pertaining to a certain recreation activity to estimate participation in that activity by people in a given geographic area, conditional on the socio-economic characteristics of these people.

Specific formulae for making estimates are given in the paper and their use is illustrated by making predictions of the number of hunters and total hunting trips by Quebec residents.

The paper concludes with a discussion pointing out some difficulties encountered in using the methods of estimation described. The reasons for having alternative methods for estimating total volume of activity (total hunting trips) is an important topic taken up in this section of the paper.

There are references in the paper to a number of papers in which further results such as (1) when such a model should be used. (2) accuracy of results, (3) structural problems with the models derived, and (4) the value of R^2 should have, have been presented.

*NOTE: **In CORDS TN using ANOVA does not refer to running a program that "partitions" variance based on the assumption that data were collected according to a designed experiment. In the terminology of 2006, one is referring to using multiple regression to *analyze the variance in a dependent variable given 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

As early as 1961 there was a paper produced which presented the results of analysis of variance, ANOVA, on how having different levels of income or belonging to a particular socio-economic category influenced a person's participation in outdoor activities. In that study, Mueller and Gurin (1961) went so far as to recognize that the same model of how participation related to socio-economic variables was not appropriate for both males and females because of what are known as interaction effects.

Other work has influenced the production of the models presented here. From within the CORD Study, one influence was the proposal by Hendry (1970) that CORD Study National Survey data should be processed by a dummy variable analysis; this is the economists way of saying that a variant of the kind of analysis described here should be undertaken. Knetsch also made a proposal that CORD Study national survey analysis should follow a strategy that was laid out by Chiccetti, Seneca and Davidson (1969; see Knetsch's commentary in Ch. 1). In other words, in the history of recreation research and in the history of the CORD Study there are suggestions that a model that may be expressed in words as follows should be used in analyzing

people's outdoor recreation participation:

Equation 1 (Form 1):

Probability of participation or frequency of participation for a person with socio-economic characteristics	=	a general participation level	+	Effect of being in a city, in a town or in in country	+	the effect of being a member of e.g a certain size household	+ etc. +error

With Y(i,J,K,L, ...) being the dependent variable, in mathematical terms the equation is:

Equation 1 (Form 2):

$Y(i,J,K,L, ...) = U+B(1,J)+B(2,K)+B(3,L)+...+\varepsilon(i)$

WHERE Y(.) is 0 or 1 for participation or nonparticipation or for a frequency model is the actual number of times that a person participated; i, J, K, L, . . ., the sub scripts of Y(.) give information about the person i who has level J of a first socio-economic variable (e.g. in Figure 1 comes from a household of some size), who has level L of a third socio-economic variable (e.g. education level), etc.

U is a general level that applies to all persons (from Table 1 for participation by male hunters for 1972 it is .234);

B(1,J) is the effect on Y(.) of having level J of socio-economic variables 1 (under 1972^2 of Table 1, B(1,2) = .006);

B(2,K) is the effect on Y of having level K of socio-economic variables 2 (e.g. in Table 1 for household size for 1972 #(2) the effect is -.002);

B(3,L) is the effect on Y of having level L of socioeconomic variable 3;

B(.)'s with first subscripts up to 9 would be necessary to define all the effects shown in Figure 1 and given in

ε(i) is an error term that has a value equal to the difference between the observed Y for person i and his predicted i (this is illustrated subsequently).

Having referred to Table 1 and Figure 1 it seems appropriate to give some general explanation about these and the related Tables 2 through 4 and Figure 2. These are from a larger document that was originally to be an appendix to TN 12, to be Rousseau's "appendix". The Table of Contents of this larger document, which was prepared by Rousseau, one of the authors of this paper, was an attachment of TN 12 in the 1976 CORDS Volume 2 draft publication but is omitted in this version since the Appendix document is no longer available. The appendix had tables and figures like those here for all activities available in the data analyzed.

That figures and tables such as appear here were prepared for all activities may lead some readers to ask what practical value there is in information like that in Tables 1 to 4? To see specifically what the coefficients indicate, an example is useful. Assume one wants to predict the probability of being a hunter in 1969 for a person who in 1969 was (1) male, (2) from a city of over 100,000, (3) married, (4) from a family of size three or four, (5) with some high school education, (6) in the age group 30 to 39, (7) with an income of $6,000 to $10,499 in 1969, (8) the head of the household, and (9) with no children under 5 at his home (10) which was a single family dwelling. To determine this probability based on the 1969 survey, one simply takes the general mean for males of .219 given in Table 1 and adds to it the relevant increments or beta

values from the first column, the column under 1969 of Table 1, to obtain the equation below, which indicates that the probability of participating is .219 - .091 + .011 + .016 + .011 + .072 + .031 - .01 - 016 + .012 = .255

The coefficients in Table 3 can be used in the same way to get a prediction of the number of times that a similar male can be expected to go hunting.

The reason for presenting the results of the analysis graphically is that the large array of numbers in Tables 1 to 4 does not give a quick impression of how probability or frequency of behaviour relates to attributes. The graphs of the regression coefficients show that (for hunting participation) there are fairly distinct trends both for males and females. Negative differentials show the low probability of hunting, for people in large cities. One also sees the shift to positive differentials for people from small communities and rural areas. With age, once one passes the age 16 at which hunting is legal, one notes that there is a decrease in hunting participation with increasing age. This decrease in participation in an activity with increasing age is something that almost all activities have in common and which, to some extent, reflects a general tendency of most people to become less active when they become older. By looking at the coefficients on the relationship between participation and numbers of persons in the household, one sees that there are not strong effects compared to the age effects or city size effects.

Tables 2 and 4 present standard deviations values for the regression coefficients that are plotted in the graphs and which are reported in Tables 1 and 3. These standard deviations give one an idea of the size of deviation between the "true" coefficient and its estimate. “True” is used because it refers to the value that the coefficient takes (1) even if answers are not accurate but (2) it is assumed the model is structurally sound (the right one to accurately predict responses). If one can assume that the distribution of estimates of the coefficients is somewhat near normal, then the characteristics of the normal distribution suggest that deviations of about 1.56 times the standard deviations reported have a relatively low probability of occurring (less than 1 chance in 10). So for example, one may note that the hunting coefficient showing the decrease in probability of a - .091 that was observed for being in a city of 100,000 in 1969 has a standard deviation of .02. This means that there is a high probability that this coefficient could be as large as .11 or as small as .071 but, in line with the point just made, the probability that it will be greater than .12 or less than .06 is quite remote. In the case of the coefficients that apply to the number of persons in the household, one may observe that all of these have an absolute value of under .02 whereas all of the standard deviations for these coefficients are over .02; this is a clear indication that one can accept the hypothesis that all of these coefficients equal zero: in other words that persons in the household need not be a variable considered in predicting hunting participation.

Given the previous statements, it should be noted that the coefficients were not computed in the most efficient way possible and therefore that the variance estimates may be larger than they “should” be. The point is that there are a number rather tricky statistical issues related to the need to carry out special weighted regressions to correct for heteroscedasticity (in another context see TN 19). For example, when one is concerned with a respondent’s probability of participating in an activity, variability ($\varepsilon(i)$ of Equation 1) depends on the individual’s probability of participation. For efficient estimation this dependency should be considered (e.g. see Scheffe 1959, pp. 19-21). As indicated in the Review of Chapter VII, carrying out these special regressions would have resulted in additional expense that was unnecessary because with the number of observations available. Based on some simulation done, with thousands of responses weighted regression apparently does not produce parameters that are perceptibly more accurate

than those produced by an unweighted analyses (see Smith and Cicchetti's work in Appendix A).

ESTIMATED PARAMETER VALUES FOR HUNTING PARTICIPATION

General means for Males		General means for Females	
1969	.219	1969	.035
1972(1)	.216	1972(1)	.017
1972(2)	.234	1972(2)	.022

		EFFECTS FOR MALES			EFFECTS FOR FEMALES		
Beta No.	Labels	1969	1972(1)	1972(2)	1969	1972(1)	1972(2)
CITY SIZE							
B (1,1)	Over 100,000	-.091	-.126	-.144	-.015	-.026	-.030
B (1,2)	30,000-100,000	-.022	.018	.006	-.026	.006	-.002
B (1,3)	10,000-30,000	.012	.090	.077	.022	-.006	.005
B (1,4)	1000-10,000	-.005	-.036	-.012	.005	.011	.013
B (1,5)	Rural	.106	.054	.074	.015	.014	.014
PERSONS IN HOUSEHOLD							
B (2,1)	One	.011	-.000	.008	.007	-.007	-.003
B (2,2)	Two	-.019	-.012	-.002	-.002	-.006	-.003
B (2,3)	Three or Four	.011	.025	.022	.004	-.011	-.006
B (2,4)	Five	-.009	-.025	-.029	-.008	.006	-.002
B (2,5)	Six or More	.006	.012	.001	-.002	.018	.014
EDUCATION							
B (3,1)	Public School, Refuse	.001	-.024	-.055	-.008	-.006	-.004
B (3,2)	Some High School	.016	.043	.064	.008	.007	.006
B (3,3)	High School Grad. or More	-.017	-.019	-.009	-.0002	-.002	-.003
AGE							
B (4,1)	10 to 17	-	-	.035	-	-	-.000
B (4,2)	18 to 29	.070	.067	.049	.034	.029	.027
B (4,3)	30 to 39	.011	.026	.014	-.012	-.013	-.011
B (4,4)	40 and Over	-.081	-.092	-.098	-.022	-.016	-.015
INCOME							
B (5,1)	Refuse, Don't Know	-.024	-.022	-.042	-.021	-.019	-.014
B (5,2)	Less than $2,999	-.059	-.043	-.033	-.006	.008	.006
B (5,3)	$3,000 to $5,999	-.006	-.006	.005	.005	-.017	-.020
B (5,4)	$6,000 to $10,499	.072	.031	.038	.005	.004	.006
B (5,5)	$10,500 or More	.016	.039	.032	.017	.023	.021
MARITAL STATUS							
B (6,1)	Single	.007	.010	.010	-.003	.014	.015
B (6,2)	Married	.031	-.007	-.008	.001	-.012	-.013
B (6,3)	Other	-.038	-.004	-.002	.002	-.002	-.002
POSITION IN HOUSEHOLD							
B (7,1)	Head (Male or Female)	-.010	.034	.030	.003	.030	.028
B (7,2)	Son or Daughter	.012	.010	.009	.0004	-.002	.001
B (7,3)	Other	-.002	-.043	-.039	-.004	-.028	-.029
CHILDREN UNDER 5							
B (8,1)	None	-.016	.004	-.002	-.0005	.007	.004
B (8,2)	Some	.016	-.004	.002	.0005	-.007	-.004
HOUSING							
B (9,1)	Single	.012			-.011		
B (9,2)	Row, Duplex, Multi	.007			.012		
B (9,3)	Apartments	-.019			-.001		

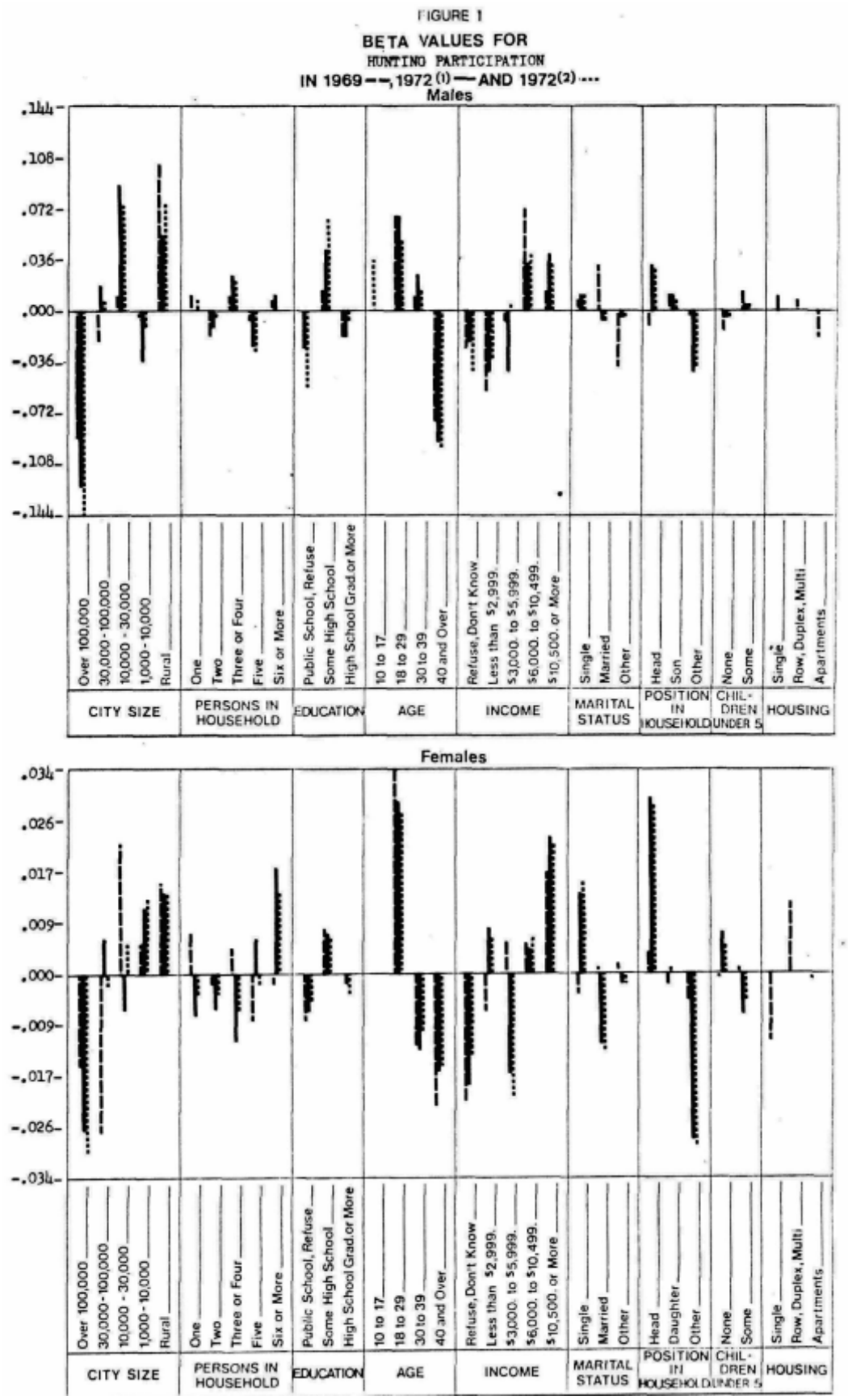
(1) 18 years of age and over

(2) 10 years of age and older

* The 1969 and 1972 Frequency figures are not comparable because frequency catagories

FIGURE 1

BETA VALUES FOR
HUNTING PARTICIPATION
IN 1969 – –, 1972 (1) — AND 1972 (2) ····



In fact, by the beginning of the second millennium one would not carry out analysis as in the 70s. One can easily and cheaply run Goldberger (1966) "type" weighted regression to correct for unequal variability (heteroscedasticity). Weighting can be used to achieve the desirable property of "preventing" estimating negative probabilities (since observations are positive large weight are used to force prediction close enough to small values so they remain positive). In the second millennium one might use a logistic model (e.g. see SAS or SPSS manuals) to estimate probabilities, if only to avoid writing code to get a Goldberger type solution. Furthermore, analysis showed that the CORDS models developed had structural problems. Results presented in TN 20 show that the kinds of simple models presented here very often explain only about half the variance that should be explained by socio-economic characteristics *if the structure of the model was correct*. The structural problem (as e.g. confirmed by analyses allowing for more complicated structure – see TN 20) was that the simple autonomous effects of socio-economic variables assumed to apply were not complicated enough to mirror reality. When the structure of a model is not correct, probabilities calculated are not correct so using those probabilities to "correct" for heteroscedasticity can result in a poorer weighting than not correcting at all. Regardless, the ANOVA method presented can be used to making predictions. The ideas presented apply when a more complicated model is estimated. Therefore, there is merit in presenting work done and noting problems with it. Doing that does not diminish the value of and need for developing structurally more adequate models.

Model structure is an important matter that needs attention. The reader may note that a hunting model is tested in Technical Note 6 and is accepted to be structurally appropriate to the 1972 data for male residents of Canada's participation in the activity hunting obtained in the 1972 National Survey of Canadians' participation in outdoor activities. In that same note it is explained why results derived in TN 29 show that the hunting model derived here is not good because the effect of supply on participation is not considered. In the next section of this paper the rationale behind using an equation, such as the one introduced in making predictions, is presented. The accuracy, really reliability since response may not accurately reflect behaviour, of the predictions that can be made using such a model is taken up in a separate paper (TN 6). As already implied, structural problems with equations (in terms of using them in modelling people's behaviour) are taken up in several other papers. The matter of whether or not interaction effects exist and their magnitude is the topic of TN 20. TN 29 deals with the matter of whether the effect of supply can and/or should be incorporated into Equation 1. TN 35 presents the results of research on what the value of R^2 should have when regression is carried out on survey data to estimate the parameters of Equation 1 (form 2) and how one can test for model validity. (See also the Review of Chapter 7 of this volume).

In the way of further introduction, the reader may find it interesting to note that the present version of TN 12 is not the original version which was released. A great deal of work on applying the analysis of variance model for making predications has been carried out by Renoux, (1973, 1975). The CORD Study research, cited above has occurred since the original TN 12 was prepared. This research has resulted in the recognition of a number of practical problems that are encountered in making computations which yield predictions and in the recognition of a number of theoretical problems. So a re-examination of the material resulted in the preparation of this present note and the revision of some other notes.

TABLE 2

STANDARD DEVIATIONS OF

ESTIMATED PARAMETER VALUES FOR HUNTING PARTICIPATION

General means for Males		General means for Females	
1969	.038	1969	.013
1972[1]	.030	1972[1]	.013
1972[2]	.028	1972[2]	.012

		EFFECTS FOR MALES			EFFECTS FOR FEMALES		
Beta No.	Labels	1969	1972[1]	1972[2]	1969	1972[1]	1972[2]
CITY SIZE							
B (1,1)	Over 100,000	.020	.019	.016	.008	.008	.007
B (1,2)	30,000-100,000	.033	.030	.025	.012	.012	.012
B (1,3)	10,000-30,000	.036	.034	.030	.015	.016	.014
B (1,4)	1,000-10,000	.029	.029	.024	.011	.012	.010
B (1,5)	Rural	.023	.021	.018	.009	.009	.008
PERSONS IN HOUSEHOLD							
B (2,1)	One	.054	.043	.044	.017	.016	.016
B (2,2)	Two	.025	.022	.021	.009	.009	.009
B (2,3)	Three or Four	.022	.020	.018	.008	.008	.007
B (2,4)	Five	.030	.029	.023	.011	.011	.010
B (2,5)	Six or More	.028	.027	.022	.010	.011	.010
EDUCATION							
B (3,1)	Public School, Refuse	.017	.017	.014	.007	.007	.006
B (3,2)	Some High School	.016	.015	.013	.006	.006	.006
B (3,3)	High School Grad or More	.017	.016	.015	.006	.006	.006
AGE							
B (4,1)	18 to 19	-	-	.027	-	-	.014
B (4,2)	19 to 29	.023	.021	.019	.008	.008	.008
B (4,3)	30 to 39	.021	.021	.026	.007	.008	.010
B (4,4)	40 and Over	.020	.019	.022	.007	.007	.009
INCOME							
B (5,1)	Refuse, Don't Know	.060	.033	.027	.021	.012	.011
B (5,2)	Less than $2,999	.031	.026	.025	.012	.011	.011
B (5,3)	$3,000 to $5,999	.024	.022	.019	.009	.009	.008
B (5,4)	$6,000 to $10,499	.023	.018	.015	.009	.008	.007
B (5,5)	$10,500 or More	.033	.024	.020	.012	.010	.009
MARITAL STATUS							
B (6,1)	Single	.035	.028	.029	.012	.013	.013
B (6,2)	Married	.035	.028	.028	.010	.010	.011
B (6,3)	Other	.041	.033	.034	.011	.010	.010
POSITION IN HOUSEHOLD							
B (7,1)	Head (Male or Female)	.041	.033	.034	.014	.015	.016
B (7,2)	Son or Daughter	.034	.030	.029	.015	.016	.014
B (7,3)	Other	.038	.037	.036	.014	.018	.016
CHILDREN UNDER 5							
B (8,1)	None	.017	.017	.014	.006	.006	.006
B (8,2)	Some	.017	.017	.014	.006	.006	.006
HOUSING							
B (9,1)	Single	.019			.007		
B (9,2)	Row, Duplex, Multi	.023			.008		
B (9,3)	Apartments	.026			.010		

(1) 18 years of age and over

(2) 10 years of age and older

* The 1969 and 1972 frequency figures are not comparable because frequency categories are not identical for both years.

TABLE 3

ESTIMATED PARAMETER VALUES FOR HUNTING FREQUENCY*

General means for Males		General means for Females	
1969	1.041	1969	.096
1972[1]	1.736	1972[1]	.093
1972[2]	1.861	1972[2]	.084

		EFFECTS FOR MALES			EFFECTS FOR FEMALES		
Beta No.	Labels	1969	1972[1]	1972[2]	1969	1972[1]	1972[2]
CITY SIZE							
B(1,1)	Over 100,000	-.406	-.978	-1.282	-.063	-.133	-.120
B(1,2)	30,000-100,000	-.276	-.065	-.014	-.086	.028	.014
B(1,3)	10,000-30,000	-.023	1.039	.689	.092	.114	.093
B(1,4)	1000-10,000	.010	-.107	.119	.016	.005	.010
B(1,5)	Rural	.695	.112	.488	.042	-.014	.004
PERSONS IN HOUSEHOLD							
B(2,1)	One	.285	-.164	.012	.032	-.051	-.036
B(2,2)	Two	-.051	-.300	-.116	-.006	-.016	-.005
B(2,3)	Three or Four	-.075	-.064	.093	.024	-.050	-.027
B(2,4)	Five	-.150	.339	.075	-.019	.055	.043
B(2,5)	Six or More	-.009	.190	-.065	-.032	-.062	.024
EDUCATION							
B(3,1)	Public School, Refuse	-.033	-.167	-.507	.003	-.018	.008
B(3,2)	Some High School	.198	.272	.487	.020	.040	.023
B(3,3)	High School Grad or More	-.166	-.105	.020	-.024	-.021	-.031
AGE							
B(4,1)	10 to 17	--	--	.457	--	--	-.014
B(4,2)	18 to 29	.682	.453	.207	.107	.124	.125
B(4,3)	30 to 39	-.227	.323	.172	-.023	-.054	-.041
B(4,4)	40 and Over	-.455	-.776	-.836	-.085	-.070	-.069
INCOME							
B(5,1)	Refuse, Don't Know	.266	.039	-.357	-.058	-.026	-.034
B(5,2)	Less than $2,999	-.370	-.276	-.154	-.029	.006	-.009
B(5,3)	$3,000 to $5,999	-.067	-.073	.096	-.002	-.040	-.049
B(5,4)	$6,000 to $10,499	.080	.052	.259	.046	-.008	.009
B(5,5)	$10,500 or More	.092	.258	.156	.044	.069	.084
MARITAL STATUS							
B(6,1)	Single	.027	-.203	-.203	-.011	.067	.067
B(6,2)	Married	.275	.168	.129	.001	-.048	-.049
B(6,3)	Other	-.302	.035	.074	.009	-.019	-.018
POSITION IN HOUSEHOLD							
B(7,1)	Head (Male or Female)	-.022	-.035	-.056	.023	.120	.110
B(7,2)	Son or Daughter	.142	-.030	.078	-.013	-.026	-.012
B(7,3)	Other	-.119	.065	-.022	-.010	-.093	-.098
CHILDREN UNDER 5							
B(8,1)	None	-.124	.143	.046	-.001	-.001	-.002
B(8,2)	Some	.124	-.143	-.046	.001	.001	.002
HOUSING							
B(9,1)	Single	.129			-.017		
B(9,2)	Row, Duplex, Multi	.014			.028		
B(9,3)	Apartments	-.143			-.011		

(1) 18 years of age and over

(2) 10 years of age and older

* The 1969 and 1972 frequency figures are not comparable because frequency categories are not identical for both years.

TABLE 4

STANDARD DEVIATIONS OF
ESTIMATED PARAMETER VALUES FOR HUNTING FREQUENCY*

General means for Males		General means for Females	
1969	.253	1969	.048
1972(1)	.287	1972(1)	.061
1972(2)	.278	1972(2)	.055

		EFFECTS FOR MALES			EFFECTS FOR FEMALES		
Beta No	Labels	1969	1972(1)	1972(2)	1969	1972(1)	1972(2)
CITY SIZE							
B (1,1)	Over 100,000	.131	.177	.157	.027	.035	.032
B (1,2)	30,000 - 100,000	.218	.283	.253	.042	.055	.053
B (1,3)	10,000 - [illegible]	.238	.322	.297	.053	.071	.065
B (1,4)	1,000 - 10,000	.194	.272	.243	.039	.053	.047
B (1,5)	Rural	.155	.202	.180	.031	.041	.037
PERSONS IN HOUSEHOLD							
B (2,1)	One	.360	.409	.437	.061	.073	.073
B (2,2)	Two	.166	.207	.214	.032	.042	.042
B (2,3)	Three or Four	.143	.188	.180	.027	.036	.034
B (2,4)	Five	.197	.273	.233	.040	.050	.044
B (2,5)	Six or More	.182	.256	.220	.036	.049	.043
EDUCATION							
B (3,1)	Public School, Refuse	.115	.166	.140	.023	.033	.029
B (3,2)	Some High School	.106	.144	.127	.021	.028	.025
B (3,3)	High School Grad or More	.114	.151	.154	.023	.029	.029
AGE							
B (4,1)	10 to 17	-	-	.274	-	-	.065
B (4,2)	18 to 29	.149	.200	.190	.029	.037	.036
B (4,3)	30 to 39	.142	.203	.256	.026	.036	.045
B (4,4)	40 and Over	.134	.184	.223	.026	.033	.042
INCOME							
B (5,1)	Refuse, Don't Know	.397	.312	.264	.075	.056	.050
B (5,2)	Less than $2,999	.205	.250	.244	.042	.052	.049
B (5,3)	$3,000 to $5,999	.157	.205	.186	.031	.039	.036
B (5,4)	$6,000 to $10,499	.152	.174	.152	.031	.034	.031
B (5,5)	$10,500 or More	.216	.224	.198	.044	.046	.041
MARITAL STATUS							
B (6,1)	Single	.232	.270	.287	.044	.058	.058
B (6,2)	Married	.234	.265	.284	.035	.048	.048
B (6,3)	Other	.270	.318	.341	.038	.046	.046
POSITION IN HOUSEHOLD							
B (7,1)	Head (Male or Female)	.269	.319	.339	.049	.070	.070
B (7,2)	Son or Daughter	.224	.287	.290	.053	.072	.065
B (7,3)	Other	.250	.356	.360	.052	.081	.074
CHILDREN UNDER 5							
B (8,1)	None	.112	.163	.137	.022	.028	.025
B (8,2)	Some	.112	.163	.137	.022	.028	.025
HOUSING							
B (9,1)	Single	.125			.025		
B (9,2)	Row, Duplex, Multi	.152			.029		
B (9,3)	Apartments	.171			.035		

(1) 18 years of age and over

(2) 10 years of age and older

* The 1969 and 1972 Frequency figures are not comparable because frequency categories are not identical for both years.

DEFINITION OF THE PROJECTION MODEL THEORY AND APPLICATION

Given the two forms of an equation, specified earlier as Equation 1, it is easy to see that for all of the individuals in a certain geographic area one could compute their probability of participating in a given activity. In fact, the equation refers to their frequency of participation in a given activity. Frequency presents special issues. One can see discussion in Ch. 1 (regarding Cicchetti, Seneca and Davidson 1969) as well as below. Still, for now ignoring model structural issues, as long as one knows the socio-economic characteristics of each individual and has related census information from which to establish trends, one can make predictions. Visualize for the moment that one has an equation for each individual. Then, for example, for males, i=1, one has the situation indicated in Figure 3. For example, in column 2 of the figure, one gets T(i)U(i) as a total because if the U(i) (a constant for males) is added with the weight for each member of the population, T(i) males, one must get T(i) U(i)'s, or T(i)U(i) as a total. The difference in the other columns is that there are both 0's and l's, so certain B's, those multiplied by l's, have weights added whereas their effects is not included for zeros. Now, for a person, some of the B's are multiplied by 1 because the person has particular characteristics (e.g. they live in a city of over 100,000 population). So, when one adds up the l's in a column without considering the B's, one finds out how many of the T(i) people (males or females) have the particular characteristic which is being considered. Obviously, for each socio-economic variable the number of people in each level of the variables adds up to the total number of people in the population being considered, T(i). So what is depicted in Figure 3 is that individual equations for people can be added up and there is no need to consider equations for each individual but only to take the regression coefficients and multiply them by certain data, for example, from the census (e.g. data on how many females in Quebec come from cities of over 100,000 or come from certain household size categories, etc.). One can make estimates without knowing individual characteristics but by knowing information such as indicated in Equation 2 following:

(2) $N(i) = U(i)T(i) + (\Sigma_k \Sigma_j B(i,k,j)n(i,k,j)) + (\Sigma_k \Sigma_j (B(i,k,j)n(i,k,j)) + \varepsilon(i,k,j)$

WHERE N(i) is the number of males (i=1) or females (i=2) participating in some activity and U(i) and B(i,.,.)'s are the parameters as defined earlier with k referring to a particular socio-economic variable and j referring to levels of the variable k.

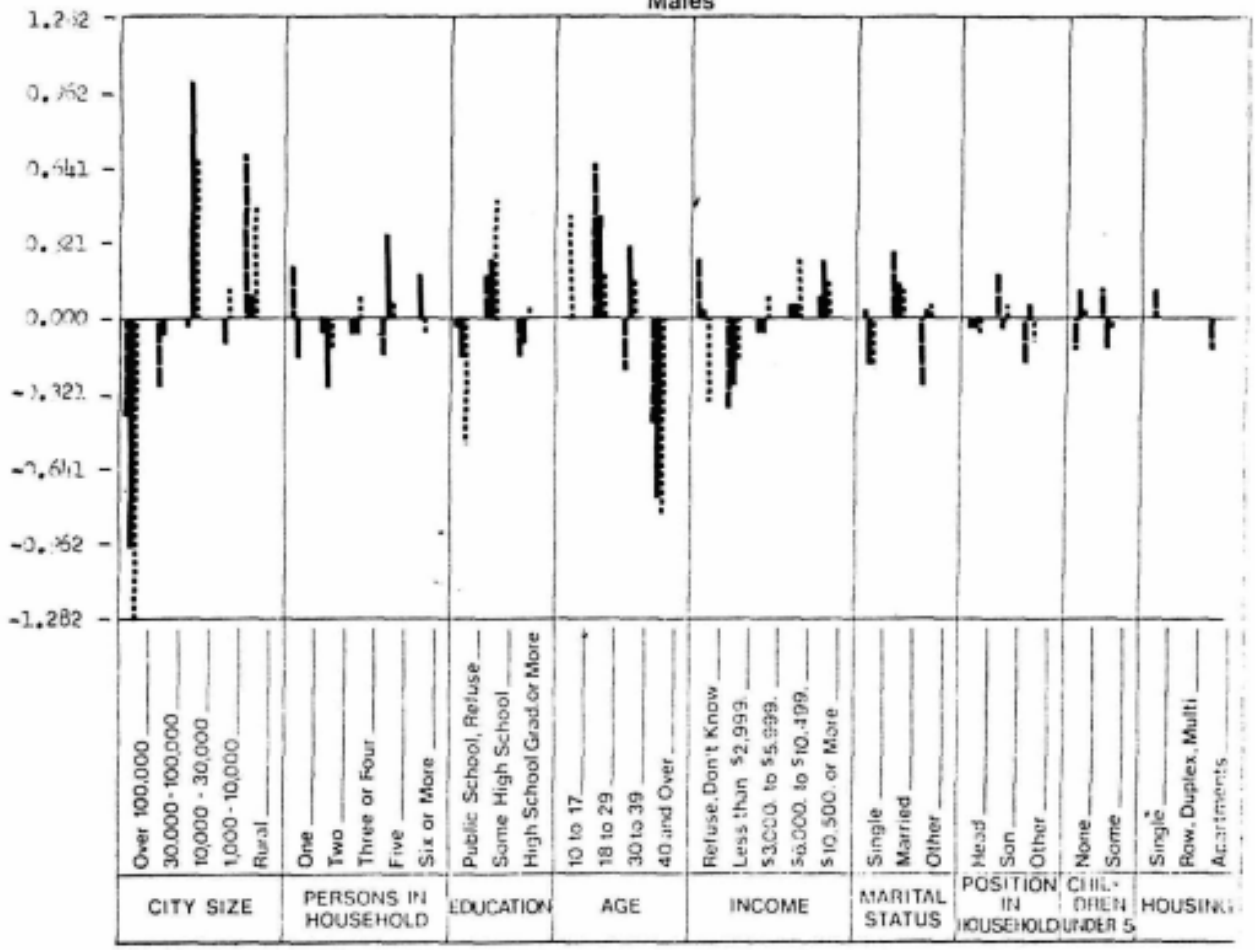
Incidentally, the subscript for gender has been introduced to stress the importance of dealing with the two genders separately for most activities (gender-activity interaction for effect). The importance of carrying out analyses where N(i)'s for the two genders are predicted independently can be very easily understood. By looking at Figure 1 one sees that female participation in hunting has coefficients which look somewhat similar to the male coefficients but the male and female coefficients are not related so that one simply adds a constant value to the female effects to get the male effects. The female effects may approximate a scaled-down version of the male effects, e.g. being a quarter to a third as large. A multiplicative relation (being ¼ or 1/3) between coefficients involves interaction effects. Renoux (1973, 1975) has presented a similar considerations in his analysis of the CORD Study data. Interactions are discussed from a different perspective in TN 20.

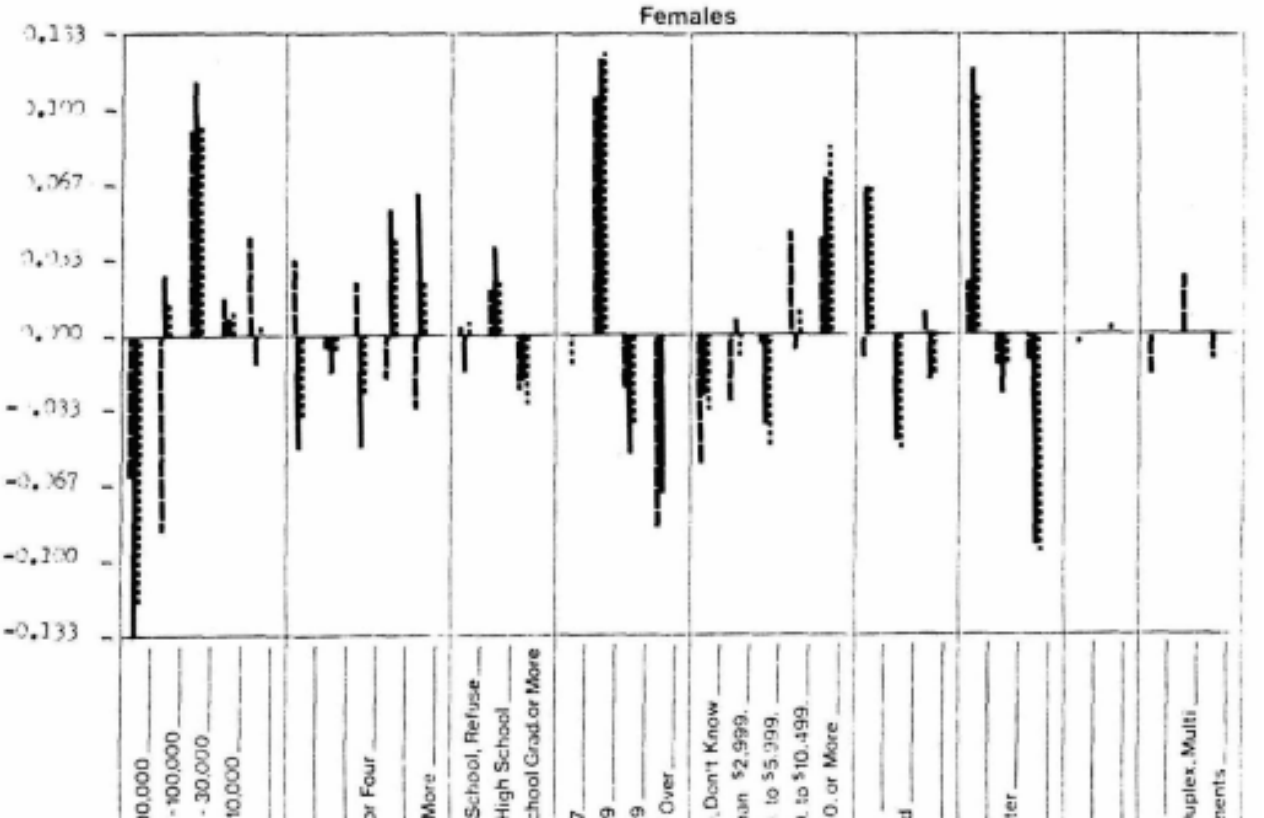
FIGURE 2

BETA VALUES FOR

[illegible] FREQUENCY

IN 1969 — —, 1972 (1) — AND 1972(2) ···





If one has used Equation 2 to calculate N(1) and an N(2), to get results for the total population one obviously need only add together male results and female results. Note that one is getting persons or person times participating. Given that one needs party trips or visits (the typical unit in "destination" models as per Chapter 2), then one needs a model based on parties as a unit or needs a way to transform estimates in persons to estimates in parties. Equation 2 being described for predicting participation implies participation measured in units based on what was estimated. To clarify one aspect of matters, there are different approaches that may be taken to predicting frequency of participation measured in "person-times". One is the approach which is implicit in the way that Equation 1 was defined and suggests that frequency be predicted as indicated in Equation 3:

(3) $TP(i) = U(i)T(i)+\Sigma_k \Sigma_j B(i,k,j)n(i,k,j)$

WHERE the B()'s are estimated based on predicting frequency and the sums are on k and j, as per Equation 2, and where TP(i) is total amount of participation in "person-times" in a given activity for males (i=1) or females (i=2).

Rather than using the equation for total participation just indicated, data can be used to obtain an average frequency of participation in hunting *by participants*. For example, this may be computed by adding up the number of trips that each participant makes and dividing by a total number of participants. Actually, the average number figure should probably be disaggregated according to rural/ urban or in some other way that is consistent with the socio-economic variables considered in the analysis. This is so that as the population changes an appropriate change can be made in the amount of participation predicted. But, that was not done here. Equation 4 is not written in such a way as to allow for this:

(4) $TP = \hat{h}(1)N(1)+\hat{h}(2)N(2)$

WHERE TP is total amount of participation in a given activity for males and females,
$\hat{h}(i)$ are average frequencies for each gender as described above,
N(i) are as defined by Equation 2.

Really, the equation shows that one multiplies each of two types of participants by their average frequencies of participation. Renoux (1973, pp. 66-67) has presented average frequency figures of 2(1) = 4.4 and 2(2) = 3.0 for hunting in Quebec.

As described in Chapter 1, Cicchetti, Seneca and Davidson (1969) have proposed that to get a frequency model that approximates the "real world" one should use two models. One should use one model to estimate numbers in certain categories participating. One determines a frequency of participation model based on data from those that actually participate. Doing this creates the possibility of capturing changes in participating at all and changes in participation by those that participate. The two model approach is not pursued here.

Given the kind of equations just indicated, there is no need to use them with population figures for the year in which the data were collected for determining the B(.)'s . Obviously one can make predictions by using the equations but changing the population data. One can also see how the "prediction" equations introduced earlier are used by seeing that in each of Tables 5 and 6 the numbers of people in various socio-economic categories are multiplied by the effects associated with given socio-economic -categories. These products have been added together in the ways specified by the equations, with the actual estimates being shown at the bottom of the tables as totals. The results of carrying out these computations are summarized in Table 7. In terms of computations, it is actually a very simple procedure to make projections using this kind

of equation.

A problem in making forecasts is determining the N( )'s, the numbers of people that are predicted to be in certain socio-economic groups at some time in the future. One should note that for the example in Tables 5 and 6 estimates were made of what the total Quebec population would be for males and for females in 1980. Estimates were also made of how many people would be in certain socio-economic categories. Care was then taken to see that the predictions as to the number of people in socio-economic categories agreed in total with the total number of people available to be in these categories. This is expressed by the constraint:

(5) $T(i) = \Sigma_k\, n(i,k,j)$

WHERE the sum is over all values of j (the levels) and the condition must hold for all k (socio-economic variables), for each I, e.g., each gender.

Figure 3: Equations for Males Participating in Hunting for a Population T(i) Considering "City Size" and "Number of People in the Household"

Probability of a person participating in hunting				City Size						No. of persons in household									
		General Mean (1)		100,000+	(2)		.....	(3)			(4)		1-2	(5)		3-4	(6)		5+	(7)
Y(1,....)	=	U	+	B(1,1)	0	+	B(1,2)	0	+	B(1,3)	0	+	B(2,1)	0	+	B(2,2)	1	+	B(2,3)	0
Y(2,....)	=	U	+	B(1,1)	0	+	B(1,2)	1	+	B(1,3)	0	+	B(2,1)	0	+	B(2,2)	0	+	B(2,3)	1
.....		...		....			....			....			....			....			....	
.....		...		....			....			....			....			....			....	
Y(j,....)	=	U	+	B(1,1)	1	+	B(1,2)	0	+	B(1,3)	0	+	B(2,1)	0	+	B(2,2)	0	+	B(2,3)	0
.....		...		....			....			....			....			....			....	
Total Estimated Participation																				

equals B(2,1) times the number of persons of sex i whose household size is 1 or 2. This is the no. of 1's in the column marked (3).

equals B(1,1) times the number of persons of sex i in the population who come from cities over 100,000. This is the number of 1's in the column marked (2).

equals U times the total population of sex i, UT(i), i=1 or 2. (indicated in column (1)). The figure is supposed to imply that there are T(i) U's in the column marked (1).

It may seem odd that a point has been made of the fact that the number of people in various levels for socio-economic categories should add to the total number of people in a population. However, this has actually been found to be a problem for several reasons. For one thing, some researchers have projected the number of people in various socio-economic categories by drawing trend lines through the numbers of people in these categories at a number of census years. From this procedure, for example, they get estimated numbers of people in cities over 100,000 at future years. From another curve, estimated numbers of people in various other city size categories are obtained in a similar way. This is done for other variables. Also, a trend line is drawn to determine the total number of people in the population being considered at some future points in time. There is no constraint that makes the prediction of "total number of people" obtained by adding together the numbers in (say) each city size or each level of education agree with the total number of people predicted to be in a population based on a trend line.

If the researcher uses "trend line" projection she/he must make some choice as to how to deal with the problem of differing totals. One simple approach is to make "trend line" projections

for the number of people in various socio-economic categories. These projections can be used to determine what part (proportion) of a total population defined independently would be, say, in each level of education. The following equation then applies:

$$\text{Corrected Number in level j of variable k} = \frac{\text{(Estimated number in level j for variable k)}}{\text{(Sum of number in each level for variable k)}} * \text{“Correct total”}$$

TABLE 6

THE PROJECTED HUNTING FREQUENCY IN QUEBEC IN 1980 BY SEX, AGE, INCOME AND URBANIZATION

	MALE			FEMALE		
	TOTAL POPULATION	GENERAL AVERAGE PER CAPITA FREQUENCY	GENERAL MEAN	TOTAL POPULATION	GENERAL AVERAGE PER CAPITA FREQUENCY	GENERAL MEAN
	2150782	1.053	2264773	2274707	0.042	95537
VARIABLES	PERSONS	EFFECT	TOTAL EFFECTS	PERSONS	EFFECT	TOTAL EFFECTS
AGE						
18-19	94,626	0.039	3,690	89,788	0.353	31,695
20-24	346,550	0.742	257,140	341,821	0.315	107,674
25-29	293,233	-0.318	- 93,248	269,524	0.186	50,131
30-34	189,092	0.332	62,779	255,290	-0.121	- 30,890
35-39	219,706	0.280	61,518	214,098	-0.128	- 27,405
40-44	198,264	0.128	25,378	195,646	-0.137	- 26,804
45.49	168,657	0.036	6,072	167,411	-0.154	- 25,781
50-64	414,605	-0.850	- 352,414	441,189	-0.153	- 67,502
65+	226,049	-0.390	- 88,159	299,940	-0.161	- 48,290
Sub Total			- 117,244			- 37,172
INCOME						
0-2999	346,547	-0.671	- 232,533	366,514	-0.044	- 16,127
2000-4499	11,537	-0.232	- 2,677	12,202	-0.087	- 1,062
4500-5999	44,394	0.325	14,428	46,952	-0.091	- 4,273
6000-7499	321,276	0.231	74,215	339,787	0.070	23,785
7500-8999	387,713	0.077	29,854	410,053	0.071	29,114
9000-10499	331,777	0.366	121,430	350,894	0.025	8,772
10500+	707,538	-0.097	68,631	748,305	0.056	41,905
Sub Total			63,914			82,114
URBANIZA-TION						
1000-999	244,701	-0.320	- 78,304	253,379	0.002	507
10000-29999	191,021	-0.020	- 3,836	198,139	0.026	5,152
30000-99999	198,309	0.206	40,852	214,330	-0.066	- 14,146
100000+	1,198,204	-0.470	- 56,315	1,309,365	-0.061	- 79,871
Rural	318,547	0.320	101,935	299,494	0.099	29,650
Sub Total			- 502,509			- 58,708
TOTAL			1,581,106			81,771
TOTAL NUMBER OF TRIPS ESTIMATED (MALES PLUS FEMALES)						1,662,877

TABLE 7: THE RESULTS OF PROJECTING HUNTING PARTICIPATION AND FREQUENCY FOR 1980 USING EQUATIONS 2, 3 AND 4

	MALES	FEMALES	TOTAL	"SOURCES"
Participation	425,742	63,267	489,009	EQUATION 2
Total Participation	1,581,106	81,771	1,662,877	EQUATION 3
Total Frequency	1,873,264[1]	189,801[2]	2,063,065	EQUATION 4

[1]This is the product of 425,742 and the average frequency for males given in the paper.
[2]This is 63,267 times the average frequency of participation of for females given in the paper.

One other example of where problems have occurred is where census information is not available on certain characteristics. Income of households may be available but in a survey one may ask for personal income. A problem can arise in trying to procure census information on income by person rather by household. What data are needed in the male and female models for which results were generated depends on what was asked in the survey. Given data needed for projection for a variable could not be obtained directly, an approximation approach was adopted accepting the fact that there will be some error.

DISCUSSION

Although it is a simple matter to make projections in the way that has been described, this does not mean that the procedure yields valid results. There are a number of different concerns. By using a particular model, one implies that the model is structurally adequate, a good/adequate approximation to reality. And, as already mentioned, if responses are not accurate the accuracy of the projections must be questioned even if variability implies the estimates are reliable enough (have a low enough standard deviation). However, earlier it was indicated that only certain matters of concern about the modelling problems were to be pursued in this paper. One may have noted that the coefficients of the model were computed for nine socio-economic variables whereas in estimation only three of these socio-economic variables were used. In statistical analysis, it is well known that the interrelationship between two variables - for example age and income - may be such that a coefficient which is used to take into account the effect of age may also reflect the effect of income. In the technical literature one refers to the collinearity problem. An examination of the correlations between the variables that were left out of the model and the variables that were included shows that some of these variables are correlated. So, there is a possibility that there may have been an over or under compensation for some factors that reflect the behaviour of the people in the population. Renoux (1973, 1975) has commented on this problem and there are numerous discussions of the problem of multi-collinearity in the statistics and the econometrics literature.

Another type of problem is that the frequency models defined here seem to be poor models. Such a comment may seem odd. However, when Renoux (1973) began to seriously investigate how to predict the total amount of hunting in Quebec, he found that results obtained using the first form of the frequency model often did not make much sense. That is why the second type of frequency model was introduced. Renoux and Beaman (unpublished) concluded that problems with the "type 1 frequency model" were caused by the drastic skewedness of the distribution of stated frequencies of participation. Cicchetti, Seneca and Davidison's 1969 proposal results in reduced skewedness because one does not have a frequency distribution with, e.g., 75% or more of participation frequency being zero. At first it was considered necessary to "correct" for this skewedness by using weights so that the larger uncertainty that one may have in

a person's statement that he participated 100 times rather than zero times would be reflected in estimates. However, soon the kind of problems with models discussed in a number of technical notes were recognized and it was decided that further work on a frequency model (e.g., pursuing the Cicchetti, Seneca & Davidson 1969 proposal) should be deferred until the research on other matters was completed (for TN 6, 20, 29 and 35).

It may not be clear that there must be data on a specific age group before it is valid to make the predictions for this age group. What is the importance of this? The data from which the regression coefficients presented here were obtained were either for 10 years of age and over (1972) or for people of 18 years and over (1969). So for hunting it is quite possible to make predictions for, say, people 16 years of age and over by assuming that they have the same regression coefficients as people 18 years of age and over, if 1969 data are to be used. If 1972 data are used, coefficients must be defined in some special way using the age specific data which do not give month of birth. However, explicit age information need not be available. So it is possible to make predictions about total number of people of all ages participating in some activities. To further illustrate this, if one knows on average how many heads of households are with each picnicking party or has some other measure that can be used to divide or multiply a predicted figure one can possibly translate adult person estimates into using parties or people (of all ages). In some cases it is possible to use socio-economic characteristics to infer that if a respondent takes part in an activity, probably a certain number of other people do. The "individual/person" weight can, for example, be increased to make estimates of all participation by multiplying by party size (depending on sampling it may be necessary to correct for numbers of adults in a household who could be in a "person" sample). In this regard care must be taken that children are not counted twice when weights for adults are increased to reflect that children will go with them on trips (e.g. each parent gets half the children). There are clearly problem areas in applications for the prediction technique that needs to be explored in terms of particular implications to make certain kinds of total use estimates that may be of interest to planners or managers.

In one case, an estimation problem may be overcome by considering that certain activities are carried out by family groups. Then family information and socio-economic information can be used to get regression equations. In other cases, average- results from an adult analysis can be multiplied by inflation factors. But, as already implied what to do in particular cases involves research problems that remain to be solved in the future and must be solved in the context of each need for a special kind of information (unit of analysis such as person, person-times, person-trips, etc.). In other words, no simple or universal answers can be given regarding: what are the best data; what is the easiest way to arrive at acceptable estimates of total participation; what is the number of people who participate in an activity; etc.

On a very different matter it may have been noticed that this paper suggests the use of a cross-sectional equation for making predictions. TN 13 points out some of the concerns that arise because the coefficients of the kind of model proposed here are derived for a certain point in time and may be changing over time. Also there are comments made in this volume about the importance of recognizing that the relationship (previously referred to as collinearity) between, for example, age and some other variables actually reflects what may be treated as a causal relation. It is truly unfortunate that no CORD Study research has pursued the matter of developing causal models, for example the kind of path-analysis models that have been described by Blalock (1964; now see LISREL or other structural equation modelling approaches as in e.g., Hayduk 1987). Anyway, it is the case that in the CORD Study, socio-economic variables that do

influence each other in a causal way and do not influence participation "independently" have been used. This approach, at best, is an approximation that can result in errors in using the model for predicting the future.

Another point worthy of discussion is that the modelling introduced treats individuals as participating in one activity or another independently of their participation in other activities. But it has been indicated in TN 10, 32 and 37 (and as well is accepted among researchers) people's participation in one activity is typically related to their participation in other activities. Any models like the present class which treats behaviour on an activity-by-activity basis may have structural problems related to the fact that activities are treated autonomously. Certainly when one considers making predictions of what will happen in the future, there is a danger that by using the model proposed here one is suggesting that what is important in determining peoples future behaviour is their socio-economic characteristics rather than their orientation to a variety of recreational activities. In fact, one important factor is definitely what happens in the development of the supply of facilities for different activities (on supply and participation see TN 29). In part, developments that take place with respect to supply affect participation, not in terms of people's participation in individual activities but in terms of how in the future they allocate their finite time according to some kind of time budget. This depends on the supply of opportunities for a variety of activities that they find in the future as opposed to the supply with which they are confronted in the present.

The development of the Blackstrap Ski Development in Saskatchewan has had a drastic effect not only on the amount of skiing in Saskatchewan but on activities that were taking place in the time now filled by skiing and obviously not now taking place, or at least not as much. Both change in supply and substitution of one activity for another (as has happened with snowmobiling, biking and cross country skiing) have played important roles in altering the behaviour of a large number of Saskatchewan residents. The supply component of this change may quite possibly be taken into account to some extent on a single activity basis for some activities by incorporating a supply factor into the kind of analysis of variance model considered here (see TN 29). But as pointed out in CORD Study TN 10, the problem of how to compute the impact of the movement into skiing on other activities which were previously participated in by a person still remains as a matter for research. There are no practical answers at this point in time.

TYPE OF DATA NEEDED FOR MAKING PROJECTIONS USING THE ANALYSIS OF VARIANCE MODEL

In previous sections of this article it has been pointed out that census data were used to make certain projections. In practice there has been a continuing problem in getting good agreement between census definitions and the definitions used in the CORD Study National Surveys on which B( )'s are based. Also, published census results often do not give information such as education by age for males, or even income of head of household by gender of household member, which is necessary if the present model is to be used with income as a variable. Education by gender by age data are needed if the fact that education means something different for young people than old people is taken into account in a slightly more general model (see TN 20). The point is that for a model such as the one introduced to be used there must be a great deal of care taken to see that survey definitions do correspond with census definitions. What is more, if special tabulations from census data are required, plans must be made well in advance of carrying out an analysis so that special tables can be produced and ready when projections are actually to be made.

The biggest complication in using the kind of model described, insofar as data are concerned, is that it is not the straightforward matter that one might suspect to make projections of what the age distribution of the population will be at some future time, or what the education distribution will be. If one gets into the matter of trying to predict the breakdown of education by age at some future time, the problems are truly very complicated. Now, it has been the case that statistical agencies responsible for making projections have refused to make the projection necessary to use the- model described because they say they cannot make the required estimates accurately. But here one enters into a rather delicate problem of whether "accurate" projections are really needed for planning purposes. It is better to plan on the basis of the best projections that can be made rather than plan based on wild guesses.

It is not a straightforward matter to decide how good projections should be before a model should be used. But one can say that for using the model described here, there is no need to spend fantastic amounts of money on getting "extremely reliable" estimates of what the age distribution or education distribution is going to be in the future. The results predicted are not going to be accurate to the same degree as the predictions on the distribution of education or age, or none of the distributions will be accurate. Demographic variables should be predicted with an accuracy (demographic variables may be accurately assessed because of their nature and quality of census data collection) somewhat greater than one feels reliability, and hopefully accuracy, implicit in the analysis of variance model, justifies. Certainly nothing is gained by having accuracy in population figures which is lost as soon as regression coefficients are applied to make predictions. (This and other matters are pursued in TN 6 and TN 20.)

CONCLUSION

This paper has presented a model for predicting participation in Outdoor Recreation. However, there has not been a wholehearted endorsement of using the procedure. Rather the paper should have provided the reader with a clear understanding of how the analysis of variance technique can be used should it appear better than other alternatives which might be used to generate the same kind of information. It might be chosen because other methods appear more costly or are not feasible for other reasons (e.g. data cannot be collected and analyzed in the time available).

An important point to recognize is that the recognition of how such a model can be used is an important research step towards developing more sophisticated models. There is certainly a need to develop such models so that extremely expensive surveys need not be carried but so that equally good results can be obtained by using predictive models along with census information.

Even now, for all the criticism that can be leveled against it, the model described here is good given that the present ability of researchers to make predictions is very limited. Massive amounts of money can be expended on surveys when the effort might better have been spent on trying to develop scenarios of what will happen in the future and thus to relate current behaviour to the way the population is likely to behave in the future. It is this point that is stressed in CORD Study TN 13. This latter note, to a certain degree, defends the use of the technique described here.