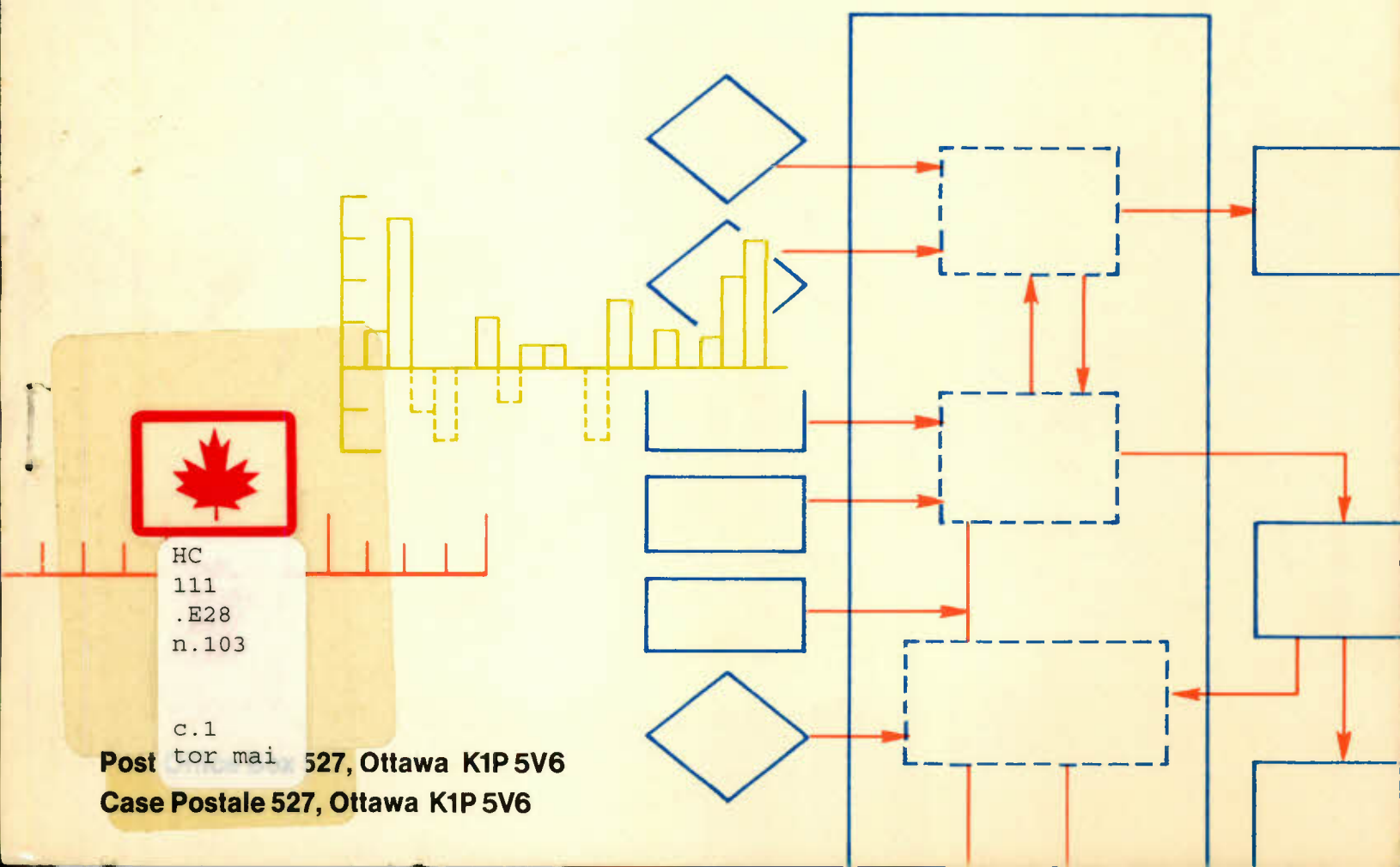
Economic Council of Canada
Conseil économique du Canada



Post [illegible] 527, Ottawa K1P 5V6
Case Postale 527, Ottawa K1P 5V6

DISCUSSION PAPER NO. 103

A Technique For Efficient Estimation Using Grouped Micro-Data

by Jac-André Boulet
and Paul Robillard

Discussion Papers are working documents available by the Economic Council of Canada, in limited number and in the language of preperation, to interested individuals for the benefit of their professional comments.

December 1977

Catalogue No. EC25-103/1977
ISBN 0-662-01444-8

TABLE OF CONTENTS

Abstract*

The proliferation of survey and data collecting organizations in many countries, and the ever-increasing number of samples containing more and more information about the individuals or organizations being queried has been accompanied by two types of problems involving the treatment of information thus compiled. Treatment of this considerable information with the usual research tools, such as multiple regression analysis, incurs very high computer costs, reaching proportions that in many cases threaten to quickly devour funds budgeted for analysis. Further, it sometimes proves difficult to use these instruments because the compiling agency, for reasons of confidentiality, does not wish to divulge the microeconomic information gathered on each of the persons or organizations queried.

The following paper proposes an approach to multiple regression analysis that avoids these two problems. The results obtained by applying this approach are in every way identical to those that would be obtained through the usual approaches to the multiple regression analysis of microeconomic data. Not only does this new method considerably reduce computer costs but it makes it possible to use micro-data without breaking the rules of confidentiality, and permits data already available from data-gathering bodies, in aggregate form, to be employed directly in analysis. The algorithm proposed in this document is an integral part of a multiple regression program now in use at the Economic Council.

---

*We wish especially to thank David W. Henderson and John Eden Cloutier for their numerous appropriate suggestions with respect to the contents of this document and also to express our gratitude to J.C. Robin Rowley for his helpful comments. We also wish to acknowledge the technical contribution of J.A. Aurèle Leduc, computer analyst at the Council.

Résumé*

Avec la multiplication d'organismes enquêteurs dans de nombreux pays et l'arrivée sur le marché d'échantillons de plus en plus volumineux et renfermant de plus en plus d'information sur les personnes ou organismes interrogées sont venus se greffer deux types de problème touchant le traitement des informations ainsi recueillies. Si l'on désire traiter ces masses d'information à l'aide d'instruments usuels aux chercheurs, comme la régression multiple par exemple, il s'ensuit des coûts d'ordinateur très élevés qui risquent dans bien des cas d'atteindre des proportions qui hypothèquent grandement les budgets consentis à l'analyse. En outre il s'avère parfois impossible d'utiliser ces instruments parce que l'organisme enquêteur ne veut pas consentir à divulguer, pour des raisons de confidentialité, les information micro-économiques telles que recueillies sur chacune des personnes ou organismes interrogées.

Dans le document qui suit les auteurs proposent une façon d'éviter ces deux problèmes dans le cas de la régression multiple. Les résultats obtenus, par la régression multiple, en appliquant leurs suggestions sont en tout point identiques à ceux que l'on obtiendrait en procédant à l'aide de l'approche usuelle, c'est-à-dire à l'aide des données micro-économiques. Enfin non seulement cet algorithme de calcul réduit-il considérablement les coûts d'ordinateur et rend-il possible l'utilisation de micro données, sans violer en rien les règles de la confidentialité, mais de plus il ouvre les portes à une exploitation nouvelle des données déjà publiées sous forme agrégée par les organismes enquêteurs tout en produisant des résultats identiques à ceux que nous aurions obtenus en procédant à l'aide des données micro-économiques. L'algorithme est déjà en usage au Conseil.

---

* Nous tenons à remercier tout particulièrement MM. David W. Henderson et John Eden Cloutier pour leurs nombreuses suggestions et leurs commentaires fort appropriés de même aussi que M. J.C. Robin Rowley. Nous voudrions enfin souligner la contribution technique de J.A. Aurèle Leduc, informaticien au Conseil.

Introduction

A great many economic studies today use multiple regression analysis. There has been a notable growth in the number of surveys over the past two decades and in the volume of census data, and researchers have employed multiple regression techniques as one of the important means of approaching the analysis of this information. In analyzing this material and other similar sorts of data, researchers tend generally to favour the use of micro-economic rather than regrouped data, because of the greater precision in the estimates and the increased power of discrimination among the various specifications that one can obtain from the former. This paper will demonstrate that it is possible in a number of cases to work with grouped data and obtain estimates that are in every way identical to those that would have been obtained if conventional techniques had been applied to the underlying micro-economic data. Further, this paper will show that the costs of carrying out such an analysis are much less than the costs would be if one employed conventional techniques.

The main thrust of this paper can be summarized as follows:

1) Given a set of observations for a variable, this variable being defined as a dependent variable in a particular analytical context, and for a number of potentially related independent variables, each of which is divided into a series of categories such that each observation falls into one category and one category only for each of these variables (i.e., the independent variables are dichotomous in form), the moment matrix [X:Y]'[X:Y]

can be constructed employing conventional techniques and multiple regression analysis performed using the matrix. However, employing the algorithm proposed in this document,[1] it is possible to construct the identical moment matrix with only the following information: the tabulation of pair-wise joint frequencies (i.e., the frequency cross-tabulation)[2] for the cross-tabulated categories of the independent variables, and the mean value of the dependent variable for each category of each independent variable.[3] Hence, the moment matrix can be obtained from grouped data, without knowledge of the underlying microeconomic data.

---

1. This algorithm is particularly useful when the independent variables are all expressed in dichotomous form. However, this approach is also applicable if one or more (but not all) of the independent variables are in continuous form. It is preferable to keep the number of continuous variables to a minimum in using this approach, as it loses its simplicity and becomes more complex with increases in the number of continuous variables. However, it should be noted that if *all* the independent variables are in continuous form, multiple regression analysis can be undertaken using the familiar approach in which the variance-covariance matrix is employed. We will return to the question of continuous variables later in the text.

2. A tabulation of pair-wise joint frequencies is very important if not virtually indispensible, in studies employing multivariate analysis. It is used, for example, to ensure that there are sufficient observations within various cross-categories of the independent variables so that the estimates will be statistically reliable, or simply to ensure that the choice of categories for each of the different independent variables is a reasonable one for the analysis being undertaken. Hence, it is often calculated as a matter of course in carrying out a conventional multiple regression analysis.

3. The mean of the dependent variable for each of the *off-diagonal* elements in a cross-classification of all categories is not required.

2) Because of the question of confidentiality of survey data collected by organizations such as Statistics Canada, researchers often cannot obtain direct access to the detailed micro-data in order to carry out their analyses. The algorithm presented in this document permits researchers to carry out efficient and unbiased multivariate analysis from the grouped data (for which, of course, the problems of confidentiality do not exist). That is, they can construct the same moment matrix and carry out the same multiple regression analyses that they can perform with the underlying micro-data using conventional techniques. Further, frequency cross-tabulations for various sets of data are sometimes published by the survey organization, which smooths the path even more for the application of this technique.

3) The use of the algorithm proposed in this document reduces, to a considerable extent, the costs of constructing the moment matrix. Using this approach, the costs of constructing the moment matrix would rarely exceed two digit dollar amounts, whereas the costs can sometimes reach the four digit dollar level employing the conventional approach -- depending, of course, on the size of the sample, the number of independent variables, and the number of categories into which these independent variables are divided.

Certain authors have already proposed methods for estimating multiple regression models using grouped data. A few examples are the works of Prais and Aitchison (1954),

Haitovsky (1966, 1973) and the latter's reference to an unpublished text by Houthakker. While these methods are of some assistance in those cases where only grouped data are available, they unfortunately provide estimates that are not as efficient as those which could be obtained from the underlying micro-data.[1] On the other hand, as mentioned above, there is no loss of efficiency in employing the algorithm proposed in this document.

Finally, it should be noted that the algorithm proposed in this document produces regression coefficients for which the weighted sum of the coefficients for all the categories of a particular independent variable is zero.[2] This constraint imposed on the regression coefficients does not result in any loss of generality.

To sum up, in the following text we propose a method for computing a moment matrix from which the coefficients for the various categories of the independent variables can be obtained by least squares estimates -- these coefficients being constrained in the particular fashion noted above. We will demonstrate that the estimates obtained by this approach are identical to those that would be provided through direct access to the micro-data and the use of conventional techniques. We will also show that this approach reduces costs considerably

---

1. See Cramer (1964), and Orcutt, Watts, and Edwards (1968).

2. Weighted by the number of observations in each category (see Boulet (1975)).

relative to the conventional approach, these costs reductions increasing with the size of the sample and the total number of categories for all the independent variables. We will illustrate the cost reduction aspect with an example derived from earlier research work, and then, in order to illustrate in a general fashion the possibilities offered by this algorithm, we will present a simple example that applied this technique to a set of grouped data which has been published by Statistics Canada. These grouped data were obtained by Statistics Canada from observations with respect to 11 million individuals.

Section 1: THE PROPOSED METHOD AND ITS PROPERTIES

(a) Defining the Problem

Let Y be a dependent variable which is a function of s explanatory factors (variables) $X_1, X_2, \ldots, X_s$. Suppose that each factor X_j may be divided into k_j mutually exclusive and collectively exhaustive categories $X_{j1}, X_{j2}, \ldots, X_{jk_j}$; (j=1, 2, ..., s). Suppose that we have N observations, and define the t^{th} observation on the i^{th} category of the j^{th} factor as X_{jit}.

Thus $X_j = (X_{j1}, X_{j2}, \ldots, X_{jk_j})$ $\forall$ j = 1, 2, ..., s is an N x k_j matrix;

$X_{ji} = (X_{ji1}, X_{ji2}, \ldots, X_{jiN})'$ $\forall$ j = 1, 2, ..., s and i = 1, 2, ..., k_j is a column vector of dimension N;

X_{jit} is a scalar variable that takes on the value 1 if the t^{th} observation on the j^{th} factor corresponds to the i^{th} category of the factor, and the value 0 otherwise.

$$\text{Then } \sum_{i=1}^{k_j} X_{jit} = 1 \quad \forall\ j = 1, 2, \ldots, s; \text{ and}$$
$$t = 1, 2, \ldots, N \tag{1.1}$$

$$\text{and} \quad X_{jit}X_{jgt} = 0 \quad \forall\ j = 1, 2, \ldots, s;$$
$$t = 1, 2, \ldots, N;$$
$$i,g = 1, 2, \ldots, k_j,\ i \neq g \tag{1.2}$$

For expositional clarity we maintain the constant term of the equation separate from the s explanatory factors. Thus define $c = (c_1, c_2, \ldots, c_N)'$

where $c_t = 1 \quad \forall\ t = 1, 2, \ldots, N$

The equation to be estimated therefore takes the form:

$$Y_t = \alpha_0 c_t + \sum_{i=1}^{k_1} \alpha_{1i} X_{1it} + \sum_{i=1}^{k_2} \alpha_{2i} X_{2it} + \ldots$$
$$+ \sum_{i=1}^{k_s} \alpha_{si} X_{sit} + u_t, \quad (t = 1, 2, \ldots, N)$$
$$= \alpha_0 c_t + \sum_{j=1}^{s} \sum_{i=1}^{k_j} \alpha_{ji} X_{jit} + u_t \tag{1.3}$$

where $\alpha_0, \alpha_{11}, \alpha_{12}, \ldots, \alpha_{1k_1}, \alpha_{21}, \ldots, \alpha_{sk_s}$ are the coefficients to be estimated; and $u = (u_1, u_2, \ldots, u_N)'$ is distributed normal $(0, \sigma^2 I_N)$. The rank of the matrix of observations in dichotomous form that appears in (1.3) follows from (1.1) and (1.2) as, $k_1 + k_2 + \ldots + k_s + 1 - s$.[1] Since there are $k_1 + \ldots + k_s + 1$ terms

---

1. It is assumed that the number of observations is greater than the number of parameters to be estimated. Since the categories for any particular variable X_j are mutually exclusive and collectively exhaustive, the rank of the observations on the k_j categories will be $k_j - 1$. This is true for all s variables. In addition to the s variables there is also a constant term.

to be estimated, we require s contraints on (1.3) for the estimation to be unique.

Let the constraints on (1.3) be given by:

$$\sum_{i=1}^{k_j} \alpha_{ji} x_{ji} = 0 \quad \forall\ j = 1, 2, \ldots, s \tag{1.4}$$

where $$x_{ji} = \sum_{t=1}^{N} X_{jit} \quad \forall\ j = 1, 2, \ldots, s; \text{ and } i = 1, 2, \ldots, k_j$$

That is, x_{ji} is the number of observations in the i^{th} category of factor j.

It is easily demonstrated that (1.4) implies that the estimated constant term is equal to the mean of the dependent variable.[1] Thus,

$$\hat{\alpha}_0 = \sum_{t=1}^{N} Y_t / N \equiv \bar{Y} \tag{1.5}$$

The remaining coefficients α_{ji} may be estimated by[2]

$$\hat{\alpha}_{ji} = \bar{Y}_{ji} - \bar{Y} - \frac{1}{x_{ji}} \left[\sum_{\substack{m=1 \\ m \neq j}}^{s} \sum_{n=1}^{k_m} \hat{\alpha}_{mn} \sum_{t=1}^{N} X_{jit} X_{mnt} \right] \tag{1.6}$$

where $\bar{Y}_{ji}$ is the mean of the dependent variable Y for the subpopulation corresponding to the occurrence $X_{jit} = 1$; $t = 1, 2, \ldots, N$. That is

$$\bar{Y}_{ji} = \sum_{t=1}^{N} \frac{Y_t X_{jit}}{x_{ji}}$$

---

1. See Scheffé (1959), p. 100.

2. See Scheffé (1959), pp. 113-114, and Boulet (1975), pp. 15-16.

Information of considerable importance can be drawn from the estimation of (1.6) and the parallel expressions for the other coefficients. In (1.6), $\hat{\alpha}_{ji}$ can be considered as the difference between the mean of the dependent variable calculated for the sub-population i of X_j and the mean of the dependent variable calculated over the population as a whole, when this difference is adjusted for structural differences between the sub-population i of X_j and the population as a whole. This adjustment is represented by the third term on the right-hand side of (1.6).[1] In other words, $\hat{\alpha}_{ji}$ represents the difference that would exist between the mean of the dependent variable for the sub-population i of X_j and the mean of the dependent variable for the whole population if the i^{th} sub-population were adjusted to reflect the mean effect of all other factors except X_j.

Estimation employing the constraints (1.4) was preferred to the more common method of omitting a category for each variable[2] since we are interested in comparing the coefficients of two or more equations of the form (1.3).[3] Although the coefficients obtained by either method may be interpreted as differences,

---

1. For a more complete discussion of this procedure, see Boulet (1975).

2. See Johnston (1972), pp. 176-181.

3. In using this technique employing constraints, a category is actually omitted for each variable in estimating the equations, but the constraints permit the direct estimation of the coefficients of the omitted categories in a residual fashion.

the base for the difference using constrained estimation is $\bar{Y}$, and hence is invariant from equation to equation. In the case where categories are omitted, the base for the difference is composed of the sum of the coefficients of the categories omitted and the constant term, and thus may vary from equation to equation. Employing the constraints (1.4) helps to facilitate the analysis of the results in the sense that the meaning of the coefficients is more easily indentifiable and the results obtained are skew-symmetric and transitive.[1]

(b) A New Method For Calculating The Moment Matrix

Imposing the constraints (1.4) on Equation (1.3) is equivalent to considering the following equation:

$$Y_t = \alpha_0 c_t + \sum_{i=1}^{k_1^*} \alpha_{1i}\left(X_{1it} - \frac{x_{1i}X_{1k_1t}}{x_{1k_1}}\right)$$

$$+ \sum_{i=1}^{k_2^*} \alpha_{2i}\left(X_{2it} - \frac{x_{2i}X_{2k_2t}}{x_{2k_2}}\right) + \ldots.$$

$$+ \sum_{i=1}^{k_s^*} \alpha_{si}\left(X_{sit} - \frac{x_{si}X_{sk_st}}{x_{sk_s}}\right) + u_t$$

$$= \alpha_0 c_t + \sum_{j=1}^{s} \sum_{i=1}^{k_j^*} \alpha_{ji}\left(X_{jit} - \frac{x_{ji}X_{jk_jt}}{x_{jk_j}}\right) + u_t \quad (1.7)$$

---

1. See Boulet and Rowley (1977).

where $k_j^* = k_j - 1$, for $j = 1, 2, \ldots, s$. We may assume, without loss of generality, that the excluded categories are $k_1, k_2, \ldots, k_s$ for factors $X_1, X_2, \ldots, X_s$ respectively. The parameters for these categories, $\alpha_{1k_1}, \alpha_{2k_2}, \ldots, \alpha_{sk_s}$, which we will call the "reference parameters", can be estimated using the constraints (1.4) as:

$$\hat{\alpha}_{jk_j} = -\sum_{i=1}^{k_j^*} \frac{x_{ji}}{x_{jk_j}} \hat{\alpha}_{ji} \qquad (1.8)$$

When equation (1.7) is estimated by ordinary least squares and the reference parameters by (1.8) the estimated reference parameters have the same properties as the parameters (coefficients) estimated by (1.7), as can be demonstrated by use of the Gauss-Markov theorem.[1]

The matrix of observations for the explanatory variables of (1.7) (less the excluded category for each variable) is written as $\mathbb{Z}$ and can be defined as follows:

$$\mathbb{Z} = [c, \mathbb{Z}_{11}, \mathbb{Z}_{12}, \ldots, \mathbb{Z}_{1k_1^*}, \mathbb{Z}_{21}, \mathbb{Z}_{22}, \ldots, \mathbb{Z}_{2k_2^*}, \ldots, \mathbb{Z}_{s1}, \mathbb{Z}_{s2}, \ldots, \mathbb{Z}_{sk_s^*}] \qquad (1.9)$$

$$\text{where } \mathbb{Z}_{jit} = X_{jit} - \frac{x_{ji}X_{jk_jt}}{x_{jk_j}} \qquad \begin{matrix} j = 1, 2, \ldots, s \\ i = 1, 2, \ldots, k_j^* \\ t = 1, 2, \ldots, N \end{matrix} \qquad (1.10)$$

$$\text{and } \mathbb{Z}_{ji} = (\mathbb{Z}_{ji1}, \mathbb{Z}_{ji2}, \ldots, \mathbb{Z}_{jiN})' \qquad \begin{matrix} j = 1, 2, \ldots, s \\ i = 1, 2, \ldots, k_j^* \\ t = 1, 2, \ldots, N \end{matrix} \qquad (1.11)$$

---

1. See Johnston (1972), pp. 126-127.

If the corresponding parameter vector, say α, is defined as

$$\alpha = (\alpha_0, \alpha_{11}, \alpha_{12}, \cdots, \alpha_{1k_1^*}, \alpha_{21}, \alpha_{22}, \cdots, \alpha_{2k_2^*}, \cdots, \alpha_{s1}, \alpha_{s2}, \cdots, \alpha_{sk_s^*})' \qquad (1.12)$$

We may re-write (1.7) in the conventional matrix form

$$Y = \mathbb{Z}\alpha + u$$

and estimate the vector of coefficients by ordinary least squares as

$$\hat{\alpha} = (\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'Y$$

It can be shown that once the moment matrix $\Omega = [\mathbb{Z}:Y]'[\mathbb{Z}:Y]$ is obtained, one can proceed directly to the estimation of $\hat{\alpha}$. The form of the moment matrix is given by (1.13).

$$\Omega = \begin{bmatrix} \mathbb{Z}' \\ Y' \end{bmatrix} [\mathbb{Z}:Y] = \begin{bmatrix} \mathbb{Z}'\mathbb{Z} & \mathbb{Z}'Y \\ Y'\mathbb{Z} & Y'Y \end{bmatrix} = \text{(see following page)} \qquad (1.13)$$

$c'c$	$c'g_{11}$	$c'g_{12}$	$\cdots$	$c'g_{1k_1^*}$	$c'g_{21}$	$c'g_{22}$	$\cdots$	$c'g_{2k_2^*}$	$\cdots$	$c'g_{s1}$	$c'g_{s2}$	$\cdots$	$c'g_{sk_s^*}$	$c'Y$
$g_{11}'c$	$g_{11}'g_{11}$	$g_{11}'g_{12}$	$\cdots$	$g_{11}'g_{1k_1^*}$	$g_{11}'g_{21}$	$g_{11}'g_{22}$	$\cdots$	$g_{11}'g_{2k_2^*}$	$\cdots$	$g_{11}'g_{s1}$	$g_{11}'g_{s2}$	$\cdots$	$g_{11}'g_{sk_s^*}$	$g_{11}'Y$
$g_{12}'c$	$g_{12}'g_{11}$	$g_{12}'Z_{12}$	$\cdots$	$g_{12}'g_{1k_1^*}$	$g_{12}'g_{21}$	$g_{12}'g_{22}$	$\cdots$	$g_{12}'g_{2k_2^*}$	$\cdots$	$g_{12}'g_{s1}$	$g_{12}'g_{s2}$	$\cdots$	$g_{12}'g_{sk_s^*}$	$g_{12}'Y$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$
$g_{1k_1^*}'c$	$g_{1k_1^*}'g_{11}$	$g_{1k_1^*}'g_{12}$	$\cdots$	$g_{1k_1^*}'g_{1k_1}$	$g_{1k_1^*}'g_{21}$	$g_{1k_1^*}g_{22}$	$\cdots$	$g_{1k_1^*}'g_{2k_2}$	$\cdots$	$g_{1k_1^*}'g_{s1}$	$g_{1k_1^*}'g_{s2}$	$\cdots$	$g_{1k_1^*}'g_{sk_s^*}$	$g_{1k_1^*}'Y$
$g_{21}'c$	$g_{21}'g_{11}$	$g_{21}'g_{12}$	$\cdots$	$g_{21}'g_{1k_1^*}$	$g_{21}'g_{21}$	$g_{21}'g_{22}$	$\cdots$	$g_{21}'g_{2k_2^*}$	$\cdots$	$g_{21}'g_{s1}$	$g_{21}'g_{s2}$	$\cdots$	$g_{21}'g_{sk_s^*}$	$g_{21}'Y$
$g_{22}'c$	$g_{22}'g_{11}$	$g_{22}'g_{12}$	$\cdots$	$g_{22}'g_{1k_1^*}$	$g_{22}'g_{21}$	$g_{22}'g_{22}$	$\cdots$	$g_{22}'g_{2k_2^*}$	$\cdots$	$g_{22}'g_{s1}$	$g_{22}'g_{s2}$	$\cdots$	$g_{22}'g_{sk_s^*}$	$g_{22}'Y$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$
$g_{2k_2^*}'c$	$g_{2k_2^*}'g_{11}$	$g_{2k_2^*}'g_{12}$	$\cdots$	$g_{2k_2^*}'g_{1k_1^*}$	$g_{2k_2^*}'g_{21}$	$g_{2k_2^*}'g_{22}$	$\cdots$	$g_{2k_2^*}'g_{2k_2^*}$	$\cdots$	$g_{2k_2^*}g_{s1}$	$g_{2k_2^*}'g_{s2}$	$\cdots$	$g_{2k_2^*}'g_{sk_s}$	$g_{2k_2^*}'Y$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$g_{s1}'c$	$g_{s1}'g_{11}$	$g_{s1}'g_{12}$	$\cdots$	$g_{s1}'g_{1k_1^*}$	$g_{s1}'g_{21}$	$g_{s1}'g_{22}$	$\cdots$	$g_{s1}'g_{2k_2^*}$	$\cdots$	$g_{s1}'g_{s1}$	$g_{s1}'g_{s2}$	$\cdots$	$g_{s1}'g_{sk_s^*}$	$g_{s1}'Y$
$g_{s2}'c$	$g_{s2}'g_{11}$	$g_{s2}'g_{12}$	$\cdots$	$g_{s2}'g_{1k_1^*}$	$g_{s2}'g_{21}$	$g_{s2}'g_{22}$	$\cdots$	$g_{s2}'g_{2k_2^*}$	$\cdots$	$g_{s2}'g_{s1}$	$g_{s2}'g_{s2}$	$\cdots$	$g_{s2}'g_{sk_s^*}$	$g_{s2}'Y$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$
$Z_{sk_s^*}'c$	$g_{sk_s^*}'g_{11}$	$g_{sk_s^*}'g_{12}$	$\cdots$	$g_{sk_s^*}'g_{1k_1^*}$	$g_{sk_s^*}'g_{21}$	$g_{sk_s^*}'g_{22}$	$\cdots$	$g_{sk_s^*}'g_{2k_2^*}$	$\cdots$	$g_{sk_s^*}'g_{s1}$	$g_{sk_s^*}'g_{s2}$	$\cdots$	$g_{sk_s^*}'g_{sk_s^*}$	$g_{sk_s^*}'Y$
$Y'c$	$Y'g_{11}$	$Y'g_{12}$	$\cdots$	$Y'g_{1k_1^*}$	$Y'g_{21}$	$Y'g_{22}$	$\cdots$	$Y'g_{2k_2^*}$	$\cdots$	$Y'g_{s1}$	$Y'g_{s2}$	$\cdots$	$Y'Z_{sk_s^*}$	$Y'Y$

where the matrix Ω is symmetric.

Within this matrix, there are four basic types of products one involving the vector c, a second involving two categories of the same factor, a third involving two categories of different factors and the fourth involving the vector Y. For those involving c, it can be shown that:

$$c'c = N \tag{1.14}$$

$$Z_{ji}'c = 0 \quad j = 1, 2, \ldots, s; \; i = 1, 2, \ldots, k_j^* \tag{1.15}$$

$$Y'c = N \cdot \bar{Y} \tag{1.16}$$

For those products involving two categories, i and n, of the same factor, j, it can be demonstrated for all j that:

$$Z_{ji}'Z_{jn} = \begin{cases} \dfrac{x_{ji}x_{jn}}{x_{jk_j}}; \; i \neq n & \\ & i, n = 1, 2, \ldots, k_j^* \\ x_{ji} + \dfrac{x_{ji}^2}{x_{jk_j}}; \; i = n & \end{cases} \tag{1.17, 1.18}$$

For those products involving two categories, i and n, of different factors, j and m, we have that:

$$Z_{ji}'Z_{mn} = (x_{ji} \odot x_{mn}) - (x_{jk_j} \odot x_{mn}) \cdot \left(\frac{x_{ji}}{x_{jk_j}}\right) - (x_{ji} \odot x_{mk_m}) \cdot \left(\frac{x_{mn}}{x_{mk_m}}\right) + (x_{jk_j} \odot x_{mk_m}) \cdot \left(\frac{x_{ji}x_{mn}}{x_{jk_j}x_{mk_m}}\right) \tag{1.19}$$

were $x_{ji} \odot x_{mn} = \sum_{t=1}^{N} X_{jit}X_{mnt}$

and $j, m = 1, 2, \ldots, s; \quad j \neq m$

$i = 1, 2, \ldots, k_j^*$

$n = 1, 2, \ldots, k_m^*$

Finally, for the vector Y, it can be shown that

$$Y'g_{ji} = (\bar{Y}_{ji} - \bar{Y}_{jk_j})x_{ji} \quad \begin{matrix} j = 1, 2, \ldots, s \\ i = 1, 2, \ldots, k_j^* \end{matrix} \qquad (1.20)$$

and $$Y'Y = \sum_{t=1}^{N} Y_t^2 \qquad (1.21)$$

It is apparent from the above expressions that to construct a moment matrix, so as to be able to estimate the regression coefficients of an equation where the independent variables are presented in dichotomous form, it is only necessary to know the mean value of Y for each category of each of the independent variables, the number of observations (i.e., the frequency) for each of these categories, and, finally, the frequency for each of the off-diagonal cells in a cross-tabulation of all the categories of all independent variables. In short, no microeconomic information is required to construct the moment matrix and estimate the regression coefficients of a particular equation. The t-statistics, F-statistics and the coefficient of multiple determination (R^2) can be obtained either from $\sum_{t=1}^{N} Y_t^2$ or the variance $(\sum_{t=1}^{N} (Y_t^2 - \bar{Y}^2)/N)$, information which should be readily available from the organization gathering the data.

In the case where certain (but not all) of the independent variables are continuous, the following additional information is required in order to construct the moment matrix: the average value of these variables for each category of those independent variables that are in dichotomous form; the variance for these continuous variables; the covariance between each of the continuous variables; and, finally, the covariance between each

of the continuous independent variables and the dependent variable (the latter three items are aspects of the variance/covariance matrix for the dependent variable and the continuous independent variables).[1]

Section 2: ILLUSTRATIONS OF THE METHOD'S ADVANTAGES

(a) An Example of the Extent of Cost Reductions

Considerable research on earnings disparities between the various ethnic and linguistic groups in Montreal in 1961 and 1971 has been carried out at the Economic Council of Canada.[2] In doing this work, it was found necessary to make a particularly intensive use of multivariate analysis. The samples employed, drawn from Census data, contained 101,708 male earners for 1961 and 227,678 for 1971. For each individual, detailed information was available with respect to their earnings, ethnic origin, language spoken, education, length of employment per annum, occupation, age, citizenship status, marital status, and whether they were self-employed or worked for others. Each of the nine independent variables (earnings being the dependent variable) was subdivided into several categories in such a way that the requirements of the analytical approach employed were met. This resulted in a total of 68 categories which implies a 70 x 70 moment matrix.[3]

---

1. See footnote 2, p. 23.

2. See Boulet (1975), Boulet and Raynauld (1977), and J-A. Boulet, "Trends in Earnings Disparities According to Ethnic Origin and Language in Montreal, 1961 to 1971", Economic Council of Canada (forthcoming).

3. The figure 70 comes from the 68 categories plus the constant term and the dependent variable.

In pursuing the conventional approach, using a program specially designed to handle large samples,[2] more than 11,000 seconds of central processing unit time on an IBM 370/168 computer were required to obtain the results shown in Table 1 -- that is, to construct the moment matrix from the detailed data on the 227,678 workers in the 1971 sample, and to estimate the regression coefficients, t-statistics, R^2, and F-statistic. Given the frequency cross-tabulation, the mean of the dependent variable for each category of the independent variables, and the variance (or $\sum_{t=1}^{N} Y_t^2$), the identical results can be obtained on the same computer in less than 7 seconds of central processing unit time (see the last two columns in Table 1) -- using the algorithm proposed in this document. Thus, the technique proposed in this document provides unbiased and efficient estimates using grouped data (i.e., the same results that would have been obtained using the underlying microdata).

---

1. This program, known as program INFINITE, was developed by the Statistical Methodology and Procedure Section, Division of Research and Statistics, Board of Governors of the Federal Reserve System, Washington, D.C.. This program, as modified at the Council, has been employed (without the algorithm) in other work besides the research noted above on earnings disparities between the various ethnic and linguistic groups in Montreal in 1961 and 1971. See, for example, Lacroix, Lemelin, and Robillard (1977).

Table 1

REGRESSION COEFFICIENTS FOR CERTAIN IMPORTANT DETERMINANTS OF EARNINGS, MALE WORKERS, MONTREAL METROPOLITAN AREA, 1971[1]

Independent Variable	Category	Using The Conventional Approach		Using The Proposed Algorithm	
		Coefficients	t-Statistics	Coefficients	t-Statistics
Weeks of Employment	1-13 Weeks	- 3,956.46	115.23	- 3,956.46	115.23
	14-26 Weeks	- 3,078.83	103.44	- 3,078.83	103.44
	27-39 Weeks	- 1,684.08	62.62	- 1,684.08	62.62
	40-48 Weeks	+ 43.62	2.12	+ 43.62	2.12
	49-52 Weeks	+ 869.79	142.28	+ 869.79	142.28
Age (Years)	15-19	- 1,627.73	40.28	- 1,627.73	40.28
	20-24	- 1,623.59	69.95	- 1,623.59	69.95
	25-29	- 804.46	40.76	- 804.46	40.76
	30-34	+ 267.77	12.24	+ 267.77	12.24
	35-39	+ 956.85	42.91	+ 956.85	42.91
	40-44	+ 1,146.42	50.79	+ 1,146.42	50.79
	45-49	+ 1,090.60	45.23	+ 1,090.60	45.23
	50-54	+ 774.11	28.36	+ 774.11	28.36
	55-59	+ 414.40	13.88	+ 414.40	13.88
	60-64	- 90.45	2.50	- 90.45	2.50
	65 or more	- 1,305.44	27.95	- 1,305.44	27.95
Period of Immigration	Born in Canada	+ 112.46	18.49	+ 112.46	18.49
	Immigrated before 1946	+ 301.51	5.96	+ 301.51	5.96
	Immigrated after 1946	- 547.80	21.41	- 547.80	21.41
Education	Primary or less	- 1,078.21	67.09	- 1,078.21	67.09
	Secondary, 1-2 years	- 533.39	32.05	- 533.39	32.05
	Secondary, 3-5 years	+ 192.78	16.35	+ 192.78	16.35
	Some University	+ 441.04	14.62	+ 441.04	14.62
	University Graguate	+ 2,124.46	85.95	+ 2,124.46	85.95
Occupation	Administrator	+ 4,986.29	146.77	+ 4,986.29	146.77
	Engineer	+ 1,338.10	30.42	+ 1,338.10	30.42
	Scientist	+ 20.85	0.20	+ 20.85	0.20
	Teacher	+ 139.04	2.84	+ 139.04	2.84
	Doctor, Dentist, etc.	+ 7,612.90	79.36	+ 7,612.90	79.36
	Optometrist, Pharmacist, etc.	+ 1,986.58	11.97	+ 1,986.58	11.97
	Magistrate, Lawyer, etc.	+ 5,277.77	43.31	+ 5,277.77	43.31
	Architect	+ 3,133.61	13.95	+ 3,133.61	13.95
	Accountant, Economist, etc.	+ 1,039.08	23.69	+ 1,039.08	23.69
	Other Professionals	- 573.64	14.94	- 573.64	14.94
	Clerk	- 967.64	46.34	- 967.64	46.34
	Salesperson	- 149.71	7.16	- 149.71	7.16
	Security Personnel	+ 18.05	0.43	+ 18.05	0.43
	Other Services	- 1,627.88	56.78	- 1,627.88	56.78
	Transportation Supervision	+ 399.73	3.14	+ 399.73	3.14
	Pilot	+ 7,415.93	31.86	+ 7,415.93	31.86
	Other Transportation	- 1,059.13	34.03	- 1,059.13	34.03
	Communications Supervisor	+ 2,492.25	5.53	+ 2,492.25	5.53
	Other Communiations	+ 231.95	1.21	+ 231.95	1.21
	Farmers, Fisherman, etc.	- 1,005.79	13.51	- 1,005.79	13.51
	Skiller Workers (Tradesmen)	- 236.91	18.68	- 236.91	18.68
	Unskilled Workers	- 660.15	13.62	- 660.15	13.62
Language	Unilingual Francophones	- 286.59	14.94	- 286.59	14.94
	Unilingual Anglophones	+ 682.03	20.85	+ 682.03	20.85
	Unilingual Allophones	- 790.89	9.18	- 790.89	9.18
	Bilingual Francophones	- 84.07	6.12	- 84.07	6.12
	Bilingual Anglophones	+ 507.75	17.14	+ 507.75	17.14
	English-Speaking Allophones	- 286.66	6.06	- 286.66	6.06
	French-Speaking Allophones	- 382.75	5.52	- 382.75	5.52
	Bilingual Allophones	+ 20.81	0.49	+ 20.81	0.49
Ethnic Origin	French	- 70.00	5.59	- 70.00	5.59
	English-Scottish	+ 201.91	6.57	+ 201.91	6.57
	Irish	+ 124.28	2.75	+ 124.28	2.75
	Scandinavian-Dutch	+ 646.37	6.34	+ 646.37	6.34
	German	+ 533.95	8.47	+ 533.95	8.47
	Italian	- 40.83	0.94	- 40.83	0.94
	Jewish	+ 759.55	18.28	+ 759.55	18.28
	Eastern European	- 171.47	3.00	- 171.47	3.00
	Other Ethnic Origin	- 464.30	12.03	- 464.30	12.03
Marital Status	Single	- 1,072.69	59.79	- 1,072.69	59.79
	Married	+ 354.83	59.79	+ 354.83	59.79
Type of Employment	Salaried	- 124.82	41.78	- 124.82	41.78
	Salaried/Self-Employed	+ 1,575.30	57.99	+ 1,575.30	57.99
	Self-Employed	- 447.07	10.55	- 447.07	10.55
	Constant Term	+ 7,122.57		+ 7,122.57	
	R^2	.4786	.4786	.4786	.4786
	F-Statistic	3,542.55		3,542.55	
	Computer Time Required[2] (Seconds)	11,450		6.60	

1. Based on observations for 227,678 workers.

2. Central processing unit time. In both cases, an IBM 370/168 computer was employed.

(b) An Example Involving the Use of Existing Aggregate Data

We noted in the introduction that one of the advantages of this algorithm is that it permits the treatment of existing and easily obtainable grouped data for which the underlying microdata cannot be directly accessed because of the understandable problems of confidentiality. Such grouped data are available, for instance, in a number of official publications of Statistics Canada. In the following material, which is intended as an example of the utility of the algorithm, we draw on data from one such publication.

Let us suppose that we wished to determine the functional relationship between the total income from all sources of all individuals in Canada receiving income, and their age, sex, period of immigration, and region of residence, using data from the 1971 Census. This relationship can be expressed as follows:

$$\text{Total income} = f(\text{sex, age, period of immigration, region}). \qquad (2.1)$$

The data necessary to determine this relationship are available in a Statistics Canada publication[1] for the 11,572,800 individuals who are identified as receiving income by the Census. The explanatory factors are divided into the 21 categories in Table 2.

---

1. Statistics Canada, *Income of Individuals by Sex, Age, Marital Status, and Period of Immigration 1971 Census of Canada*, Cat. No. 94-760, Vol. III, Part 6 (June, 1975), Table 5.

Table 2

CATEGORIZATION OF INDEPENDENT VARIABLES
FOR CANADIANS WITH INCOME, 1971

Independent Variables	Categories
Sex	Male (S1) Female (S2)
Age (Years)	15-19 (A1) 20-24 (A2) 25-34 (A3) 35-44 (A4) 45-54 (A5) 55-64 (A6) 65 or more (A7)
Period of Immigration	Born in Canada (I1) Immigrated before 1946 (I2) Immigrated between 1946 and 1955 (I3) Immigrated between 1956 and 1960 (I4) Immigrated between 1961 and 1965 (I5) Immigrated between 1966 and 1971 (I6)
Region of Residence	Atlantic Region (R1) Quebec (R2) Ontario (R3) Prairie Region (R4) British Columbia (R5) Yukon and Northwest Territories (R6)

From the information contained in the publication it is possible to calculate the frequency cross-tabulation, and to determine the mean of the dependent variable for each category (see Table 3). With this information, as has been described, it is possible to proceed to the calculation of the moment matrix.

Table 3

FREQUENCY CROSS-TABULATION FOR CHARACTERISTICS OF CANADIANS WITH INCOME, 1971

	S1	S2	I1	I2	I3	I4	I5	I6	A1	A2	A3	A4	A5	A6	A7	R1	R2	R3	R4	R5	R6
S1	6807515[1] 6538[2]		5372535	454755	406670	219935	123035	230580	507385	870770	1433850	1270330	1116640	831450	777085	613815	1819465	2493155	1135790	729545	15750
S2		4765285 2883	3733095	411810	237925	139810	84835	157815	403590	718645	817995	680305	681110	528190	935440	390135	1200405	1876520	771680	517830	6715
I1			9105630 4975						835570	1383810	1811985	1507885	1415840	1056325	1094210	956655	2684295	3098960	1477810	867215	20690
I2				866565 4266					0	0	12410	35910	101005	177010	540225	21775	85995	345560	242640	169770	820
I3					644595 6233				17960	70705	92710	158795	173465	83365	47600	10995	82090	378425	86635	85445	965
I4						359745 5818			26115	37135	75405	123665	59100	22000	16325	4410	59220	205365	40015	50080	660
I5							207870 5686		11500	21585	82520	53805	22140	9365	6960	3160	43690	119125	17650	23870	375
I6								388395 5051	19830	76190	176810	70570	26205	11575	7215	6950	64575	222230	42720	50995	930
A1									910975 1214							83485	208685	351345	165435	99940	2085
A2										1589415 3577						143345	448165	583440	252545	157670	4255
A3											2251845 5898					182530	644230	844075	342765	231130	7110
A4												1950635 6876				150620	519780	759355	312130	204040	4695
A5													1797750 6598			149895	455380	695265	297085	196925	3210
A6														1359640 5748		124565	340050	504655	232735	155785	1850
A7															1712525 3048	169515	403570	631535	304770	201870	1265
R1																1003950 3985					
R2																	3019870 4969				
R3																		4369675 5459			
R4																			1907470 4556		
R5																				1247375 5255	
R6																					24465 5514

1. Frequency count

2. Mean of the dependent variable, total income, for this particular category.

Source: Statistics Canada, *Income of Individuals by Sex, Age, Marital Status, and Period of Immigration*, Cat. No. 94-760, 1971, Table 5, and estimates by the authors.

In order to calculate the t-statistics for the coefficients, the F-statistic, and R^2 we require as well the variance, or $(\sum_{t=1}^{N} Y_t^2)$. This information could have been obtained from Statistics Canada; however, since this is an illustrative exercise, we have unemployed $1.53N\cdot\bar{Y}^2$ in place of $(\sum_{t=1}^{N} Y_t^2)$ in order to obtain first approximation estimates for the t-statistics, F-statistic, and R^2. The multiplier 1.53 is obtained from the ratio of $(\sum_{t=1}^{N} Y_t^2)$ to $N\cdot\bar{Y}^2$ observed in the previous example for Montreal, and is indicative ($\pm.2$) of the sort of ratio one might generally expect for the income or earnings of a representative group of individuals.[1] The nature of the relationship between $(\sum_{t=1}^{N} Y_t^2)$ and the t-statistics for the coefficients is such that, in this case, only where the t-statistic indicates that the coefficient is a "borderline" case from the view-point of significance, or that it is insignificant, is there any real danger in attaching importance to the value of the coefficient.

The results are presented in Table 4. The calculations took 2.64 seconds of central unit processing time. The constant term, as has been noted earlier, is the mean of the dependent variable -- i.e., the mean total income from all sources for all individuals receiving income in the year studied. The coefficients are the calculated amounts by which a category will, on average,

---

1. Rough approximation based on published data suggest $1.53 \pm .20$ to be a reasonable figure in this case. No t-statistics in this range change from being significant to insignificant (or vice versa). Values below 1.53 raise the R^2, the F-statistic and the t-statistics, and values above lower them. An example of a case where the value of the ratio would be pushed up to about 2.00 is if half the population with income had an income of $1 and the other half an income of $10,065 -- a rather unlikely case.

affect the income of an individual relative to the overall mean income. Thus, under the additivity hypothesis, the calculated average income for a male, aged 35-44 years, who immigrated between 1946 and 1955, and lives in the Atlantic region was $5,033 + $1,380 + $1,597 + $249 - $1,032 = $7,227, all other factors not taken into account here being considered equal.

Table 4

REGRESSION COEFFICIENTS FOR SELECTED DETERMINANTS OF TOTAL INCOME GOING TO THOSE INDIVIDUALS RECEIVING INCOME, OBTAINED USING THE PROPOSED ALGORITHM, CANADA, 1971[1]

Independent Variables	Categories	Coefficients	t-Statistics
Sex	Male	+ 1,379.76	2,044.34
	Female	- 1,971.07	2,044.34
Age (Years)	15-19	- 3,724.02	1,353.26
	20-24	- 1,315.69	651.02
	25-34	+ 707.90	426.17
	35-44	+ 1,597.31	889.11
	45-54	+ 1,413.80	755.28
	55-64	+ 621.67	282.24
	65 or more	- 1,525.87	726.80
Period of Immigration	Born in Canada	+ 7.32	16.83
	Immigrated before 1946	- 43.14	13.82
	Immigrated between 1946 and 1955	+ 248.51	74.50
	Immigrated between 1956 and 1960	- 1.19	0.27
	Immigrated between 1961 and 1965	- 50.84	8.55
	Immigrated between 1966 and 1971	- 459.55	105.55
Region	Atlantic Region	- 1,032.00	394.51
	Quebec	- 165.43	121.45
	Ontario	+ 471.74	452.95
	Prairie Region	- 445.06	246.15
	British Columbia	+ 255.62	110.72
	Yukon and Northwest Territories	+ 179.18	10.32
	Constant Term	5,033.00	
	R^2	.4499	
	F-Statistic	556,759	
	Computer Time Required (Seconds)[2]	2.64	

1. Based on observations for 11,572,800 individuals receiving income.

2. Central processing unit time. An IBM 370/168 computer was employed.

Since the objective of this exercise employing published data is to illustrate the usefulness of the proposed alogrithm, it should be noted that the results obtained are both reasonable and informative. Without going very deeply, there are several points that stand out. Income, for instance, increases with age, reaching its peak for the 35-44 year age group, and then falls off thereafter. With respect to the effect of the period of immigration on income, the results would seem to indicate that the immigrant must go through a period of adaptation before being able to benefit fully from his or her own characteristics. It is also instructive to note the amplitude of the difference that the sex of the individual causes, as well as the difference that living in one region as opposed to another incurs.

Conclusion

We wish to stress, that although the use of this algorithm in multiple regression analysis results in considerable cost reductions, reductions which increase with sample size and the number of total categories for the explanatory variables, this is not necessarily its main advantage. In our opinion, a major benefit to be derived from this algorithm is that it permits multiple regression analysis of grouped data that gives results identical to those that would be provided by direct access to and analysis of the micro-data (i.e., there is no loss of efficiency). Such grouped data are often available in published or unpublished form from data gathering agencies, and, of course, the use of such data creates no problems with respect to confidentiality.

In most cases, there should be no problem in obtaining the variance or $(\sum_{t=1}^{N} Y_t^2)$ from data-gathering agencies so as to be able to calculate the t-statistics for the coefficients, the F-statistic for the equations, and the coefficient of multiple determination, R^2. Nonetheless, we strongly suggest that the data-gathering agencies should publish one or the other along with the frequency cross-tabulations and the means for the various categories that are often published, in order to facilitate this sort of analysis.[1]

Finally, we should underline the fact that the algorithm proposed in this document already forms an integral part of a multiple regression program in use at the Economic Council of Canada.[2] The underlying program (excluding the algorithm), known under the name INFINITE, was originally developed at The Statistical Methodology and Procedures Section, Division of Research and Statistics, Board of Governors of the Federal Reserve System, in Washington and modified at the Council.

---

1. For those cases where the variance, or $(\sum_{t=1}^{N} Y_t^2)$, cannot be acquired readily with respect to already published data, it is often generally possible to provide a reasonable estimate for $(\sum_{t=1}^{N} Y_t^2)$, as noted above, in order to calculate first approximation values for the t-statistics for the coefficients, the F-statistic, and R^2. The relationship between the estimate and $N \cdot \bar{Y}^2$ will depend on the nature of the distribution of the observations on the dependent variable.

2. A "user's manual" for this multiple regression program is being prepared at the Council. This manual will discuss efficient means of carrying out multivariate analysis (a) when the independent variables are all in dichotomous form, (b) when they are all continuous in form, or (c) when some are in dichotomous and some in continuous form.

INFINITE was designed to handle the multiple regression analysis of large numbers of observations. Where programs, such as MASSAGER, store all the basic information on the units of observation before calculating the moment matrix, INFINITE proceeds through iteration, thereby tying up less space in the computer's memory banks. Thus, INFINITE was already less expensive than most conventional programs, and, as can be seen from Table 1, the use of the algorithm in conjunction with INFINITE further reduces costs by an extremely large amount.

BIBLIOGRAPHY

1. BOULET, J-A., "L'analyse des disparités de revenus: un cadre méthodologique de recherche". Economic Council of Canada, Discussion Paper No. 34, July 1975.

2. BOULET, J-A. and A. RAYNAULD, "L'analyse des disparités de revenus suivant l'origine ethnique et la langue sur le marché montréalais en 1961". Economic Council of Canada, Discussion Paper No. 83, March 1977.

3. BOULET, J-A. and J.C.R. ROWLEY, "Measurement of Discrimination in the Labour Market: A Comment", *Canadian Journal of Economics*, Vol. X, No. 1 (February 1977), pp. 149-154.

4. CRAMER, J.S., "Efficient Grouping, Regression and Correlation in Engel Curve Analysis". *Journal of Statistical Association*, Vol. 59 (1964), pp. 233-250.

5. HAITOVSKY, Y., "Unbiased Multiple Regression Coefficients Estimated from One-Way-Classification Tables When the Cross-Classification are Unknown." *Journal of American Statistical Association*, Vol. 61 (1966), pp. 720-728.

6. HAITOVSKY, H.S., "Regression Estimation From Grouped Observations." Edited by Allan Stuart. Griffen's Statistical Monographs and Course 33. New York: Harper, 1973.

7. HOUTHAKKER, H.S., "The Combined Use of Differently Grouped Observations in Least Squares Regressions." Cited in Haitovsky (1973).

8. JOHNSTON, J., *Econometric Methods*. New York: McGraw-Hill, 1972.

9. LACROIX, R., C. LEMELIN and P. ROBILLARD, *Champ de spécialisation et revenu*, Department of Economics, University of Montreal, Discussion Paper No. 7706 (July 1977).

10. ORCUTT, G.H., H.W. WATTS and J.B. EDWARDS, "Data Aggregation and Information Loss." *American Economic Review*, (September 1968).

11. PRAIS, S.J. and J. AITCHISON, "The Grouping of Observations in Regressions Analysis." *Review of International Statistical Institute*, Vol. 1 (1954), pp. 1-22.

12. SCHEFFÉ, H., *The Analysis of Variance*. New York: John Wiley and Sons, 1959.