TN 6: THE ACCURACY OF ESTIMATES OF PARTICIPATION AND FREQUENCY OF PARTICIPATION USING ANALYSIS OF VARIANCE MODELS

By J. Beaman and M. Alvo

PROLOGUE

The 1976 version of this article included a lot of material relevant to that time period. Such material has been dropped. The article while presenting the same ideas as the original article gives information relevant to 2006. This is done without any consideration of the appropriateness in 2006 of using a methodology that was proposed in the early 70s.

ABSTRACT

The purpose of this paper is to present general expressions for the *statistical* error to be expected when making predictions of frequency/amount of participation and of numbers of participants in recreational activities using linear regression models (in CORDS also called analysis of variance, ANOVA, models) and to present survey planning and secondary analysis implications of having estimate reliability formulae.

When a multivariate regression model is developed on the basis of statements made by individuals about participating or about how much they participate in certain activities (see TN 12), it is reasonable to consider using the model in various ways. One predicts results by multiplying regression coefficients by “constants” and summing these products. Results on the variability that is to be expected in estimates is presented in equation form. Model structure being a good approximation to reality is discussed as (1) affecting the value of estimates and (2) making variance estimates more than a fiction created using statistical models where they do not apply. How variability in an estimate can be used to consider what reliability can be expected in a similar survey of a different size is considered. An example is presented based on total number of male participants in hunting that can be expected in Quebec.

The value of using model estimation t get desirable variability for forecasting and for estimation of participation for areas for which survey results (a) are not available or (b) are too variable to be useful (e.g., because of small sample size) is considered. How to make variability computations is describe, in particular guidance is given on getting information need in variance estimation using statistical programs available in 2006.

PURPOSE

One purpose of this note is to show how the reliability of estimates made using a regression model can be assessed, given the model used can be accepted as being structurally sound (being a good approximation to reality) . Another purpose is addressing application of the method in terms of it being likely to produce valid and useful variability estimates.

INTRODUCTION

One of the models used as part of the CORDS can be expressed in a simple forms (see e.g. TN 12 Equation 1, Form 1). Equation 1a is a verbal version of Equation 1b.

Equation 1A:

Probability or frequency of Participating in an outdoor recreation activity	=	General Mean	+	Income Effect	+	Age Effect	+	Edu-cation effect	+	Urban-ization Effect	+	Family Comp-osition effect	+Error

Equation 1B: $Yr_i=\mu+\beta_{1,J}+\beta_{2,K}+\beta_{3,L}+\beta_{4,M}+\beta_{5,Q}+\varepsilon_r$

WHERE i is the person being considered;

Y_r= the probability or frequency of a respondent, *r*, participating in a particular activity (for estimation of participation Y(r)=1 is participated and =0 is did not);

μ = the constant for a particular activity (general mean)

1, 2, 3, 4 and 5 refer to 5 variables but there could be V_n n>0 each with m_n >1 levels

J, K, L, M and Q are valid level for variables 1 to 5 respectively (Here level values are treated as categorical).

$\beta_{1,J}$, $\beta_{5,Q}$, etc. = the beta/regression coefficients that apply to each category that defines the socio-economic status of a person being considered.

Values of "β" coefficients that were estimated for hunting are shown in Figure l. For some detail on how to interpret and use the coefficients, one may refer to TN 12. In summary, one obtains a probability or "amount" by summing coefficients that apply to a person. For participation in hunting for a person of moderate income one sees from the figure that effects of about 0.07 would be added for income and age. If one had the graph for females one would see the patterns of effects tend to be much smaller, as is the general mean, μ. This implies that you cannot just get the probability for females by a gender adjustment to the probability for a male with similar attributes. A model only including effects for "individual" variables is referred to as using a main effects model. When one has, e.g., age effects by gender (an effects showing how men and women differ by age in participation), one refers to using a model with interaction effects.

\The best known use of the kind of "effects" model just introduced (pre 1970) was an application by Mueller and Gurin (1961). It was originally recommended that the model be applied in the CORDS by Hendry (1969). Subsequently TN 12 was prepared. As well, two works involving its use have been prepared in Quebec (Renoux 1973, 1975).

What is important to this research report is recognizing that equations given in TN 12 for making predictions (see Equations 2, 3 and 4) can be described as what statisticians call estimable functions. Here all that matters is that, based on estimates made using a model like the one expressed in Equation 1, one makes an estimates by multiplying coefficients by sizes of populations to which the coefficients apply. This combination gives an estimate of "total amount of participation" or of total participants. Equation 2 is the equation for expected number of participants or amount of participation for a population of size N with $n_{v,c}$ "net" people in categories defined by levels *c* of variables *v*.

(2) $\hat{E}(T_A)=\hat{\mu}N+\sum_{v=1}^{m}\sum_{c=1}^{l(v)}n_{v,c}\beta_{v,c}$

WHERE the sums are over levels, *c*, of variables, *v*, for *m* variables each with *l*(v) levels.

$\hat{E}(T_{\boldsymbol{A}})$= expected number of participants or amount/frequency of participation in a particular activity, ***A*****,** by a population

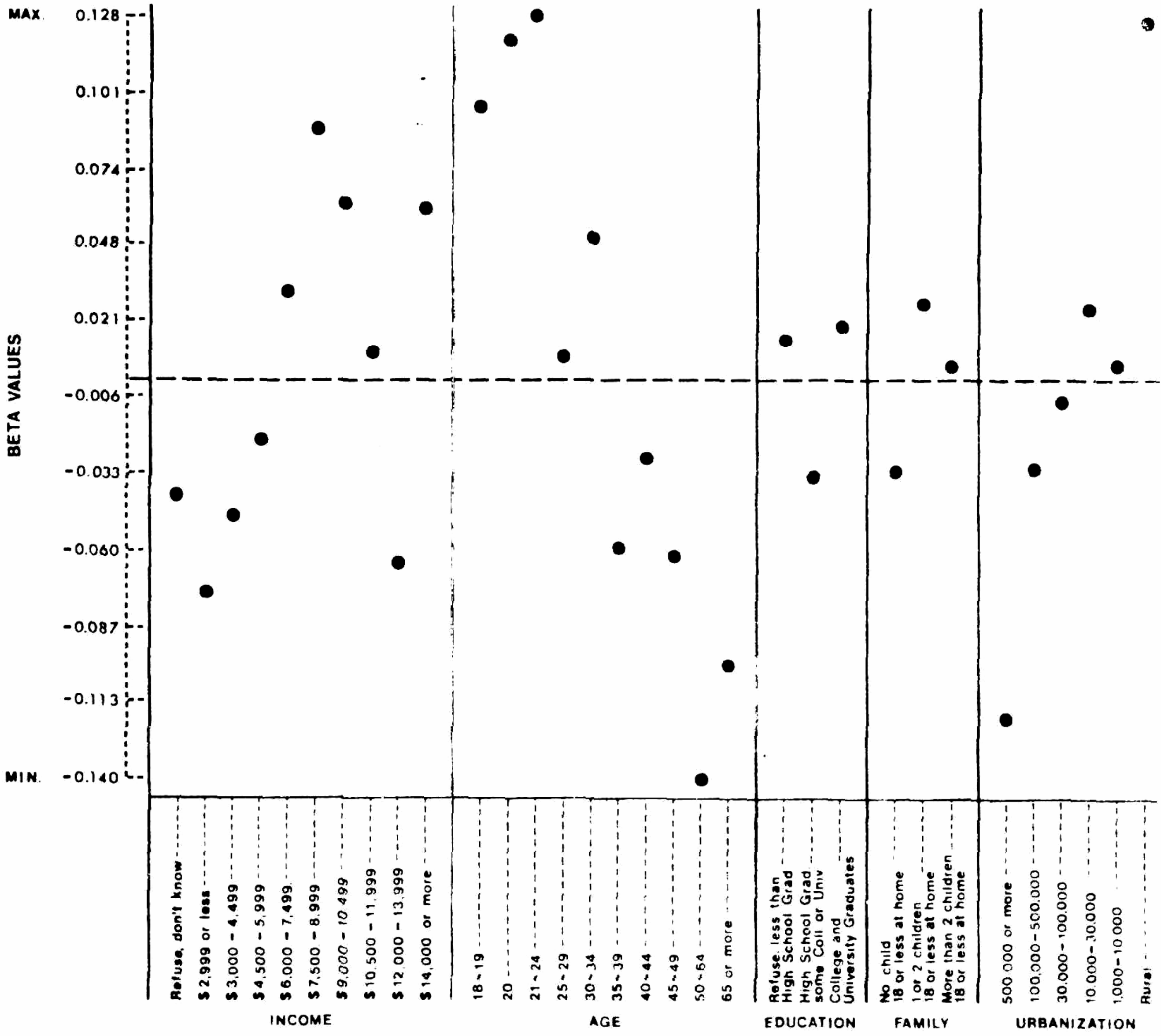
$\hat{\mu}$ = the general mean for ***A***

N = the total size of the population of concern

$\beta_{v,c}$ = the effect of being in a category x of a socio-economic variable *v* (for ***A***)

$n_{v,c}$ = the "net" numbers of people in the population to whom $\beta_{v,c}$ applies ("net" refers to the need in some regressions to consider people in an "omitted" level of a variable).

Figure 1: Relationship Between Participation/Non-Participation In Hunting And Selected Socio-Economic Characteristics (For Males)



THEORY REGARDING MEASURING ERROR/RELIABILITY

In words Equation 2 states that if you have estimated regression coefficients ($\hat{\mu}$ and $\beta_{v,c}$) for participation or "amount" then to get an estimate $\hat{E}(T_A)$ of participation or amount for a particular population (e.g. numbers hunting or days/trips hunting by Quebec males), you multiply the population size by the estimated overall mean ($\hat{\mu}$) and then for each category considered in the model perform a similar operation ($n_{v,c}\beta_{v,c}$). Finally, you add up the results. Justification for doing this is in TN 12. In that note, as in the figure, there is the implicit assumption that coefficients applying to levels of a variable sum to zero. That is that they have been obtained in such a way that sets of regression coefficients that apply to levels of a variable (or combinations for interactions) meet certain constraints (see constraints and the "H" matrix in Scheffe 1959, pp. 15-19). The ways coefficients are often computed (e.g., leaving out one level and declaring presence or absence of a level, including the missing one, by dummy variables having particular values) affects how one computes variability in Ê(T*A*) Regardless, the general idea that matters is that you are adding an effect the number of times that there are people that it applies to.

Independent of details about estimation of the model used, a critical assumption is that the model used is a good approximation to reality. Therefore, if combinations of levels of variables affecting behaviour need to be considered (e.g., combinations of gender and age) then interaction effect regression coefficients must appear (e.g., $\beta_{v1,c1;v2,c2}$). One could consider combinations of levels of *v1* (e.g., gender) and *v2*.(e.g., age). In doing this for gender and age you would need to know population sizes such as the numbers of males and females in each age group ($n_{v1,c1;v2,c2}$). In this article's Appendix, we clarify what population figures would be used with regression coefficients estimated in a particular way using two modern (in use in 2006) regression programs.

Since here one is always considering a sum of products of regression coefficients and population/subpopulation sizes, to make this presentation as simple as possible to understand, Equation 2 is rewritten as Equation 3. In Equation 3 one sees a single sum in which regression coefficients are multiplied by population/subpopulation sizes. Again, specific information on how a regression estimation is made determines how β_{dv} and n_{dv}.

The form of Equation 3 makes it clear that this paper does not pursue regression using continuous variables (e.g., y=constant+coefficient*age or *income). This is because for such regression, in *formal mathematical/statistical terms,* zero people are associated with a particular value of a variable. The continuous variable formulation implies that a variable, such as age or income, takes on an infinite number of values *all of which are considered to be possible responses*. Therefore, a finite number of people is only associated with an interval. For example, zero people are associated with $100.00 while a number is associated with dollars, d, in $100-x<d<$100+x where x>0.

$$(3)\quad \hat{E}(T_A) = \sum_{dv=0}^{m} n_{dv}\beta_{dv}$$

WHERE the sum is over $\hat{\mu} = \hat{\beta}_0$ and the *m* other regression coefficients $\hat{\beta}_{dv}$

Ê(T*A*)= expected number of participants or amount of participation in a particular activity, ***A***

n_{dv} = the numbers of people in the populations to whom $\hat{\beta}_{dv}$ applies

One purpose in using Equation 3 is that it facilitates introducing the matrix formulation of least squares regression. Equation 4 expresses Equation 3 using a vector/matrix notation. For those who are not familiar with matrix equations, while presenting results using them, we interpret these results so all readers should understand what the equations mean/imply. In this regard, Equation 4 is just a compact way of writing the sum expressed in Equation 3.

(4) Ê(T*A*)= $n'\hat{\beta}$

where n is the 1*x*n matrix (vector) of population "sizes" and $\hat{\beta}$ is the m*x*1 matrix (vector) of parameter estimates.

The "'" symbol is read as "transpose." When one writes *n*' it refers to the horizontal vector or 1*x*m (read 1 by m) matrix ($n_{1,}$ n_2 , n_{m-1} , n_m). Without the transpose the vector/matrix *n* is vertical (m rows and 1 column). Matrix multiplication involves multiplying corresponding elements in rows and columns (e.g. $n_1\beta_1$, $n_2\beta_2$ etc.) and adding up these products. That is how you get the sum shown in Equation 3.

Returning to more mundane matters, in this paper a point is made of using Ê(T_A) and referring to estimating the expected number participating or to the total participation in "activity" ***A*** as Ê(T_A). In using a sum of population sizes multiplied by estimated coefficients ($\hat{\beta}_{dv}$ rather than "exact ones", β_{dv}), one can take as given that one is using a regression model similar to the one in Equation 1b to determine regression coefficients and then using the coefficients determined to make an estimate, Ê(T_A). Therefore, what one gets is estimated "expected" participation *given survey results*. For any number of reasons (Expo 67 year, storms, floods or Olympic Games), actual participation in a given geographic area in a given year could vary from what *on average* is expected . Here we are just concerned with "random" statistical variability. In that regard, given there is nothing exceptional about the year of a survey, the expected value which is predicted using a model can be more appropriate for planning or policy than observations that reflect circumstances peculiar to a given place in a particular year. Regardless, the main point is that Ê(T_A) is an estimate based on estimated values of regression coefficients. Therefore, it is subject to statistical variability because the coefficients are. One can think of the amount of variability in Ê(T_A) reflecting the values that would arise if $\hat{\beta}_{dv}$ were determined by one survey, then another and another. Here this variability is measured as variance, Var(Ê(T_A). Var(Ê(T_A) is by definition the average squared deviation of Ê(T_A) from its overall "average" value E(T_A), average of (Ê(T_A)-E(T_A))^2.

Moving back to estimating variability, one is concerned with how Ê(T_A)varies around E(T_A). Here we are specifically concerned with the expected/average value of Var(Ê(T_A))=**E**(Ê(T_A)-E(T_A))^2. **E**( ) is read expected value of. Now, a person might think from their statistics class(es) that he/she can take standard deviations in $\hat{\beta}_{dv}$ values, Std($\hat{\beta}_{dv}$) (see Tables 2 and 4 of TN 12), and estimate Var(Ê(T$_A$)) as indicated in Equation 5. Equation 5 is based on the variance of a sum of independent random variables, X_i, times constants, C_i, being *the sum of the variances in those random variables*(Var(X_i)=(Std(X_i))^2)=$\sigma_{X(i)}{}^2$ *multiplied by the squares of the constants,* C_i^2. However, in general, the $\hat{\beta}_{dv}$ obtained in a regression are correlated. Therefore, one must take that correlation into account in estimating Var(Ê(T$_A$)). This is done by considering the covariance between regression coefficients.

Let Σ_β be the matrix/array of covariances, $Cov(\hat{\beta}_n, \hat{\beta}_k)$, between regression coefficients. As indicated in Equation 6, in a covariance matrix at row *n* and column *n* one finds $Var(\hat{\beta}_n)$ while for *n*≠*k* one finds $Cov(\hat{\beta}_n, \hat{\beta}_k)$ in row *n* and column *k* and also in column *k* and row *n*. Given that the regression coefficients are uniquely determined (for practical purposes the regression "works" in getting $\hat{\beta}_{dv}$), Σ_β exists. Furthermore, its inverse Σ_β^{-1} exists.

You can think of the inverse matrix, Σ_β^{-1}, as allowing one to correct for large variances by "normalizing" all variances to 1 much as you would divide values of a variable by the variable's standard deviation to create a variable with a standard deviation of 1. However, as well as compensating for differing variability in different coefficients, the matrix contains information necessary to compensate for covariance between regression coefficients. Statistical theory leads to Equation 6 as the correct way to determine $Var(\hat{E}(T_A))$. The matrix equation is just specifying that one must multiply Σ_β^{-1} by the vector of population values. One sees that one multiplies both from the right and from the left. If Σ_β^{-1} was a diagonal matrix (0 for n≠k) with $1/Var(\hat{\beta}_n)$ on the diagonal, the multiplication would result in the sum given in Equation 5. Multiplication of a 1*x*m matrix by a m*x*m by a m*x*1 yields a 1*x*1 matrix which is just a number/scalar (e.g., $Var(\hat{E}(T_A))$).

Equation 5: $Var(\hat{E}(T_A)) = \sum_{dv=1}^{m} n_{dv}^2 Var(\hat{\beta}_{dv})$

Equation 6: $Var(\hat{E}(T_A)) = n'\hat{\Sigma}_{\beta}^{-1} n$ where the "^" indicates $\hat{\Sigma}_{\beta}^{-1}$ is estimated.

Equation 7: $$\Sigma_{\beta} = \begin{bmatrix} Var(\beta_1) & \cdots & Cov(\beta_1, \beta_k) & \cdots & Cov(\beta_1, \beta_n) & \cdots & Cov(\beta_m, \beta_n) \\ \vdots & \ddots & & & & & \vdots \\ Cov(\beta_k, \beta_1) & & Var(\beta_k) & & Cov(\beta_k, \beta_n) & \cdots & Cov(\beta_k, \beta_m) \\ \vdots & \cdots & & \ddots & & & \vdots \\ Cov(\beta_n, \beta_1) & & Cov(\beta_n, \beta_k) & & Var(\beta_n) & \cdots & Cov(\beta_n, \beta_m) \\ \vdots & & & & & \ddots & \vdots \\ Cov(\beta_m, \beta_1) & \cdots & Cov(\beta_m, \beta_k) & \cdots & Cov(\beta_m, \beta_n) & \cdots & Var(\beta_m) \end{bmatrix}$$

It is important to note that the value of $Var(\hat{E}(T_A))$ being near "the correct" variability in T_A depends (1) on the model estimated being structurally appropriate to the data used in estimating regression coefficients and (2) the survey responses being unbiased (e.g., responses not being systematically low because of partial recall). We use "the correct" because there are at least two correctness issues. A model can be appropriate for biased responses (e.g. systematically low) and thus $Var(\hat{E}(T_A))$ is correct with respects to $\hat{E}(T_A)$ but not the actual correct value, say $C(T_A)$. E(T_A) and $C(T_A)$ differ by the bias, $\boldsymbol{B_A}$. If the model is not appropriate for the data, the $Var(\hat{E}(T_A))$ determined is "mathematically correct" but not meaningful because if the model does not really explain the data, estimated values have a somewhat to totally spurious relation to correct values (to what people actually do). In practical terms, given the model comes close to predicting what people do, one can have confidence in using $Var(\hat{E}(T_A))$. If the model is not appropriate to the data, then the mathematics of computing variability and making predictions is just playing with numbers.

The matter of "goodness of approximation" is taken up in CORD Study TN 20. Actually, the conclusion reached there is that a simple main effect model, a model in which e.g. gender, age and other effects are computed without recognizing that there are interactions between these, can be expected to be a poor approximation to reality when one knows there are interactions such as when an activity is male oriented (e.g. hunting with interaction effects as described above). Again, one introduces interaction effects so there are regression coefficients that e.g. allow for age specific difference in hunting participation between males and females. Equation 6 applies when there are interaction and when there are not. The modeler has the problem of getting the model right.

VARIABILITY IN Ê(T_A), RELIABILITY AND BIAS

Using terminology and notation that is used in many statistical applications, Equation 6 results in one knowing the standard deviation $s = Var(\hat{E}(T_A))^{1/2}$ of Ê(T_A). Now, sometimes by chance Ê(T_A) will be very close to E(T_A), its average over many independent surveys administered in the same way. Other times, simply by chance, there may be a quite substantial difference between Ê(T_A) and E(T_A). The size of *s* conveys information about how likely deviations of a given size are.

The common way of expressing variability in estimates is to give a probability that the estimate is within a certain range. Symbolically, Equation 8 expresses the idea of $\hat{E}(T_A)$ differing from $E(T_A)$ by more than d having a probability less than p. A common expression is that for a normal distribution $P(|x-\mu|>1.96\sigma)<.95$. Because theory and practice has provided justification, it is common to assume that $\hat{E}(T_A)$ is roughly normally distributed about $E(T_A)$ with standard deviation s. When normality is accepted probabilities can be based on the normal distribution. Regardless, of whether one assumes normality, one can use expressions like Equation 8 to express desired or actual “reliability”.

Equation 8: $P(|\hat{E}(T_A)-E(T_A)|<d)>p$ where $d>0$ and p is a probability (e.g. .95 or 95%).

In relation to Equation 8 and “reliability” statements, we are not mentioning accuracy because how near $E(T_A)$ is to $C(T_A)$ is generally not known. As mentioned earlier, even if a model is structurally appropriate for data, being accurate depends on responses not being biased: From a sample survey we do know either $E(T_A)$ or $C(T_A)$. Therefore, we do not know $\boldsymbol{B}_A=E(T_A)-C(T_A)$. Determining bias in estimates based on statements about behaviour, survey responses, is not generally a trivial matter. To pursue this briefly, for hunting we may be able to verify that the number estimated to purchase a particular hunting license corresponds, or does not, to the number of those licenses sold. However, people may hunt with different licenses. Some people may have one license while other people have several. Therefore, even given a license is required for any hunting, one cannot “audit” survey estimates of number of hunters against license sales, other than to check that the number of people saying they hunt does not exceed sales of required licenses. Such a check does not allow determining bias though it could establish that there is bias.

To continue from this point, consider that the interval around $\hat{E}(T_A)$ defined in $P(|\hat{E}(T_A)-E(T_A)|<d)>p$ is clearly a function of the degree of confidence chosen, p, as well as of the variability of the estimate. In that regard, Equation 9 is just Equation 8 with d replaced by cs, *a constant times the estimated standard deviation of* $\hat{E}(T_A)$. Now, think about s being determined from a survey of size n. Given how s varies with sample size, Equation 10 express how to compute s_2, if one obtains s_1 in a survey of size n_1 and wants to know the expected size of s_2 in a another very similar survey of size n_2. Of more importance is that one can, for example, use the results from a survey to determine how large a survey must be to achieve a given reliability. We provide an example of this below.

Equation 9: $P(|\hat{E}(T_A)-E(T_A)|<cs)>p$

Equation 10: $s_2=(n_1/n_2)^{1/2}s_1$

Prior to moving on to making estimates for Quebec male hunters, it is important to recognize that when s is determined, you may know something good or bad. Given the model is an adequate approximation to reality, if s is, say, 50% $\hat{E}(T_A)$ ($50\%=100*s/\hat{E}(T_A)$), your results are so much in error as to be useless. For example there is about a 2.5% chance that $E(T_A)$ is zero and a 2.5% chance it is greater than $2\hat{E}(T_A)$. Based on Equation 10, you would need a survey 25 times as large (n_2/n_1) to get an “s_2” that is 10% of $\hat{E}(T_A)$. Then, given that $P(|\hat{E}(T_A)-E(T_A)|<1.96s_2)>95\%$, you would have a 95% chance that $\hat{E}(T_A)$ was within about 20% of $E(T_A)$.

From a practical perspective, it is worth noting that desired reliability has been discussed in terms of how s, or a multiple of it relates to a prediction. People tend to use criteria like “I want it within 10%.” Given such a statement the analyst requires a feeling as to whether “needing it within” refers to that close 90% of the time, 95% or if some other p is appropriate. In this regard, when you do not yet have a prediction it can be important to consider its likely magnitude and select “acceptable error” as a percent of this. Pulling a number out of the air like 75 that turns out to be 25% of $\hat{E}(T_A)$ would mean you would be starting out to ge estimates that are too variable to be useful. Unless you go into data collection with information allowing determination of sample size, for all your effort you may come out with data that yields unacceptably poor predictions. Proper planning is critical to getting adequately reliable predictions. Another matter in which sample size matters is deciding whether secondary data available for analysis has an adequate sample size to be worth using. Given one survey, Equation 10 provides the basis for making a decision. So, you can view this paper as about more than measuring reliability after data are collected.

A QUEBEC EXAMPLE

Suppose that the amount of participation in hunting by the male population of Quebec is to be predicted for the year 1975. Let the survey data used to estimate regression coefficients be the 1972 CORDS National Survey on Canadians' participation in outdoor activities (documented in CORDS Volume III). The income, urbanization and age categories used in a regression model are given in Table 1. You can view this analysis as providing an alternative to trying to base estimates for Quebec male hunting on tabulations made just using the data for males in Quebec. We consider variability for such sampling below.

The idea in using a model would be to get a smaller "*s*" because the model is based on data for 3000 respondents. Only participated or not is considered (1) because it is adequate to illustrate the matter of concern and (2) because using a very simple frequency model would in all likelihood be setting a bad example for such modelling (see specifically Cicchetti 1973, pp 33-38; also TN 29).

Table 1 actually provides the estimates of the regression coefficients and the population size values needed to estimate the number of male hunters in Quebec in 1975. There, one also sees a column containing the products of population with regression coefficients. When results in that column are added, you get $\hat{E}(T_A)$, as indicated by the last line of the table. In summary, you see that from 1,980,000 males 16+ it is estimated that there are 803,000 hunters.

TABLE 1: INFORMATION OF RELEVANCE IN ESTIMATING NUMBERS OF MALE PARTICIPANTS IN HUNTING IN 1975*

Population by attribute	Symbol for n_{dv}	1000's β_{dv} apply to	β_{dv}	Value	$n_{dv}\beta_{dv}$
MalePopulation					
16+	N	1,980	$\beta_0=\mu$	.438	867.
Income					
(1) 0-$5,999	n(1,1)	685	$\beta_{1,1}$	-.076	52.
(2)$6k-$10,499	n(1,2)	810	$\beta_{1,2}$	.050	40.
(3) $10,500+	n(1,3)	485	$\beta_{1,3}$	.027	13.
Urbanization				-.154	
(1) 100,000+	n(2,1)	1,093	$\beta_{1,1}$		-59.
(2) 10k-99,999	n(2,2)	333	$\beta_{1,2}$	.056	18.
(3) rural-9,999	n(2,3)	554	$\beta_{1,3}$	-.001	-1.
Age					
(1) 18-24	n(3,1)	398	$\beta_{1,3}$	.083	33.
(2) 25-34	n(3,2)	448	$\beta_{1,1}$	.063	28.
(3) 35-49	n(3,3)	553	$\beta_{1,2}$	.008	4.
(4) 50+	n(3,3)	581	$\beta_{1,3}$	-.154	-89.
TOTAL($\hat{E}(T_A)$)					803.

By carrying out estimation of *s* based on Equation 7 one determines that it is about 28. In percentage terms this is about 3.5% of estimated participation (=100*28/803). Therefore, in percentage terms, % error in hunting participation at the .05 level (p=.95) has a range of about 6.6% (= (1.96) (27.6) (100)/ 803). In other words, (1) *given that the model is a reasonable approximation to reality,* (2) *given that the forecast of population to 1975 is perfectly accurate*, and (3) *given no exceptional circumstances associated with the survey year(1972) or 1975*, one is 95% sure that the number of male hunters in Quebec in 1975 was between 750,000 and 850,000.

DISCUSSION

Now, one may wonder why one did not just use Quebec data from the "all Canada" survey to determine $\hat{E}(T_{\boldsymbol{A}})$. Well, the sample for the survey was about 3000 of which about 1500 were males. This is important because the regression was just run on data for males. If one was just going to use data for Quebec males for tabulations, there were about 400 Quebec males. Based on statistical theory, the expected variance in the proportion is about $p(1\text{-}p)/400$ where p is set to 0.4. Thus, one obtains an estimated standard deviation of about .025 ($\approx(.4*(1\text{-}.4)/400)^{1/2}$). As a percent of the 0.4 this is about 6% (=100*.025/.4). By using the Canada sample which is about 4 times as large and making some assumptions, we got 3.5%. Thus by modelling one has obtained an estimate that is "twice as reliable" as expected from using the Quebec data. This result is exactly what would be expected based on Equation 10 (standard deviation in half when the sample is quadrupled). However, if there are differences in hunting participation between Quebec and the rest of Canada that are not accounted for by the model, this "more reliable" estimated could be found to be significantly different from the correct hunting participation for Quebec. We do not pursue this matter (but, analysis can be carried out since the 1972 data in SPSS format are still available from the University of Waterloo as of 2006-see contact information in the CORDS web posting).

We have not yet commented specifically on regression results. The R^2 for the regression was just .084. However, the F=216.5 with 7 and 1263 degrees of freedom has probability less than .001. The regression only being for 1271 respondents (7+1263+1) is a consequence of missing data. Since socio-economic variables are used in the model and only hunted or not would not be needed in using data on Quebec males to get a proportion, one sees that missing data influences results. However, if you were going to forecast a 1972 survey result to 1975 using Quebec data, you would need a method and it might require you know some socio-economic attributes of Quebec males. TN 36 pursues the matter of the meaning of R^2 in regression models such as Equation 1. Obviously, coefficients do not contribute much to explaining variability for participation but even making a 4% or 5% correction can matter.

From the example you may have noted that for participation estimation, when R^2 is low one can estimate var($\hat{E}(T_{\boldsymbol{A}})$) adequately without dealing with covariance matrices. We could just use the "$p(1\text{-}p)/n$" formula. Unfortunately no such very simple "trick" applies when one is estimating amount of participation. Also, the "trick" does not apply when estimates are for segments for which effects (β_{dv}) result in a "p" for the segment that is quite different than the one for the population.

In summary, results show clear advantages of making estimates using models *when one is relatively sure the model approximates behaviour by the population of concern.* One is making big reliability gains by being able to transfer reliability achieved by having a large sample to estimates made for a smaller population. However, it has already been noted that problems can be found with some of the simple CORDS models developed. Highly significant F-values and patterns of coefficients that clearly make sense (Figure 1 and Figures in TN 12) can occur when supply and interaction effects are not being considered or not being properly considered. One can see TN 20 and TN 29 regarding interaction effects and supply effects. If models with higher R^2 can be developed and one has real confidence that these capture most effects influencing behaviour, modelling will be a really useful too. In as much as regression models are used, this variability research will then be very important because it addresses appropriate measurement of reliability of estimates.

Moving to a rather more technical matter, in the original version of this paper a reason for introducing estimation for *participate or not* was to illustrate making one round of computations to get preliminary estimates of probabilities of participation for respondents. A respondent's probability, p_r, then was taken as an estimate in their variability in responding. Based on the binomial distribution, if 1 indicates participation and 0 is "not", sometimes we will observe 1 and sometimes zero. The variance associated with 0-1 responses for respondent r is taken to be $p_r(1-p_r)$. This is the variability expressed by ε_r in Equation 1b (variance that should be taken into account in estimating regression coefficients). Different observations having different variability is known as heteroscedasticity. A useful consequence of weighting is that if some p_r are estimated to be negative or greater than 1 weighting can be used to move them back to valid values. Since observations get a weight $1/(p_r(1-p_r))$ "resetting" p_r that are outside 0-1 near enough to zero or 1 gives them enough "influence" so revised estimates are within 0-1. However, the need to set high weights to force many estimates into the 0-1 range can be an indication that an invalid model is being used. One can find further comment and references in the Smith and Cicchetti "Appendix A" .We do not pursue this matter further since in 2006, one might well use a logistic model rather than a weighted least squares model to estimate participation/non participation (e.g. see SPSS manuals [base and Advance Statistics] or the SAS Stat manual for examples and references on logistic models and weighted least squares regression. See TN 19 for a non "p_r" application of weighted least squares regression).

In the end of the review material for Chapter VII of this volume there is a brief introduction to the matter of how to determine whether some regression models as structurally valid, as good approximations to reality. The point made there is that when iteration is used as described in the last paragraph, the differences between predictions and observations should be distributed in a particular way, *if the model is really appropriate to the data.* Testing of model structural adequacy and finding "adequate" models is clearly an important matter if one is going to use models to predict participation and "amount" e.g., rather than getting into massive surveys to adequate numbers of respondents in small areas so that analysis for these areas can be based on their own respondents.

Two points can be noted here regarding the hunting model being a reasonable approximation to reality. Firstly, results presented in TN 29 show that the hunting model is probably deficient because the effect of supply on participation is not considered in the model. Yet it is noted in TN29 that the effect of supply on participation in hunting (in fact for all 19 activities for which information was available in the data used in this study) could not be demonstrated with data on 4,000 people. There it is shown that (1) supply distribution is important in explaining participation in hunting and fishing and (2) all that was needed to demonstrate this was a larger data set. In this other technical note the great importance of considering supply factors, if large biases in estimates is to be avoided, is illustrated. Secondly, just above a point has been made here of the potential value of using weighted regression. This regression leads to a test for model structure since if probabilities are correct, probabilities have a relation to estimated variability.

Finally, from a planning, management and research planning perspective the results presented here are important. It is not hard to find planning reports or documents prepared for management estimates are provided with no mention of "uncertainty" in them. There can be great inefficiency in carrying out research because the proper questions are not asked about how accurate results needs to be and thereby the question does not arise as to whether projections could be made by a method like the one described here using secondary analysis rather than going through laborious, expensive and time consuming primary data collection. How many times is it that data are collected but estimates do not apply in a certain situation because the universe for a study was not quite right, because the data are a couple of years out of date etc. ? How often is a survey carried out without adequate consideration of what reliability is going to be achieved even when existing surveys could be used to assess expected reliability? A great waste is getting data because somebody decides a sample of 400, 4000 or 40,000 is the right size and this is found not to be correct. Budget may dictate sample size but if the sample will not yield adequate reliability, why bother?

Researchers may often work under the misimpression that planners and managers need results which are far more accurate than what they actually need. Data on how many people actually do use facilities may be plotted to justify program expenditures, but in terms of planning for adequate facilities to begin operation of a park, for the future given new offerings, etc., being accurate within plus or minus 10% may not even be realistic. Most road designs have a leeway of about 50%. This allows both for "peaks" in use and for unexpected expansion of use without the road becoming a problem. Plans to "overbuild" to meet a relatively unlikely level of demand can be cheap compared to dealing with out growing capability. Anyway, it is not the researcher's/analyst's role to decide on scale but having a proper understanding of what planners/managers information needs are and seeing that implications of analysis are properly interpreted is a responsibility of researches/analysts.

In part the points made above relate to using the kind of model presented in designing good research. A part of designing good research is not carrying out research unless it will produce useful results and ideally yield them relatively effectively. Being effective can be not doing a lot of work when rather simple computations are all that are justified. In applied research, this may mean doing secondary analysis when it will give adequate results for planning or management decision making. In this regard it is important to note that the kind of results presented here are important in making a decision as to whether estimates should be made or a survey should be carried out.

Part of good decisions making can involve considering the level of reliability that can be achieved given the funds that are available and appropriate to a project must be considered. Obviously, a $5,600 data collection effort is not a good expenditure for making a decision about carrying out a $5,000 project. To reiterate a point, if a researcher can elicit a good feel for necessary reliability for management and planning, she/he can consider what information is sufficient for decisions making. Furthermore whether millions or thousands of dollars are involved, the question which must ultimately be asked is, if a survey is necessary, how large is large enough?

Of particular importance in relation to the last point is that the prediction of more reliable results by bigger and "better" surveys does not occur beyond a point. There is no point in paying a high price to collect data to have a model which gives ±1% results for the present when the concern is 10 years in the future and a prediction of n_{dv} has a high probability (e.g. 10% or 20%)

of being in error by ±10% or more. Given uncertainty in predicting the population, a model that gives estimates to around ±5% (p=.95) might be reasonable. Based on Equation 10 this less reliable model would be developed with one twenty-fifth of the data and therefore at a much lower cost. Of course, if data for the "more accurate" model must be collected to meet some other need, then the cost picture changes.

CONCLUSION

The results presented in this article show it is not a terribly difficult matter to obtain estimates of the reliability of participation or "amount" predictions that can be made using regression models. Whether one is dealing with a main effect model (Equation 1) or one considers interactions between variables, some modern regression programs give one access to the covariance matrix that is necessary to make predictions (see article's appendix). Nevertheless, given that a very low R^2 is attained in a regression, one may find that, except for estimates for a population much different than the survey population, calculating estimate variability using the sample mean proportion participating (p) gives good results (i.e., s=$(Np(1-p)/n)^{1/2}$ where N is the population "projected to" and n is the sample size). In particular, as considered in TN 12, when one is not just predicting for a population (e.g., in an area because survey data yields zero or few cases in the area) but forecasting to a future time, it is important to remember that Equation 6 applies when population sizes (n_{dv}) can be considered constant. When forecasting population contributes to variability, one approach to realistic consideration of what can be expected is to get reliability estimates for different (low, expected and high) values of the n_{dv} and use these in considering how variable estimates are really likely to be.

Not only must one be concerned about population values (n_{dv}), as already noted several times, using structurally inadequate models can cause analysis to just be "playing with numbers". Model structure should be of particular concern when estimates become heavily dependent on model parameters (e.g., because one is making estimates for segments of a population for which model parameters are large).

From a practical perspective being able to make "appropriately" reliability estimates relates to the researcher being able to determine with the planners, managers or others for whom she/he is doing research, using the general level of reliability needed for certain purposes. When data are "needed now" carrying out a survey (or other gathering any primary data) to determine what the participation level of population *is/will be* may not be an option because there is no time. If data are to guide decision making and estimates are to be made, for example, using models in secondary data, considering estimates likely reliability prior to doing a lot of modelling work is important to seeing that effort is not wasted. Regarding modelling, there is a choice to model and then one regarding how sophisticated modelling should be (e.g., ignoring interaction effects and/or supply factors (TN 29). Regardless, if using the reliability estimation methods for modelling presented here one finds that acceptably reliable estimates cannot be expected, it should be accepted that modelling should not occur. Decision should not be made using estimates that are likely to be misleading. In that regard, it should be recognized that a goal of research is showing when information is not good enough to use as well as providing quantitative input to decision making.

APPENDIX: REGRESSION FORMULATION, POPULATION VALUES AND GETTING THE COVARIANCE MATRIX TO ESTIMATE *s* FROM 2 REGRESSION PROGRAMS

The article has mentioned that how a regression is formulated determines the values of n_{dv} to use in estimating $\hat{E}(T_{\boldsymbol{A}})$ and $s = \mathrm{Var}(\hat{E}(T_{\boldsymbol{A}}))$. It has also been noted that in 2006 one can obtain Σ_β in electronic form from regression programs. This appendix deals briefly with both these matters.

In considering regression formulations, only two options are considered and only simple cases of these are introduced. Given that we are considering categorical variables each variable has 2 or more levels (e.g., see levels in Table 1). This means that for each respondent one can create dummy variables that have a value of 1 if a person has a level of an attribute and has another value if the person does not have that attribute. In fact this idea can be extended to having a given levels of 2 or more variables. For 2 levels of gender and 5 of another variable there are 10 combinations so one could have 10 indicator variables. The problem with this approach for model estimation is that knowing e.g. gender=1 means it is not 0. Scheffe (1959) introduces a constraint matrix, *H*, into analysis so that one can represent presence or absence of a level of an attribute or of a combination of levels for several attributes by zeros and ones. However, someone setting out to use regression in SPSS (1993 see e.g. actual code pp.) or SAS Proc Reg (SAS 1990, Ch. 36) is likely going to set up a regression in one of two ways illustrated in Figures 2 and 3.

In the central part of Figures 2 and 3 one sees columns in which 1 indicates the presence of an attribute. This central part is called the design matrix while on the left you see the dependent variable vector/matrix and on the right you see the parameter vector/matrix. In this regard, the figures given the matrix relation $Y=X\beta$ (Equation 1b is $Y=X\beta+\varepsilon$). The difference between the figures is that in one you see -1 in the design matrix, X. The first (left) column of the X matrix is the "General mean" column. In it you see just ones since the general man applies to everybody. Under income for respondent 1 (associated with y_1), you see 1 for income level 1 showing she/he has that level of income. Having that level means you see 0 for level 2. There is no column for level 3. Rather, for respondent 3 (associated with y_3), you see that the respondent has income level is 3 because -1 appears for levels 1 and 2. This shows that income is not in these levels. For regression results the effect of using -1 is that one is making $\beta_{\text{Income},3} = -(\beta_{\text{Income},1}+\beta_{\text{Income},2})$.

The only difference between Figure 2 and Figure 3 is that where there are -1 in Figure 2 there are zeros in Figure 3. When -1 is not used then one does not have the $\beta_{\text{Income},3} = -(\beta_{\text{Income},1}+\beta_{\text{Income},2})$ relation. This means that plots estimated regression coefficients would not be like the one in Figure 1. Instead, the regression coefficient for the missing level is 0 and other coefficients take values consistent with this.

For both figures the really important matter to note is that if a variable has *k* levels then in the design matrix only *k*-1 columns appear for the variable. Again, this is so that "the linear dependence" that arises e.g. from knowing that a person having 1 for one level of a variable means they cannot have 1 for another. In a similar way one can have columns for interactions and use 0/1/-1 for interactions. However, leaving out columns and allocating values 0/1/-1 values is more involved than for main effects. For example, if one were to extend the Figure 2-3 example to include income by urbanization interaction (3 by 3 levels), one does not just introduce 9 columns for the 3*3 combinations of levels. Nor does one use 9 minus 1 columns mimicking what was done with single variables. Readers who do not know what to do should consult a regression text.

If you want to carry out a regression, and your data has levels of variables, you can e.g. use the SAS or SPSS programming language to produce data as given in one of the figures. Your data "table"/array will include the dependent variable as a column and include the columns of the design matrix other than the general mean. You do not include that column because the regression program creates it (unless you use a special "no intercept" command to suppress its creation). Regardless, if you have the columns of the design matrix other than the general mean column, both SPSS and SAS will produce a Σ_β matrix that includes the general mean under the generic title "intercept".

Figure 2: A display showing how one can picture the "design" matrix for and dummy variable structure for a regression in which 1 indicates an attribute applies to a person and 0 that it does not apply

Participate or not (or amount)		General mean	Income		Urbanization			Age		Regression coefficients
			(1)	(2)	(1)	(2)	(1)	(2)	(3)	
y_1		1	1	0	0	1	1	0	0	μ
y_2		1	0	1	1	0	0	1	0	$\beta_{Income,1}$
y_3		1	-1	-1	-1	-1	0	0	1	$\beta_{Income,2}$
y_4		1	0	1	1	0	-1	-1	-1	$B_{Urbanization,1}$
y_5	=	1	1	0	0	1	0	1	0	$\beta_{Urbanization,2}$
y_6		1	-1	-1	-1	-1	0	0	1	$\beta_{Age,1}$
etc.		1	0	1	0	1	1	0	0	$\beta_{Age,2}$
y_{n-1}		1	1	0	1	0	-1	-1	-1	$\beta_{Age,2}$
y_n		1	-1	-1	-1	-1	0	1	0	

Figure 3: A display showing how one can picture the "design" matrix for and dummy variable structure for a regression in which 1 indicates an attribute applies to a person and -1 that it does not apply

Participate or not (or amount)		General mean	Income		Urbanization			Age		Regression coefficients
			(1)	(2)	(1)	(2)	(1)	(2)	(3)	
1		1	1	0	0	1	1	0	0	μ
1		1	0	1	1	0	0	1	0	$\beta_{Income,1}$
0		1	0	0	0	0	0	0	1	$\beta_{Income,2}$
0	=	1	0	1	1	0	0	0	0	$B_{Urbanization,1}$
0		1	1	0	0	1	0	1	0	$\beta_{Urbanization,2}$
0		1	0	0	0	0	0	0	1	$\beta_{Age,1}$
1		1	0	1	0	1	1	0	0	$\beta_{Age,2}$
0		1	1	0	1	0	0	0	0	$\beta_{Age,2}$
1		1	0	0	0	0	0	1	0	

As the discussion above has indicated, when you want to compute s=Var(Ê(T$_{\boldsymbol{A}}$)), it is important to know the nature of the X in Y=Xβ. This is because when an SPSS or SAS Proc REG is run, the Σ_β that a computer creates is for the coefficients for the "model" specified. In other words, for Figures 2 and 3 either $\beta_{Income,3}$= -($\beta_{Income,1}$+$\beta_{Income,2}$) applies or the coefficient for the missing level is zero.

If your data are weighted and you use the weight in the regression (use a WEIGHT statement), the weights are properly incorporated into Σ_β by SPSS and SAS regression programs. Specifically consider using SAS PROC REG and "linear regression" as accessed using the menu features in SPSS 10 or later. Given that you get data into the form for one of these programs, to estimate *s* you face four challenges. One is getting the Σ_β in a usable electronic form. For SAS PROC REG in the "PROC" statement you specify "OUTEST={SOME DATA SET NAME}" and also specify COVOUT. In SPSS10 and above, using the ANALYZE "tab" you can select Regression and "subselect" Linear. In linear regression you set up your regression and use the SAVE button to allow you to specify the data set into which you want the information stored that lets you access Σ_β. To use Σ_β, even when you have it in electronic form, is not trivial. Both SPSS and SAS offer matrix processing options. Your second challenge is using SAS/SPSS programming and the matrix language to get Σ_β^{-1}. You must perform the inverse operation.

To execute the Equation 6 multiplication you need the *n* matrix. Having it in electronic form requires further SPSS/SAS programming. If you chose the Figure 2 model setup, then you need to calculate $n_{variable,level}$ for the computation by finding the actual population in the variable-level category, $\check{n}_{variable,level}$, and also finding the actual population in the level of the variable that is excluded. $n_{variable,level}$ is the difference as expressed in Equation 10. The reason you would be using the Figure 2 version of regression is so that you can easily get plots like Figure 1. You do this based on computing the effects associated with missing levels by Equation 11. If you use the Figure 3 version of regression formulation, then $n_{variable,level}$ is exactly what seems apparent, the size of the population in the given level of the variable (e.g. number of people in age group 1 is $n_{age,1}$.

Equation 10: $n_{variable,level} = \check{n}_{variable,level} - \check{n}_{variable,level\ excluded}$
Equation 11: Effect missing level=-Σ(effects of other levels)

Given that you specify the *n* matrix/vector, your fourth and final challenge is using the matrix programming language to actually calculate $Var(\hat{E}(T_A)) = n'\hat{\Sigma}_\beta^{-1}n$ (Equation 6).

What has been described is not really complicated. If implemented with support of someone who is good at SPSS/SAS, getting "generic" code going is easy. With generic code, modifying the code for new cases is trivial when one just changes number of variables and their levels. If one starts to move from main effect models to more elaborate models, then the biggest challenge may be learning to correctly program the design matrix (X in Y=Xβ).