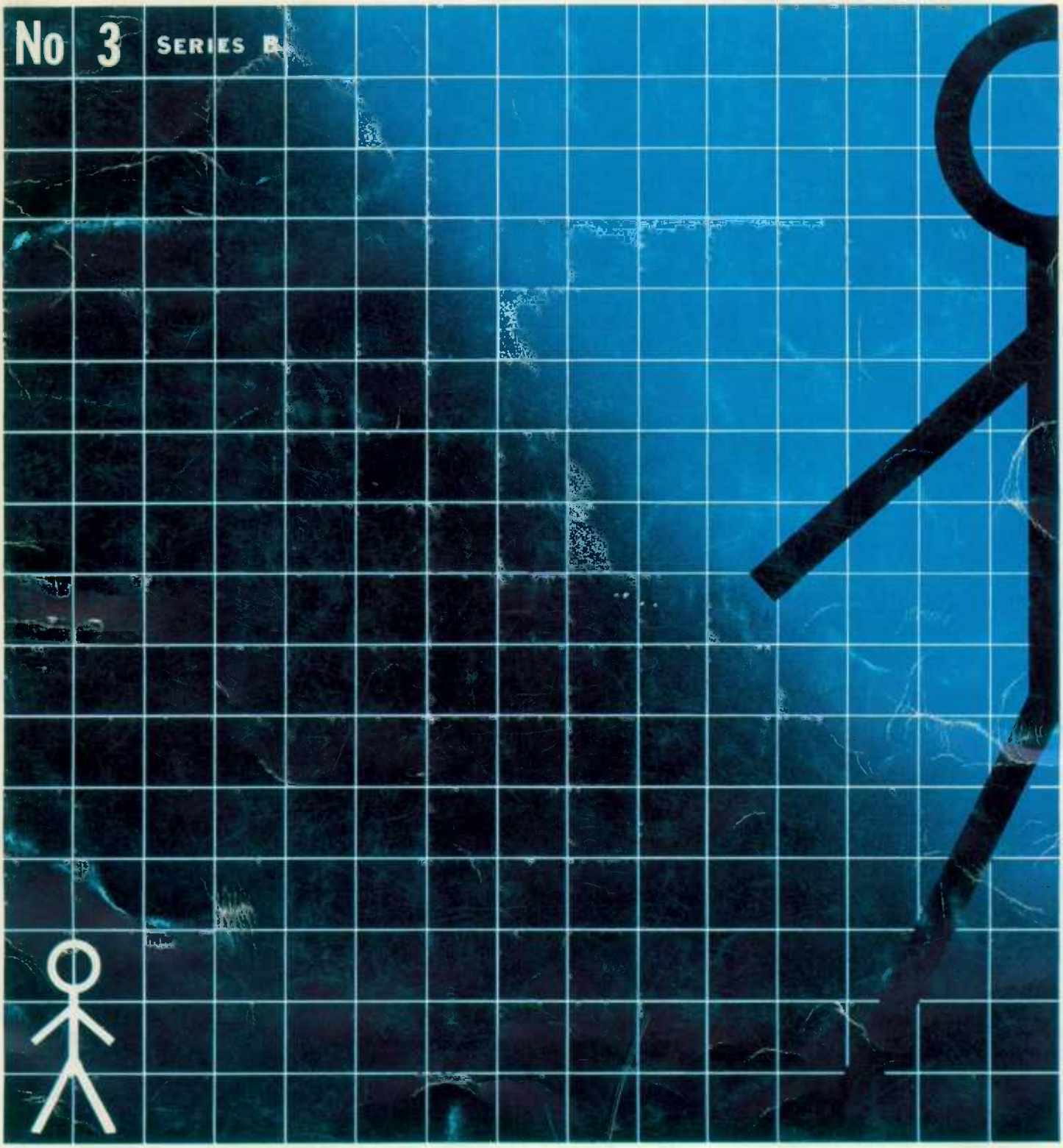
CATALOGUE No.
71-515
OCCASIONAL

CANADA

special labour force studies



some methods of analysing cross-classified census data—the case of labour force participation rates

DOMINION BUREAU OF STATISTICS

DOMINION BUREAU OF STATISTICS
Special Manpower Studies and Consultation Division

SPECIAL LABOUR FORCE STUDIES

Series B, No. 3

Some Methods of Analysing Cross-Classified Census Data
The Case of Labour Force Participation Rates

by

N. H. W. DAVIS

Published by Authority of
The Minister of Industry, Trade and Commerce

July 1969
1500-505

Price: 75 cents

SPECIAL LABOUR FORCE STUDIES

Catalogue number

71-505 **No. 1 Educational Attainment of the Canadian Population and Labour Force 1960-65** – Occasional – Price: 75 cents

Estimates based on supplementary questions appended to the **Labour Force Schedules** of February 1960 and 1965, including relationship between educational attainment and labour status and activity, and a comparison between native-born Canadians and post-war immigrants.

71-505F **N° 1 Niveau d'instruction de la population canadienne et de la main-d'oeuvre: 1960-1965** – Irrégulier – Prix: 75 cents

Estimations d'après des questions supplémentaires annexées aux **questionnaires de la main-d'oeuvre** de février 1960 et 1965, y compris le rapport entre le degré d'instruction et le niveau et l'activité de la main-d'oeuvre, et une comparaison entre les Canadiens de naissance et les immigrants d'après-guerre.

71-506 **No. 2 Annual Work Patterns of the Canadian Population/1964** – Occasional – Price: 75 cents

Annual work experience of the Canadian Population is compared with data from monthly surveys, and includes an analysis of long-duration unemployment, and part-year and part-time work.

71-507 **No. 3 The Job Content of the Canadian Economy/1941-61** – Occasional – Price: 75 cents

A review of the theory and measurement of Job Content together with an attempt to estimate the kinds of jobs, in the Canadian economy, by function and levels. A comparison is made with the United States.

71-508 **No. 4 Geographic Mobility in Canada/October 1964 — October 1965** – Occasional – Price: 75 cents

Migration of the Canadian population between municipalities is analysed by age, sex and region. For male migrants, aged 17-64, labour force status and reasons for leaving are also included.

71-509 **No. 5 Women Who Work: Part 1** – Occasional – Price: 75 cents

An evaluation of the relative importance of age, marital status, and education as factors influencing the participation of women in Canada's work force. The study is based on special tabulations of 1961 Census data.

71-510 **No. 6 Labour Force Characteristics of Post-war Immigrants and Native-born Canadians: 1956-67** – Occasional – Price: 75 cents

Differences between the labour force participation rates of post-war immigrants and native-born Canadians are compared taking into account differences in the age, sex, marital status, regional and educational attainment distributions of the two population groups.

71-511 **Series B – No. 1 The Demographic Background to Change in the Number and Composition of Female Wage Earners in Canada, 1951 to 1960** – Occasional – Price: 75 cents

An evaluation of demographic change over the 1951-1961 decade and its impact on the composition and number of female wage earners in 1961.

71-512 **No. 7 Educational Attainment in Canada: Some Regional and Social Aspects** – Occasional – Price: 75 cents

An examination of regional and occupational differences in educational attainment in Canada and its relation to migration is followed by a study of intergeneration changes in educational attainment.

71-514 **Series B – No. 2 Women Who Work: Part 2** – Occasional – Price: 50 cents

An evaluation of the influence of age, education of the wife, education of the husband, child status and residence on the participation of married women in Canada's work force, based on tabulations of the 1961 Census data.

71-515 **Series B – No. 3. Some Methods of Analysing Cross-Classified Census Data – The Case of Labour Force Participation Rates** – Occasional – Price: 75 cents.

A number of different statistical techniques are applied to cross-classified labour force participation rates obtained from the 1961 Census of Canada. The results are examined and compared.

Note: Ces bulletins seront bientôt disponibles en français.

Remittances should be in the form of cheque or money order, made payable to the Receiver General of Canada and forwarded to the Publications Distribution, Dominion Bureau of Statistics, Ottawa, or to the Queen's Printer, Hull, P.Q.

The Queen's Printer, Ottawa

FOREWORD

This third report in the Special Labour Force Studies (Series B) is devoted to consideration of some methodological aspects of manpower analysis—in this instance, the analysis of labour force participation of married women. A number of different statistical techniques are applied to the same body of cross-section data and the results examined and compared. It is intended as a reference work for and companion study to the substantive analyses in this area which have been presented both in the Special Labour Force Series and elsewhere.

These studies are prepared under the direction of Dr. Sylvia Ostry, Director, Special Manpower Studies and Consultation.

WALTER E. DUFFETT,
Dominion Statistician.

AUTHOR'S PREFACE

This study could not have been completed without the co-operation and assistance of many members of the Bureau and others. In particular, I would like to express my thanks to Mr. Leslie Szabo for overseeing the computer output of the model which is non-linear in the parameters, to Mr. Alan Sunter for his careful reading of an early draft and for many valuable discussions, and to Mrs. P. Morse (Department of Agriculture), Dr. Byron Spencer (McMaster University) and Mr. Dennis Featherstone for their comments on earlier drafts of the study. Self-imposed limitations on the scope of the study did not allow all their suggestions to be incorporated and any errors of omission, or commission, are of course entirely my own.

Lastly, I would like to express my thanks to Dr. Sylvia Ostry for her encouragement during the preparation of the study and my appreciation to Mrs. G. Walker who typed numerous drafts and whose help throughout was beyond the call of duty.

TABLE OF CONTENTS

List of Tables

List of Charts

I. INTRODUCTION

The purpose of this study is to consider ways in which labour force participation rates, obtained from the 1961 Census of Canada, cross-classified by a set of demographic, social or economic factors, can be related functionally to the different levels of those factors. It will be concerned with the examination and development of statistical models designed to represent these relationships and the testing of these models. The analysis of the results will, therefore, be in no way as detailed, in subject matter content, as would be required in a full empirical treatment of the variables.

Before proceeding it may be relevant to consider why it was felt that such a study is needed at all. Other studies have already successfully examined the relationship between demographic, social and economic variables and employment characteristics of the population using statistical techniques which have "explained" a considerable proportion of the variation in labour force participation rates between different sub-groups in the population.[1] To be justified, a study of this kind must therefore either (1) throw up evidence of serious weaknesses in the methods employed in these studies, or (2) it must show how significant improvements can be made in the predictive and analytical power of the model used. It was the latter condition only which was a consideration in the preparation of this study. It will be shown that simple additive models in which all the independent variables are expressed as dummies[2] and which yield results that are also simple to interpret, can be used to explain much of the observed differences in the labour force participation rates particularly when the independent factors (variables) are few in number.[3] It will also show how the analysis of variance, discussed in some detail on page 12, adds considerably to the analytical power of researchers when the data is in the form of a completely balanced factorial arrangement.[4] This will more often than not be the case when the data has been obtained from the census. But moving away from the simple type of model to ones which impose conditions on the form which the dependent variable can take—conditions which are inherent in the nature of the variable itself—results are obtained which are, in the author's view, not only conceptually more correct but also "explain" more of the overall variation in labour force participation rates. It will however, also be seen that the increase in the "power" of the more complicated models will not always be such as to justify their use in preference to the more simple models.

The Study is divided into six sections. Following this introduction, Section II—Quantal Response will briefly discuss how labour force participation rates, as a statistical variable, differ from other variables and will show why this is worth taking into account in the construction of the model. Section III, which will examine the simple additive model, will begin with an introduction to the methods of analysis of variance and dummy variable regression analysis which will also be of use in later sections in the study. The reason why so much more time and space is devoted to analysis of variance than to dummy variable regression is found in the author's view that analysis of variance in its widest sense (that is, not confined to the simple partitioning of the mean squares in regression analysis) is much neglected in books on econometrics and economic statistics, even though it often provides an ideal means of examining the data to obtain some notion of the "contribution" that independent variables make, both individually and in association with other variables, to the variation in the dependent variable. The technique is therefore of considerable value without the question of "testing hypotheses", which is its *raison d'être*, necessarily being considered. On the other hand, problems encountered in the use of dummy variables in regression analysis applied to economic data are now fairly well documented, and only a brief description of this approach is given.

The models developed in Sections IV and V will stem from the removal of some of the simplifying assumptions implicit in the simple additive model. A summary—Section VI—will then conclude the study with the exception of a bibliography of methodological studies.

[1] See Sylvia Ostry, *The Female Worker in Canada*, one of a series of Labour Force Studies in the Census Monograph Programme, Ottawa, Queen's Printer, 1968, and Dominion Bureau of Statistics, Special Labour Force Studies No. 2, Series B, *Women Who Work*: Part 2, by John D. Allingham and Byron Spencer, Ottawa, Queen's Printer, 1968.

[2] See Section III, page 16.

[3] See Sylvia Ostry, *op. cit.* and John D. Allingham and Byron Spencer, *op. cit.*

[4] The term completely balanced factorial design refers to that arrangement of the data in which an equal number of observations are available for all factor/level combinations.

II. QUANTAL RESPONSE

It is quite common, in the application of statistical methods to problems encountered in many of the applied sciences, to think in terms of a stimulus and a response.[5] In biological assay, for example, an insect is subjected to a dose of an insecticide (stimulus) and it either lives or dies (the response). Or a family with a given set of socio-economic characteristics may or may not have an automobile. In both of these examples the response is said to be **quantal**, that is, it is of the type "all or nothing"—the insect must live or die and the family must either own or not own an automobile. It should be clear then, from the above, that the problem which is to be considered in this study is one of how to deal with data obtained from a "quantal response" situation—an individual can only be either in, or not in, the labour force.[6]

[5] See D.J. Finney, *Probit Analysis, A Statistical Treatment of the Sigmoid Response Curve*, Cambridge University Press, 1952, and J. Aitchison and J.A.C. Brown, *The Lognormal Distribution*, Cambridge, 1957.

[6] For the purpose of this study the question of the intensity with which an individual can be in the labour force, i.e. whether the person is working, or wishes to work for 5 hours or 50 hours a week, will be ignored.

Now consider the situation where data relating to the labour force status of the population has been obtained from a census. In these cases the concept of the individual can still be retained but the data also reduced to a manageable size by considering all persons together who have common social and economic characteristics. All persons then, who have the same age, sex, marital status, etc., can be grouped and treated as a single observation. Obviously such a group of persons will no longer have an associated dependent variable of zero or one since it is to be expected that some of the members of that class will be in the labour force and some will not. If n_{1i} is the number in the ith class who are in the labour force and n_{0i} is the number who are not, then P_i, the labour force participation rate of the ith class is defined as

$$P_i = \frac{n_{1i}}{n}$$

where $n = n_{1i} + n_{0i}$ and $0 \leqslant P_i \leqslant 1$

A group of persons can now be thought of as being "subjected", say, to a given level of education and the response rate is then the percentage of the group who are in the labour force. Similarly the income level of the husband may be thought of as the stimulus, albeit a negative one, in the case of married women: it would be expected that the labour force participation rate of a group of married women whose husbands had incomes of, say, $5,000 a year would be different, and on *a priori* reasoning higher, from that of another group for whom the husband's income was $10,000 a year.

Even though controlled experiments cannot be carried out, censuses do provide the means by which data can be so ordered as to classify the population by those socio-economic variables which are known, or which are assumed to influence the likelihood of individuals entering the labour force.

What then is so special about this type of variable and how does it differ from other variables arising out of a quantitative rather than a quantal response situation? The answer to this question is found in the nature of the variable itself and by considering the way in which the variable changes in response to changes in the levels of the independent variables. By definition a proportion or percentage cannot lie outside the range of 0 to 1, or 0 to 100, so that any model which attempts to explain a dependent variable in that form should also be constrained to yield estimates which lie within the appropriate range. It would, furthermore, seem illogical to assume that the effect of a change in the level of an independent variable, on a variable which is constrained in this way, should be the same at all levels of the other factors being examined. Thus it would be reasonable to assume that the expected increase in the proportion of married women going out to work due to an increase in their level of educational attainment from say "some secondary" to "some university" would be different for two groups which had respectively 10 per cent and 50 per cent labour force participation rates at the lower educational level.[7] In other words some interaction between the effects of the factors would be expected.[8]

The data used throughout this study and on which alternative models have been tested consists of the observed participation rates in 54 cross-classifications of married women in urban Ontario, obtained from the 1961 Census, categorised by one factor (with three levels) which defines the 'child status' of the family; one factor (also of three levels) which defines the wife's level of educational attainment; and one factor, income of husband, which has six levels on a continuous scale. These factors and their levels are summarised below.

Factor	Levels
Child status	No children; some children under six years of age; no children under six years of age.
Education	Completed elementary or less; some high school or completed high school; some university or obtained a degree.
Income of husband (per annum)	Less than $1,000; $1,000 - $2,999, $3,000 - $4,999; $5,000 - $6,999; $7,000 - $9,999; $10,000 and over.

It should be noted that, when the data are organised in this way, the number of observations is defined by the product of the number of levels in each factor: thus 3 x 3 x 6 = 54. In order that the reader can compare subsequent results with the original observations, Table 1 contains, for each of the 54 factor/level combinations, the number of married women in the labour force, the number of married women in the population 14 years of age and over and the associated labour force participation rate.

The choice of this sub-set of data was dictated by three considerations. First, for reasons which will become clear later, no zero or 100 per cent participation rates were wanted. Secondly, the range of participation rates should not be so limited as to bring into question the applicability of the techniques to the analysis of participation rates in general. And thirdly, at least two types of independent variables – quantitative and qualitative – should be included. In all other respects the sub-set was selected from a number of possible sub-sets of data which could have satisfied these three conditions. It must be stressed, however, that these conditions were only imposed to facilitate the presentation of the methods employed: in no way do they imply that serious limitations exist in the use of the methods when these restrictions are removed.

---

[7] These different labour force participation rates, at the lower of the two educational levels, could exist because the other factors – age, husband's income etc., will not always be the same.

[8] The technique of analysis of variance, which can be used to detect the presence of significant interaction effects, is discussed in Section III, page 12.

TABLE 1. Population, Labour Force and Labour Force Participation Rates of Married Women in Urban Ontario by Child Status, Income of Husband and Level of Educational Attainment, June, 1961

Child status and income of husband	Education								
	Elementary			Secondary			University		
	Population	Labour force	Participation rate	Population	Labour force	Participation rate	Population	Labour force	Participation rate
	'000			'000			'000		
Children under 6:									
$10,000 and over	292	22	7.53	2,807	129	4.60	1,132	75	6.63
7,000-$9,999	823	50	6.08	6,032	349	5.79	1,177	92	7.82
5,000- 6,999	4,518	458	10.14	16,666	2,168	13.01	1,016	173	17.03
3,000- 4,999	11,785	2,051	17.40	21,488	4,583	21.33	622	174	27.97
1,000- 2,999	4,486	1,181	26.33	4,178	1,164	27.86	130	41	31.54
Under $1,000	723	167	23.10	671	200	29.81	56	20	35.71
No children under 6:									
$10,000 and over	539	71	13.17	4,104	418	10.19	1,234	130	10.53
7,000-$9,999	1,250	188	15.04	5,858	1,232	21.03	678	193	28.47
5,000- 6,999	5,381	1,303	24.21	12,284	4,265	34.72	584	304	52.05
3,000- 4,999	12,458	4,025	32.31	14,089	6,426	45.61	472	255	54.03
1,000- 2,999	4,462	1,701	38.12	2,659	1,395	52.46	129	82	63.57
Under $1,000	786	302	38.42	571	284	49.74	44	16	36.36
No children:									
$10,000 and over	386	69	17.88	1,838	269	14.64	420	85	20.24
7,000-$9,999	795	112	14.09	2,614	850	32.52	402	175	43.53
5,000- 6,999	3,060	745	24.35	6,465	3,125	48.34	576	358	62.15
3,000- 4,999	9,591	3,032	31.61	12,968	7,458	57.51	632	415	65.66
1,000- 2,999	4,791	1,689	35.25	3,879	2,230	57.49	239	148	61.92
Under $1,000	762	297	38.98	581	342	58.86	50	36	72.00

Source: 1961 Census.

III. SIMPLE ADDITIVE MODELS

It would be rare today for any research to break entirely new ground and earlier studies of a given subject will have undoubtedly provided some clues as to the expected relationship between interdependent variables. The analysis of labour force participation rates is no exception in this respect. However, it will often be the case that empirical research is being undertaken to explore the relationship between variables which, **in combination**, have not been examined before. When this is so it would seem to be desirable that the first step in the analysis should be to determine not only the relative importance of the effect which each independent variable has on the variation in the dependent variable, but also whether the effect of each variable changes significantly in response to changes in the other variables. The next step in the analysis is, then, to estimate the magnitude of any significant effects.

To conform to standard terminology, variables, in this section, will be called factors, and two types of effect or influence have to be considered – the main effect of a factor and the interaction effects. The main effect of, say, the child status factor is defined as the differences between the contributions of each of the three levels of this factor to the variation in the participation rates. The presence of a two-factor interaction arises, for example, when the three child status effects are not the same over, say, the three levels of the wife's education. With three factors there are, therefore, three two-factor interactions.

If child status is designated as factor A; the educational attainment of the wife as factor B and the husband's income as factor C then the three two-factor interactions are defined as AB, AC and BC and the one three-factor interaction is ABC.

Now consider a generalised three-factor model in which A has r levels, B, s levels and C, t levels which postulates that

$$P_{ijk} = \xi + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijk} \qquad (1)$$

where $i = 1 \ldots\ldots r$
$j = 1 \ldots\ldots s$
$k = 1 \ldots\ldots t$

and where α_i β_j and γ_k are the true contributions of the levels i, j, k of factors A, B and C, to the variation in P, and $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$ and $(\alpha\beta\gamma)_{ijk}$ are the true contributions of the three two-factor and one three-factor interactions. ξ is a constant term and e_{ijk} an error term which, for the purposes of various tests, is assumed to be normally distributed with mean zero and variance σ^2.

There is no unique solution to equation (1) because the true contribution of, say, the levels of factor A are not independent of the constant term ξ. And similarly for other factors. In least squares terminology there are more unknowns to be found than there are independent normal equations. But what can be estimated are the main effects and interaction effects which were defined above as the differences between the contributions of the factor levels.

This is done by placing certain constraints on the values which the terms in equation (1) can take. There are two ways of doing this which, although essentially the same, provide alternative approaches to the arithmetic and to the understanding of the results. The first, variance analysis, provides a simple way of completing the first stage of the analysis referred to on page 11 while at the same time yielding the data for the second stage. Dummy variable regression analysis provides a simple means of obtaining estimates of the effects, but only if all main and interaction effects are specified in the model does it provide all the answers yielded by variance analysis.

Variance Analysis

If $\bar{P}$ is the average of all the participation rates then with a factorial arrangement the expected value of $\bar{P}$ is μ, the true overall mean. And from equation (1) it follows that

$$\mu = \xi + \bar{\alpha} + \bar{\beta} + \bar{\gamma} + \overline{\alpha\beta} + \overline{\alpha\gamma} + \overline{\beta\gamma} + \overline{\alpha\beta\gamma} \qquad (2)$$

Substituting for ξ in equation (1) gives

$$P_{ijk} = \mu + (\alpha_i - \bar{\alpha}) + (\beta_j - \bar{\beta}) + (\gamma_k + \bar{\gamma}) + \ldots\ldots + ((\alpha\beta\gamma)_{ijk} - \overline{\alpha\beta\gamma}) + e_{ijk}$$

or

$$P_{ijk} = \mu + A_i + B_j + C_k + (AB)_{ij} + (AC)_{ik} + (BC)_{jk} + (ABC)_{ijk} + e_{ijk} \qquad (3)$$

where μ = the true mean participation rate

A_i = the true mean participation rate for which factor A is at the ith level minus the true mean

B_j = } defined in the same way as for A_i
C_k = }

$(AB)_{ij}$ = the true mean participation rate for which factor A is at the ith level and factor B is at the jth level minus $(\mu + A_i + B_j)$

$(AC)_{ik}$ = } defined in the same way as for $A_i B_j$
$(BC)_{jk}$ = }

$(ABC)_{ijk}$ = the true mean participation rate for which factor A is at the ith level, B is at the jth level and C is at the kth level, minus

$$(\mu + A_i + B_j + C_k + (AB)_{ij} + (AC)_{ik} + (BC)_{jk})$$

It follows from the definitions of the terms given above that

$$\Sigma A_i = \Sigma B_j = \Sigma C_k = \Sigma\Sigma(AB)_{ij} = \Sigma\Sigma(AC)_{ik} = \Sigma\Sigma(BC)_{jk} = \Sigma\Sigma\Sigma(ABC)_{ijk} = 0$$

It was stated above that μ is the expected value of $\bar{P}$; similar expressions[9] can also be formed for the remaining terms on the right hand side of equation (3).

It will also be stated without proof[10] that

$$E\left[\text{st}\,\Sigma(\bar{P}_i - \bar{P})^2\right] = \text{st}\,\Sigma A_i^2 + (r-1)\sigma^2 \qquad (4)$$

where σ^2 is the variance of the error term e_{ijk}. The expression on the right hand side of equation (4) when divided by the appropriate degrees of freedom, (r – 1), is the expected value of the mean square of the deviations of the r true means of factor A about the overall true mean, and its estimate is[11]

$$\text{st}\,\Sigma(\bar{P}_i - \bar{P})^2 \Big/ (r-1)$$

Similar expressions can be derived for the mean squares of the other main effects and interactions and these are summarised in Table 2 below.

Before discussing a little more of the methodology and interpretation of analysis of variance the calculations required in Table 2 have been applied to the data in Table 1 and these are given below in Table 3.

From Tables 2 and 3 it can be seen how the total sum of squares has been divided between the main effects and interactions. It can also be seen from the last column of Table 2 that the expectation of the mean squares is, in each case, the sum of two terms, one of which is the variance of the error term. If it was required to provide an independent estimate of the error variance then replicate observarions would be needed in each cell and differences between these estimates would provide the required variance estimate. But in the absence of replicate observations the significance of the effects can still be tested if some *a priori* assumption is made about the presence of the higher order interaction effects. For, in the example, if it is assumed that there is no ABC interaction then the estimate of

$$\frac{\Sigma\Sigma\Sigma(ABC)^2_{ijk}}{(r-1)(s-1)(t-1)} + \sigma^2$$

provides an independent estimate of σ^2, since $\Sigma\Sigma\Sigma(ABC)^2 = 0$, and can be used as such.[12]

---

[9] See O.L. Davies (ed.), *The Design and Analysis of Industrial Experiments*, London, 1954, page 281.

[10] *Ibid.*, page 282.

[11] The reason why the sum of squares $\Sigma(\bar{P}_i - \bar{P})^2$ is multiplied by st is because each $\bar{P}_i$ is the mean of st observations.

[12] It is also permissable to pool estimates of the mean squares of high order interactions to provide an estimate of the error variance providing that the one or more mean squares selected for this purpose were chosen before the data were analysed, i.e. there should be again an *a priori* assumption that the interactions are not significant.

TABLE 2. Expectation of Mean Squares in Analysis of Variance

Source of variation	Sum of squares	Degrees of freedom	Expectation of mean square
Main effects:			
A	$st\,\Sigma(\bar{P}_i - \bar{P})^2$	r-1	$\frac{st\,\Sigma A_i^2}{r-1} + \sigma^2$
B	$rt\,\Sigma(\bar{P}_j - \bar{P})^2$	s-1	$\frac{rt\,\Sigma B_i^2}{s-1} + \sigma^2$
C	$rs\,\Sigma(\bar{P}_k - \bar{P})^2$	t-1	$\frac{rs\,\Sigma C_k^2}{t-1} + \sigma^2$
Interactions:			
AB	$t\,\Sigma\Sigma(\bar{P}_{ij} - \bar{P}_i - \bar{P}_j + \bar{P})^2$	(r-1)(s-1)	$\frac{t\,\Sigma\Sigma(AB)_{ij}^2}{(r-1)(s-1)} + \sigma^2$
AC	$s\,\Sigma\Sigma(\bar{P}_{ik} - \bar{P}_i - \bar{P}_k + \bar{P})^2$	(r-1)(t-1)	$\frac{s\,\Sigma\Sigma(AC)_{ik}^2}{(r-1)(t-1)} + \sigma^2$
BC	$r\,\Sigma\Sigma(\bar{P}_{jk} - \bar{P}_j - \bar{P}_k + \bar{P})^2$	(s-1)(t-1)	$\frac{r\,\Sigma\Sigma(BC)_{jk}^2}{(s-1)(t-1)} + \sigma^2$
ABC	$\Sigma\Sigma\Sigma(P_{ijk} + \bar{P}_i + \bar{P}_j + \bar{P}_k - \bar{P}_{ij} - \bar{P}_{ik} - \bar{P}_{jk} - \bar{P})^2$	(r-1)(s-1)(t-1)	$\frac{\Sigma\Sigma\Sigma(ABC)_{ijk}^2}{(r-1)(s-1)(t-1)} + \sigma^2$
Total	$\Sigma\Sigma\Sigma(P_{ijk} - \bar{P})^2$	rst-1	

TABLE 3. Variance Analysis of Labour Force Participation Rates

	Sum of squares	Degrees of freedom	Mean squares	Variance ratio[1]
Main effects:				
Child status	5,560.0	2	2,780.0	121.5
Education	2,260.9	2	1,130.5	49.4
Income of husband	7,860.5	5	1,572.1	68.7
Second-order interactions:				
Child status/education	744.7	4	186.2	8.1
Child status/income	642.5	10	64.2	2.8
Education/income	662.3	10	66.2	2.9
Third-order interaction:				
Child status/education/income	457.8	20	22.9	
Total	**18,188.8**	**53**		

[1] Obtained by dividing the mean squares by the mean square of the third-order interaction.

If the main effects and interactions did not exist, i.e. variations in child status, education and husbands income had no effect on the wife's participation rate, then all the mean squares would also be independent estimates of σ^2.[13] It therefore follows that if any factor or interaction does affect the level of the participation rate then the observed mean square will be larger than the estimate of σ^2. The test of significance employed is the standard F test for the ratio of two variances.

In the last column of Table 3 the ratios of mean squares of the main effects and second-order interaction effects, to the mean square of the third-order interaction (here assumed to be an estimate of σ^2) are given. Tables of the F distribution for the appropriate degrees of freedom can then be referred to to see if the ratio is significant. Table 4 compares the observed ratios with those expected from the F distribution with the degrees of freedom shown in the table.

These ratios show that not only are the main effects highly significant, as was to be expected in this case, but the interaction of child status with education is also significant at the one per cent level. In other words the hypothesis that the effect of child status on the wife's labour force participarate is the same at different levels of her education has to be rejected. The interaction effects of child status with income, and of education with income, also appears from the table to be significant at the five per cent level. For completeness, however, and to further illustrate the use of analysis of variance in analysing cross classified data, Table 5 below contains a complete list of the 54 observations showing how they can be built up from the main and interaction effects defined by equation (3).

To illustrate the effect of the child status/education interaction the nine estimates of $\mu + A_i + B_j + (AB)_{ij}$ obtained from Table 5 are plotted on Chart 1 together with the average education, $(\mu + B_j)$, effect. This suggests a reason for the significance of this interaction. When there are children under six in the family the effect of the wife's education on the likelihood of her being in the labour force is small relative to the importance of this factor when there are no children.

It might appear from the above description of the technique of analysis of variance that it provides many of the answers to the problems that are encountered in the analysis of cross sectional data providing that the data can be arranged in the form of a "factorial design". If this is so it is because

[13] *Ibid.*, page 256.

TABLE 4. Variance Ratios and F Distribution

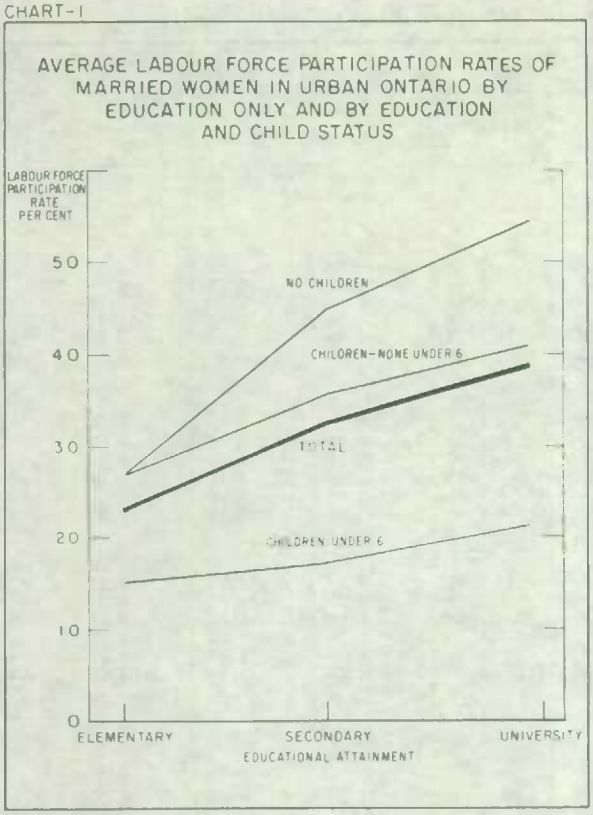
Effect	Variance ratio[1]	Degrees of freedom[2]	F distribution level of significance		
			10%	5%	1%
Child status	121.5	2 and 20	2.59	3.49	5.85
Wife's education	49.4	2 " 20	2.59	3.49	5.85
Husband's income	68.7	5 " 20	2.16	2.71	4.10
Child status/education	8.1	4 " 20	2.25	2.87	4.43
Child status/income	2.8	10 " 20	1.94	2.35	3.37
Education/income	2.9	10 " 20	1.94	2.35	3.37

[1] Mean square of effect divided by the mean square of the third-order interaction.
[2] The appropriate F distribution is defined by the degrees of freedom associated with the two variances.

TABLE 5. Contribution of Main Effects and Interaction Effects to Original Observations

Factor level combinations A. Child status B. Education C. Husband's income	Constant term = overall mean	Estimated main effects			Estimated interaction effects				Original observations
		A_i	B_j	C_k	$(AB)_{ij}$	$(AC)_{ik}$	$(BC)_{jk}$	$(ABC)_{ijk}$	
Children under 6:									
Elementary:									
$10,000 and over			- 8.42	- 19.71	+ 5.76	+ 8.20	+ 9.57	- 5.63	7.53
7,000 - $9,999			- 8.42	- 12.05	+ 5.76	+ 0.85	+ 0.79	+ 1.39	6.08
5,000 - 6,999			- 8.42	+ 0.36	+ 5.76	- 4.73	- 3.79	+ 3.20	10.14
3,000 - 4,999			- 8.42	+ 7.85	+ 5.76	- 3.38	- 3.74	+ 1.57	17.40
1,000 - 2,999			- 8.42	+ 12.42	+ 5.76	- 1.60	- 2.19	+ 2.60	26.33
Under 1,000			- 8.42	+ 11.13	+ 5.76	+ 0.65	- 0.63	- 3.15	23.10
Secondary:									
$10,000 and over			+ 1.11	- 19.71	- 1.80	+ 8.20	- 3.01	+ 2.05	4.60
7,000 - $9,999			+ 1.11	- 12.05	- 1.80	+ 0.85	- 0.70	+ 0.62	5.79
5,000 - 6,999	31.42	- 13.66	+ 1.11	+ 0.36	- 1.80	- 4.73	- 0.87	+ 1.18	13.01
3,000 - 4,999			+ 1.11	+ 7.85	- 1.80	- 3.38	+ 1.10	- 1.31	21.33
1,000 - 2,999			+ 1.11	+ 12.42	- 1.80	- 1.60	+ 0.99	- 1.02	27.86
Under 1,000			+ 1.11	+ 11.13	- 1.80	+ 0.65	+ 2.48	- 1.52	29.81
University:									
$10,000 and over			+ 7.31	- 19.71	- 3.95	+ 8.20	- 6.55	+ 3.57	6.63
7,000 - $9,999			+ 7.31	- 12.05	- 3.95	+ 0.85	- 0.07	- 2.03	7.82
5,000 - 6,999			+ 7.31	+ 0.36	- 3.95	- 4.73	+ 4.65	- 4.37	17.03
3,000 - 4,999			+ 7.31	+ 7.85	- 3.95	- 3.38	+ 2.64	- 0.26	27.97
1,000 - 2,999			+ 7.31	+ 12.42	- 3.95	- 1.60	+ 1.19	- 1.59	31.54
Under 1,000			+ 7.31	+ 11.13	- 3.95	+ 0.65	- 1.84	+ 4.65	35.71
No children under 6:									
Elementary:									
$10,000 and over			- 8.42	- 19.71	+ 0.85	- 3.44	+ 9.57	- 0.13	13.17
7,000 - $9,999			- 8.42	- 12.05	+ 0.85	- 0.89	+ 0.79	+ 0.31	15.04
5,000 - 6,999			- 8.42	+ 0.36	+ 0.85	+ 2.19	- 3.79	- 1.43	24.21
3,000 - 4,999			- 8.42	+ 7.85	+ 0.85	+ 1.68	- 3.74	- 0.36	32.31
1,000 - 2,999			- 8.42	+ 12.42	+ 0.85	+ 4.51	- 2.19	- 3.50	38.12
Under 1,000			- 8.42	+ 11.13	+ 0.85	- 4.07	- 0.63	+ 5.11	38.42
Secondary:									
$10,000 and over			+ 1.11	- 19.71	+ 0.06	- 3.44	- 3.01	+ 0.73	10.19
7,000 - $9,999			+ 1.11	- 12.05	+ 0.06	- 0.89	- 0.70	- 0.95	21.03
5,000 - 6,999	31.42	+ 3.03	+ 1.11	+ 0.36	+ 0.06	+ 2.19	- 0.87	- 2.58	34.72
3,000 - 4,999			+ 1.11	+ 7.85	+ 0.06	+ 1.68	+ 1.10	- 0.64	45.61
1,000 - 2,999			+ 1.11	+ 12.42	+ 0.06	+ 4.51	+ 0.99	- 1.08	52.46
Under 1,000			+ 1.11	+ 11.13	+ 0.06	- 4.07	+ 2.48	+ 4.58	49.74
University:									
$10,000 and over			+ 7.31	- 19.71	- 0.93	- 3.44	- 6.55	- 0.60	10.53
7,000 - $9,999			+ 7.31	- 12.05	- 0.93	- 0.89	- 0.07	+ 0.65	28.47
5,000 - 6,999			+ 7.31	+ 0.36	- 0.93	+ 2.19	+ 4.65	+ 4.02	52.05
3,000 - 4,999			+ 7.31	+ 7.85	- 0.93	+ 1.68	+ 2.64	+ 1.03	54.03
1,000 - 2,999			+ 7.31	+ 12.42	- 0.93	+ 4.51	+ 1.19	+ 4.62	63.57
Under 1,000			+ 7.31	+ 11.13	- 0.93	- 4.07	- 1.84	- 9.69	36.36
No children:									
Elementary:									
$10,000 and over			- 8.42	- 19.71	- 6.61	- 4.76	+ 9.57	+ 5.75	17.88
7,000 - $9,999			- 8.42	- 12.05	- 6.61	+ 0.04	+ 0.79	- 1.72	14.09
5,000 - 6,999			- 8.42	+ 0.36	- 6.61	+ 2.53	- 3.79	- 1.78	24.35
3,000 - 4,999			- 8.42	+ 7.85	- 6.61	+ 1.68	- 3.74	- 1.21	31.61
1,000 - 2,999			- 8.42	+ 12.42	- 6.61	- 2.93	- 2.19	+ 0.92	35.25
Under 1,000			- 8.42	+ 11.13	- 6.61	+ 3.42	- 0.63	- 1.97	38.98
Secondary:									
$10,000 and over			+ 1.11	- 19.71	+ 1.72	- 4.76	- 3.01	- 2.77	14.64
7,000 - $9,999			+ 1.11	- 12.05	+ 1.72	+ 0.04	- 0.70	+ 0.34	32.52
5,000 - 6,999	31.42	+ 10.64	+ 1.11	+ 0.36	+ 1.72	+ 2.53	- 0.87	+ 1.43	48.34
3,000 - 4,999			+ 1.11	+ 7.85	+ 1.72	+ 1.68	+ 1.10	+ 1.99	57.51
1,000 - 2,999			+ 1.11	+ 12.42	+ 1.72	- 2.93	+ 0.99	+ 2.12	57.49
Under 1,000			+ 1.11	+ 11.13	+ 1.72	+ 3.42	+ 2.48	- 3.06	58.86
University:									
$10,000 and over			+ 7.31	- 19.71	+ 4.88	- 4.76	- 6.55	- 2.99	20.24
7,000 - $9,999			+ 7.31	- 12.05	+ 4.88	+ 0.04	- 0.07	+ 1.36	43.53
5,000 - 6,999			+ 7.31	+ 0.36	+ 4.88	+ 2.53	+ 4.65	+ 0.36	62.15
3,000 - 4,999			+ 7.31	+ 7.85	+ 4.88	+ 1.68	+ 2.64	- 0.76	65.66
1,000 - 2,999			+ 7.31	+ 12.42	+ 4.88	- 2.93	+ 1.19	- 3.01	61.92
Under 1,000			+ 7.31	+ 11.13	+ 4.88	+ 3.42	- 1.84	+ 5.04	72.00

CHART-1



the model used is really a very simple one and in the example only three factors were examined. When more factors are included the analysis necessarily becomes more involved and the interpretation more difficult. With three factors there are three two-factor interactions, with four factors there are six and with five factors there are ten. There are also ten three-factor interactions with five factors compared, with only the one in the example used in this study. But even if it is not proposed to complete the analysis in the form described above but rather to use the approach described in the next section, it is still worthwhile to perform an initial analysis of variance when there are a large number of factors. This permits an investigation to be made of the way in which the total variation (sums of squares) has been made up from the main and interaction effects. This is useful because if a factor persistently appears in interactions which have large mean squares it might be desirable to control for this factor in the subsequent analysis.[14]

Dummy Variable Regression

There may be situations where it is unnecessary to use the full analysis of variance approach. For example, it might be assumed, from earlier studies, that interaction terms are not likely to be significant. Or it may be that the factor or factors which are associated with large interaction effects have in some way been controlled for. In these cases an alternative approach to the analysis can be used.[15]

Suppose that the interaction terms in equation (2) are discarded from the model. The model is then

$$P_{ijk} = \xi + \alpha_i + \beta_j + \gamma_k + e_{ijk} \qquad (5)$$

where ξ, α_i, β_j, γ_k and e_{ijk} are as before.

Now if α_i is replaced by

$$\alpha_1 x_1 + \ldots\ldots + \alpha_i x_i + \ldots\ldots + \alpha_r x_r$$

where x_i has the value of 1 when factor A is at the ith level and 0 otherwise, and β_j and γ_k are similarly replaced by

$$\beta_1 y_1 + \ldots\ldots + \beta_j y_j + \ldots\ldots + \beta_s y_s$$

$$\text{and } \gamma_1 z_1 + \ldots\ldots \quad \gamma_k z_k + \ldots\ldots \quad \gamma_t z_t$$

then since r = 3, s = 3 and t = 6, equation (5) can be rewritten

$$P = \xi + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 z_4 + \gamma_5 z_5 + \gamma_6 z_6 + e \qquad (6)$$

where ξ = a constant term

$x_1 = \begin{cases} 1 \text{ when child status is at the first level} \\ 0 \text{ otherwise} \end{cases}$

$x_2 = \begin{cases} 1 \text{ when child status is at the second level} \\ 0 \text{ otherwise} \end{cases}$

$x_3 = \begin{cases} 1 \text{ when child status is at the third level} \\ 0 \text{ otherwise} \end{cases}$

$y_1 = \begin{cases} 1 \text{ when education is at the first level} \\ 0 \text{ otherwise} \end{cases}$

.
.
.
.
.

[14] See Sylvia Ostry, *op. cit.*

[15] See E. Malinvaud, *Statistical Methods of Econometrics*, Vol. 5 in the series: Studies in Mathematical and Managerial Economics, Chicago, 1966, A.S. Goldberger, *Econometric Theory*, New York, 1964, E. Melichar, *Least-Squares Analysis of Economic Survey Data*, Board of Governors of the Federal Reserve System, (mimeo), J. Johnston, *Econometric Methods*, New York, 1963.

$$z_1 = \begin{cases} 1 \text{ when income is at the first level} \\ 0 \text{ otherwise} \end{cases}$$

.
.
.
.
.

$$z_6 = \begin{cases} 1 \text{ when income is at the sixth level} \\ 0 \text{ otherwise} \end{cases}$$

Now, for reasons mentioned on page 12, there is no unique solution to this equation. But as also mentioned earlier, what can be estimated are the differences between contributions of the different levels of a factor. The constant term can then be arbitrarily fixed to facilitate the interpretation of the results. A number of approaches to the arithmetic have been designed to produce the desired results using standard regression analysis programmes. One of these is described below.

The number of parameters to be estimated can be reduced to the number of independent normal equations by constraining one coefficient in each factor to be equal to zero, say

$$\alpha_2 = \beta_2 = \gamma_2 = 0$$

so that equation (6) becomes

$$P = \xi' + \alpha'_1 x_1 + \alpha'_3 x_3 + \beta'_1 y_1 + \beta'_3 y_3 + \gamma'_1 z_1 + \gamma'_3 z_3 + \gamma'_4 z_4 + \gamma'_5 z_5 + \gamma'_6 z_6 + e \qquad (7)$$

Equation (7) is now a reparameterised version of the model given by equation (6). Each coefficient now measures the expected difference between the participation rates of married women associated with that level of the factor and of those associated with the omitted level. The estimated values, a'_i, b'_j and c'_k (i, j and k ≠ 2), of the coefficients and of the constant term can then be obtained in the normal way.

However, in order to simplify the presentation it will often be found useful to modify the results so that the constant term takes on the value of the overall mean of the observations. This can easily be done as follows:

For the first factor obtain a value

$$k_a = \frac{a'_1 + a'_3}{3}$$

where the denominator is the number of levels of that factor; then obtain k_b and k_c for the other factors in the same way. It can now be stated that[16]

$$\xi' + k_a + k_b + k_c = \bar{P}$$

In this way the overall mean can be selected as one estimate of ξ in equation (6) with the associated set of estimates, a_i, b_j and c_k, of the coefficients in equation (6) found using the following rule

$$a_1 = a'_1 - k_a$$

$$a_2 = -k_a$$

$$a_3 = a'_3 - k_a$$

and similarly for the other factors.

Since $\Sigma a_i = \Sigma b_j = \Sigma c_k = 0$ the interpretation on the results is that the constant term is the mean participation rate and the coefficients the expected departures from that mean associated with the different levels of each factor.

It only remains to consider briefly the relationship between the proportion of the total variation "explained" by the dummy variable regression models and the components of the variance analysis model given in Table 2 on page 13.

First, consider again the model postulated as equation (6) in which all factor/levels are represented by dummy variables. The percentage of the total variation explained by each of the factors separately is given by the formula for the sum of squares attributable to the main effects of each factor shown in Table 2. It follows that the total explained variance obtained from the regression model described by equation (7) is equal to the sum of the first three terms in column 2 of Table 2. The residual sum of squares obtained from equation (7) will, as has been stated earlier, be the sum of squares due to the interaction effects or last 4 terms in column 2 of Table 2.

Now consider the situation where one of the independent factors in the equation is represented by a continuous variable. In this case it is still necessary for one of the coefficients among each of the remaining sets of the dummy variables to be constrained to zero and, if desired, the results obtained can still be adjusted in the way described above to provide coefficients for all the factor/levels. Because the constant term in such a model will not become the overall mean after following such a procedure, the interpretation of the results will not be so simplified.[17] However, in this study, where both dummy and continuous variables have been used, the procedure outlined above has been followed so as to maintain a consistent form of presentation. The sum of squares attributable to the factor represented by a continuous variable will now be lower than if dummy variables had been used. The difference between these two sums of squares can then be thought of as a measure of how far the

[16] See references cited in Footnote (15) and in particular E. Melichar, *Least-Squares Analysis of Economic Survey Data*, *op. cit.*

[17] It is interesting to note that the coefficients obtained for a set of dummy variables, after they have been adjusted in the way proposed earlier, in an equation containing at least one continuous variable, are still the deviations from the overall mean even though the overall mean does not appear in the equation.

assumptions, made about the form which the continuous variable takes, reflect the free form obtained from the use of dummy variables.[18]

Before leaving this section it is necessary to mention a point concerning the standard errors of the estimates obtained from the regression analysis of dummy variables. The coefficients obtained from fitting the dummy variable regression model of equation (7) are the differences between the contributions of the factor at the designated levels and that at the omitted level. The standard error of the coefficients are the standard errors of these differences. These standard errors are therefore unchanged when the coefficients are deconstrained to the form of equation (6) since the differences between the coefficients do not change. Furthermore, if the data are arranged in the form of a completely balanced factorial design, as is the case in this study, then the standard errors obtained will be identical for all levels of the same factor and will be the standard error of the difference between the coefficients of any two levels of that factor.[19]

Table 6 displays the results of fitting the dummy variable regression model of equation (6) to the data used in this study. It will be noted that over 86 per cent of the total variation in labour force

[18] In certain cases, using the principles of orthogonal polynomials (see Davies, *op. cit.*), it is possible to obtain the partitioning of the sum of squares for a factor directly from the variance analysis.

[19] Since the differences between the contributions of the different factor levels are the only effects which can be estimated the statements made in this paragraph are not surprising. They have been included because it is a point which appears to have been neglected in the literature.

TABLE 6. Regression Equations of Labour Force Participation Rates of Married Women in Urban Ontario

		Coefficients of					
	Constant	Child status		Education		Husband's income	
Model I							
						$'000	
$R^2 = 0.8622$	31.421	Children under 6	– 13.661	Elementary	– 8.421	10+	– 19.709
$N = 54$		No children under 6	+ 3.025	Secondary	+ 1.107	7-10	– 12.046
		No children	+ 10.636	University	+ 7.313	5- 7	+ 0.357
						3- 5	+ 7.849
						1- 3	+ 12.417
						0- 1	+ 11.133
(Standard error of coefficients)			(2.516)		(2.516)		(3.559)
Model II							
$R^2 = 0.8262$	45.377	Children under 6	– 13.661	Elementary	– 8.421	Continuous variable	
$N = 54$		No children under 6	+ 3.025	Secondary	+ 1.107	$'000 natural scale	– 2.263
		No children	+ 10.636	University	+ 7.313		
(Standard error of coefficients)			(2.705)		(2.705)		(0.216)
Model III							
$R^2 = 0.7372$	43.731	Children under 6	– 13,661	Elementary	– 8.421	Continuous variable	
$N = 54$		No children under 6	+ 3.025	Secondary	+ 1.107	$'000 on Log_2 scale	– 6.328
		No children	+ 10.636	University	+ 7.313		
(Standard error of coefficients)			(3.326)		(3.326)		(0.845)

participation rates is explained by this model which would suggest that for practical purposes it would meet the needs of most research workers. It was noted above that each point estimate is made up by the addition of four terms: a constant term which is the average of all the observations and one term for each factor depending on the level of that factor. Thus the estimated labour force participation rate of married women in urban Ontario who:

(1) have children less than 6 years of age;

(2) have only elementary school education;

(3) have husbands whose income was between $5,000 and $7,000 a year;

is (see Table 6, Model 1)

31.42 - 13.66 - 8.42 + 0.36 = 9.70 per cent

which compares very favourably with an actual figure of 10.14 per cent.

Before proceeding, however, it may be worthwhile to compare the results obtained from this dummy variable regression model and those from the variance analysis model. First compare the coefficients given in Table 6 Model I with the contributions of the main effects shown in Table 5. That they agree is as they should – by definition: but this comparison makes it possible to see what effects have not been picked up by assuming the relationship takes the form given by equation (6). In particular the significant interaction effect of child status with educational attainment noted on page 14 and in Chart 1 is ignored.

But another repercussion of the failure to include significant interaction terms, or to control for them, which was referred to earlier, is the effect that this has on the significance which can be attached to the coefficients. For example, on page 12 it was shown that if it were assumed that the third-order interaction effects were not significant then the value of the mean square obtained for this interaction was an independent estimate of the variance of the error term. And from the formula for the standard error of the difference between two means it can be shown that on this basis a difference of 4.7 percentage points is required between the participation rates at the different income levels before they could be said to be significant at the 5 per cent level. However, using the standard error obtained from pooling the mean squares of all the interaction terms participation rates at any two income levels must differ by more than 7.2 percentage points before the null hypothesie is rejected. It so happens that the increase in the difference required for significance, under the two sets of assumptions, from 4.7 to 7.2 would not, given the data used in this study, change the general conclusions concerning the effect of the income of husbands on the wife's labour force participation rate. However, it is clear that this may not always be the case.

Another illustration of the disadvantage of using a dummy variable model without interaction terms can be provided. If in the example worked out above, while holding the other factor/levels constant, the husband's income is changed from $7,000 - $9,999 to $10,000 and over, then the new estimated participation rate becomes:

31.42 - 13.66 - 8.42 - 19.71 = -10.37 per cent

which is not only clearly way off the observed percentage of 7.53 but is obviously an absurd result.

It is evident therefore, that although wives with husbands whose incomes are $10,000 a year and over have, on average, a propensity to be in the labour force some 20 percentage points lower than that of wives with husbands earning between $5,000 and $6,999 a year, this relationship does not hold over all the combinations of the other factor/levels. But before considering in some detail how this particular problem can be partially taken care of while still retaining the basic form of a simple additive model, two other forms of this model were tested.

Child status and, in the absence of a numeric scale, education have to be included in the model as dummy variables, but this does not apply to the income variable. Certainly the use of dummy variables to represent the income categories, or levels, does allow the resulting coefficients to take a completely free form and thus provide better estimates of the observed factor/level combinations. However, it may be that research workers are interested in estimating the participation rates associated with other income levels within or outside the range of observed values. This can be done, by interpolation or, with doubtful justification, extrapolation from the coefficients obtained in Model I. Alternatively, income can be treated as a continuous variable. The latter procedure is the one used in this Study.

The second model included the husband's income as a variable on a linear, or natural scale and in the third model the logarithm of the husband's income was used. The reason for using the logarithm of the husband's income will be explained later. Because the classification by income was only available for those class intervals given earlier, the mid-points of these intervals were taken to represent the points on a continuous income scale. However, as this was not possible for incomes of $10,000 and over, other evidence was obtained which showed that the average income for this class was likely to be close to $16,000 and this figure was used. Also, instead of taking logarithms to the base 10, those to the base 2 were used as a matter of convenience (when the incomes were in $'000's), so that the resulting coefficients would be the expected change in the participation rates arising out of a doubling in the income of the husband. The form of these two models, Models II and III, were therefore

$$P = \xi_1 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \gamma z + e \qquad (8)$$

and $$P = \xi_2 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \gamma \log z + e \quad (9)$$

where the x's and y's are as defined above in equation (6), and z is the income of the husband.

The parameters, or coefficients, of Models II and III were obtained using the method of least squares after one of each of the "child status" and "educational attainment" levels had been constrained to be equal to zero. And for reasons mentioned earlier the results given in Table 6 for these Models are again those calculated after "deconstraining" the coefficients by the method described above. But, as indicated, the constant term will not now become the overall mean of the participation rates.

Compared with Model I, which explained 86.2 per cent of the total variation in labour force participation rates, Models II and III are rather less efficient, explaining 82.8 and 73.7 per cent respectively. The contributions of child status and education to the expected value of the participation rate and hence to the variation in participation rate are unchanged. It is the removal of the **free form** from the income variables which has caused the loss of efficiency in Models II and III. Perhaps the simplest way to show the full effect of the changes in the form of the models is to reconstruct the variance analysis shown in Table 3, but adding the newfound components due to the specified linear form of the income variable in Model II and the log form in Model III. The results are summarised in Table 7 below. The point to note is that, although in both Model II and III an assumed form has been placed on the income variable, the total sum of squares attributable to this factor is unchanged. All that has happened is that the sum of squares has now been partitioned, in the way shown in the table. As was stated at the end of the last section, this provides some measure of how far the chosen, or assumed, form of the continuous variable approximates to the **free form** provided by Model I. But whereas in the variance analysis approach it is possible to retain the separate components to see whether the non-linear component is, itself, significant, in the regression approach used above the non-linear and non log-linear components were respectively included with the two and three-factor interaction sum of squares to make up the composite "error" sum of squares.

Only a few lines need now be spent on the interpretation of the coefficients given in Table 6 Models II and III. In Model II the coefficient of the income factor suggests that the labour force participation rate of the wife declines by 2.26 percentage points for every $1,000 increase in the husband's income, while from Model III the conclusion would be that a doubling of the husband's income would cause a drop of 6.33 percentage points in the wife's participation rate. The estimated labour force participation rates for the same factor/level combinations as was calculated for Model I on page 19 are:

MODEL II

$$45.38 - 13.66 - 8.42 - (2.26 \times 6) = 9.74$$

MODEL III

$$43.73 - 13.66 - 8.42 - (6.33 \times 2.585) = 5.29$$

where 2.585 is the $\log_2 6$.

TABLE 7. Analysis of Variance including Non-linear Components of the Main Effects

	Sum of squares	Degrees of freedom	Mean squares
Main effects:			
Child status	5,560.0	2	2,780.0
Education	2,260.9	2	1,130.5
Income of husband:			
Model II:			
(Linear component)	(7,206.2)	(1)	(7,206.2)
(Non-linear component)	(654.3)	(4)	(163.6)
Or Model III:			
(Log-linear component)	(5,588.1)	(1)	(5,588.1)
(Non log-linear component)	(2,272.4)	(4)	(568.1)
Total = Model I	**7,860.5**	**5**	**1,572.1**
Residual = interaction effects:			
(Second-order interaction effects)	(2,049.5)	(24)	(85.4)
(Third-order interaction effects)	(457.8)	(20)	(22.9)
Total	**2,507.3**	**44**	**57.0**
Total	**18,188.8**	**53**	

These estimates obviously look reasonable when compared with the actual value of 10.14 per cent but they have only been shown here by way of illustration and it will be left to the reader to see that when the level of the income variable is changed the estimated participation rates can become negative.

The last part of this section has shown that when all the factor/levels are represented by dummy variables the model can produce a "good fit" of the data together with providing easily interpreted results. This was clearly the case when it was applied to the data used in this study. However, there are certain shortcomings in the method. Interaction terms are likely to be ignored and estimates can easily be obtained which are outside the theoretical bounds of the variable. Furthermore these problems would appear to be aggravated when it is wished to include one of the factors as a continuous variable.

The next section will now consider two ways in which one of the disadvantages of the models discussed in this section, namely the problem of constraining the estimates to lie between the acceptable range of zero to 100 per cent, can be overcome. It will also be shown that these "transformations" also reduce the size of the interaction effects thus suggesting that the two main disadvantages of the simple additive model mentioned earlier may not be independent but simply different aspects of the same problem – that of formulating a more conceptually correct labour force participation rate model.

IV. PROBIT AND LOGIT TRANSFORMATIONS

Probit Transformation

The factors which influence the decision of individuals to enter the labour force are numerous. But it can be assumed that, in the last analysis it is on the individual's (or family's) reaction to the social and economic environment that the ultimate decision rests.

If these environmental pressures are thought of as stimuli and if the intensity of a given stimulus can be represented on a continuous scale, then the probit approach is premised on the assumption that each individual in the population has a threshold[20] value of the stimulus such that for a higher value he will always respond and for a lower value he will never respond. Let the distribution of these threshold values in the population be described by a probability density function $f(x)$, $(-\infty < x < \infty)$. The probability that a person chosen at random from a population has a threshold value in the range x_0, $x_0 - \Delta x$ is given by $f(x_0)\,\Delta x$ and the proportion of the population with a threshold value above x_0 is given by

$$P(x_o) = \int_{x_o}^{\infty} f(x)\,dx \qquad (10)$$

If the population is the population of married women, the stimulus is the husband's income and the response is withdrawal from (or failure to enter) the labour force. An individual's threshold value, then, is the level of the husband's income above which the individual will withdraw from (or will not enter) the labour force and equation (10) gives the participation rate of married women with husband's income greater than x_o.

[20] The term "threshold" was originally used by psychologists who can claim to have first used probit analysis although they did not recognise its full potential at the time. Biometricians who developed the technique prefer to use the term "tolerance" in biological assay work.

The probit technique now assumes that the underlying distribution of threshold values is a normal distribution $N(\mu\sigma^2)$ with mean μ and variance σ^2, or can be made so by some transformation of the stimulus (i.e. income variable).

We then have, for equation (10):

$$P(x_o) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_o}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad (11)$$

The probit, y, of a proportion P is defined by Finney[21] as "the abscissa corresponding to a probability P in a normal distribution with mean 5 and variance 1" or, more precisely:

$$P = \frac{1}{\sqrt{2\pi}} \int_{y-5}^{\infty} e^{-\frac{1}{2}u^2} du \qquad (12)$$

If the right hand sides of equations (11) and (12) are now equated it can be shown that:

$$y - 5 = \frac{x_o - \mu}{\sigma}$$

or more generally:

$$y = a + bx_o$$

where a and b are constants.

Thus if the observed proportions are replaced by their probits (obtainable from tables) we can fit a model which is linear over the range $-\infty$ to $+\infty$ by the standard techniques: we can then reconvert the fitted model to proportions as required. Observe that the probits corresponding to proportions of 0 and 1

[21] See D.J. Finney, *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve*, *op. cit.*

do not exist. Chart 2 illustrates the relationship between a percentage or proportion and its probit.

There is, however, no reason why probit analysis should be confined to a single independent variable such as the income of the husband. Indeed, referring back to equation (5) on page 16, a new equation can be postulated in which the probit of the labour force participation rate of married women is now assumed to be made up of a constant term plus a contribution associated with the child status of the family, plus a contribution associated with the wife's level of education and a contribution which will vary with the income of the husband. Three models, using the probit transformation were examined: Models IA, IIA, and IIIA which correspond in form to the three simple additive models discussed in the previous section. In Model IA the income of the husband is allowed to take a free form defined by dummy variables so that no normalising assumption is required. In Model IIA income takes a linear form which assumes that the distribution of thresholds is normal along the natural income scale. In Model IIIA income is included in a log form which assumes that the distribution of thresholds is normal along a scale defined by the logarithm of the income.

It may be appropriate at this stage to consider briefly what would, in the example used in this Study, be a normalising transformation for the income scale. Table 1 gives some clues here. The shape of the curve of labour force participation rates along the income scale can, in all six cases, be seen to be a reverse s-shape. The participation rate is fairly stable at a high level over the first three income groups, from say, no income to an average of $4,000 a year. From this point it falls fairly rapidly to the class of $7,000 - $9,999 and then the rate of decrease in the participation rate slows down between this class and the $10,000 and over class. Now it has already been mentioned that the open ended class has a probable average income value of something approaching $16,000 a year, so that it can be seen that for the reverse s-shape to be nearly symmetrical, as it has to be to approximate to a normal (reversed) ogive, the "tail" of the distribution in the upper income range has to be considerably shortened. This can be done quite easily by taking the logarithm of the income as the measurement scale over which the distribution is viewed. It was for this reason that the logarithm of the income has been introduced into models examined in this Study. In addition to the three regression analyses a full analysis of variance was carried out on the transformed variable. The results of the analysis of variance are given in Table 8, and that of the three regression analyses in Table 9. Comparing the results of the analysis of variance applied to the original participation rates given in Table 3, with that of their probit transformation, it can be seen that the interaction effects which were relatively large when the simple additive model was used have been reduced in size by the use of the probit transformation. However, if the three-factor interaction mean square is again taken as an estimate of σ^2 then the two-factor interactions are still significant.

The amount of variation explained by each of the three regressions was in all cases greater than when the corresponding model was applied to the untransformed data. When income was included as a dummy variable the value of R^2 rose from 0.862 in Model I to 0.902 in Model IA so that in this form, and using the probit transformation, less than 10 per cent of the variation now remains unexplained. With income as a linear variable in Model IIA, R^2 was 0.864 and when log income was used in Model IIIA it was 0.759. In the latter two cases the improvement in the explained variance was 3.8 and 2.2 percentage points respectively. The reason why log income should fail to give a superior result over the

CHART-2

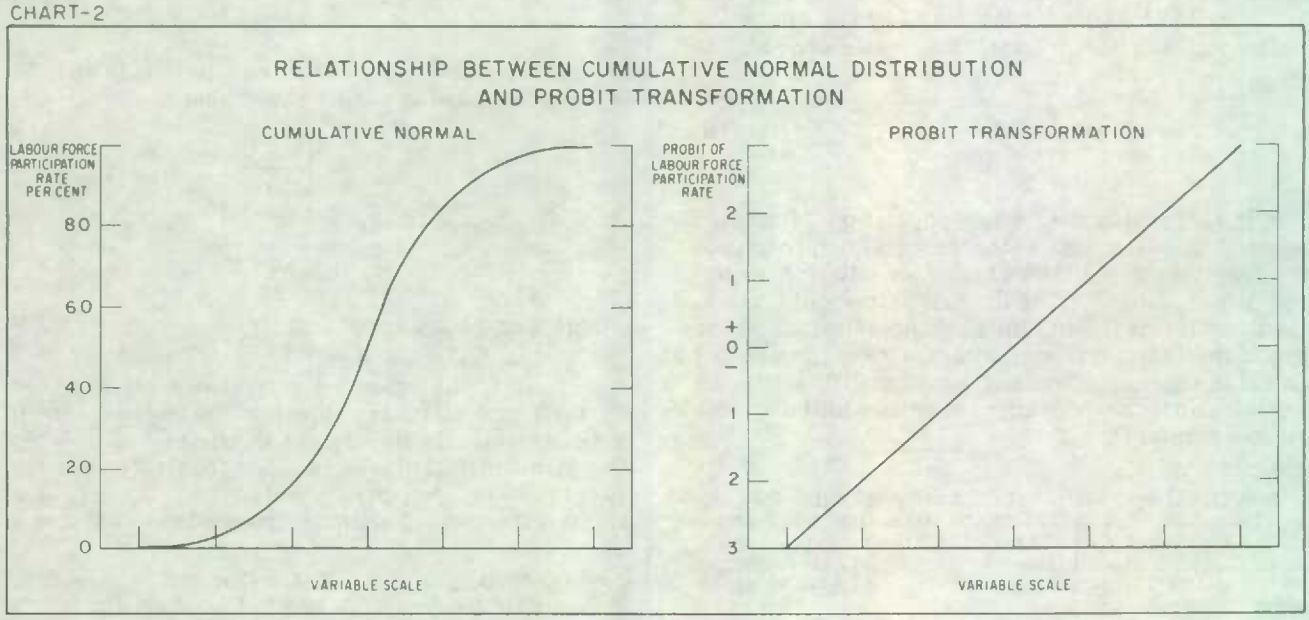


TABLE 8. Analysis of Variance with Probit Transformation

	Sum of squares	Degrees of freedom	Mean squares	Variance ratio[1]
Main effects:				
Children status	5.7872	2	2.8936	204.78
Education	1.7584	2	0.8792	62.22
Income of husband	8.7189	5	1.7438	123.41
Second-order interactions:				
Child status/education	0.4992	4	0.1248	8.83
Child status/income	0.3929	10	0.0393	2.78
Education/income	0.5940	10	0.0594	4.20
Third-order interaction				
Child status/education/income	0.2826	20	0.0141	
Total	**18.0330**	**53**		

[1] Obtained by dividing the mean squares of the main effects and second-order interaction effects by the mean square of the third-order interaction.

TABLE 9. Regression Equations of Probits of Labour Force Participation Rates of Married Women in Urban Ontario

		Coefficients of					
	Constant	Child status		Education		Husband's income	
Model I A							
						$'000	
$R^2 = 0.0919$	4.436	Children under 6	– 0.445	Elementary	– 0.231	10+	– 0.666
$N = 54$		No children under 6	+ 0.114	Secondary	+ 0.021	7-10	– 0.396
		No children	+ 0.332	University	+ 0.209	5- 7	+ 0.028
						3- 5	+ 0.267
						1- 3	+ 0.399
						0- 1	+ 0.369
(Standard error of coefficients)			(0.067)		(0.067)		(0.095)
Model II A							
$R^2 = 0.8641$	4.902	Children under 6	– 0.445	Elementary	– 0.231	Continuous variable	
$N = 54$		No children under 6	+ 0.114	Secondary	+ 0.021	$'000 natural scale	– 0.076
		No children	+ 0.332	University	+ 0.209		
(Standard error of coefficients)			(0.075)		(0.075)		(0.006)
Model III A							
$R^2 = 0.7591$	4.844	Children under 6	– 0.445	Elementary	– 0.231	Continuous variable	
$N = 54$		No children under 6	+ 0.114	Secondary	+ 0.021	$'000 and Log_2 scale	– 0.210
		No children	+ 0.332	University	+ 0.209		
(Standard error of coefficients)			(0.100)		(0.100)		(0.025)

linear income form is again not difficult to see but an explanation of this will be left to the next section. The estimates obtained from the probit regression equations were converted back to proportions and the differences between these and the original observations were used to obtain a measure of how much of the variation in the participation rates had been explained by the probit models. Table 10 summarises the results of this exercise and compares them with the measures of explained variation obtained directly from the probit and also from the simple additive regression equations.

TABLE 10. Percentage of Total Variation Explained by Six Regression Models

Form of dependent variable from which variation is calculated →	Form of dependent variable in regression equations		
	Proportions[1]	Probits	
	Proportions[1]	Probits	Proportions[1]
	per cent		
Form of income variable:			
Dummies	86.2	90.2	90.4
Linear	82.6	86.4	88.2
Logarithmic	73.7	75.9	73.3

[1] Labour force participation rates.

From this table it can be seen that the amount of the variation in the original participation rates which is explained in the participation rates obtained from the estimated probits is, in the case of Model IA and IIA, even greater than the amount of variation in the original probits explained by the probit regression. But perversely, when the logarithm of income was used, the efficiency of the estimating equation falls (when the probits are converted back to proportions) to even less than that of the equivalent additive model.

The results obtained from Model IA will now be used to show how the probit transformation works, and why on *a priori* grounds, it should give results superior to those obtained from the equivalent simple additive model. On page 19 it was shown that the labour force participation rate of married women in urban Ontario who had children under six, elementary school education, and husbands with an income between $5,000 and $6,999 was estimated, from the simple additive model using dummies for the income variable, at 9.70 per cent compared with an actual of 10.14 per cent. But when, keeping the other variables unchanged, the husband's income is in the $10,000 and over class the estimated participation rate of the wives fell to the absurd figure of -10.37 per cent compared with an actual of 7.53 per cent.

Now in order to obtain the estimates for the same socio-economic groups obtained from the probit transformation in Model IA, it is first necessary to calculate the estimate of the appropriate probits. From Table 9 these are found to be:

$$4.436 - 0.445 - 0.231 + 0.028 = 3.788$$

and

$$4.436 - 0.445 - 0.231 - 0.666 = 3.094$$

which convert to 11.27 per cent and 2.83 per cent respectively. Taken together these estimates are now much closer to original ones. But more important, from the point of view of an example in the use of probits, is the effect that this transformation has on the change in the estimated labour force participation rate arising from a change in the income variable. It will be remembered that the conclusion obtained from the simple additive model was that the average labour force participation rate of married women with husbands earning over $10,000 a year was just over 20 percentage points lower than that of wives with husbands earning between $5,000 and $6,999 a year and that this applies regardless of the other social or economic characteristics of the married women. The probit model, on the other hand, says that there is a fall of 0.694 in the probits of the participation rates between the same income groups. And, as was seen in the above example, this only represented a drop of about 8½ percentage points when the higher estimate was just over 11 per cent. But if the higher estimate had been equivalent to a participation rate of 50 per cent (equal to a probit of 5) then a fall of 0.694 in the probit would be equivalent to a drop in the participation rate of more than 25 percentage points.

The next section will now examine an alternative transformation which, although doing essentially

the same job as the probit, is derived from a completely different set of assumptions. A comparison of the two transformations will naturally be left to the end of the section.

Logit Transformation

An alternative approach, then, to the problem of non-additivity and that of constraining the estimated participation rate to keep between the bounds of reality, is to consider the variable directly rather than the forces which generate the variable as is done in the case of the probit.

If P is again the labour force participation rate then it would seem reasonable to assume that the ratio of the rate of change of P to the value of P is a function of how far P can go before reaching its upper limit. If the relationship is assumed to be linear and the independent variable is again assumed to be the income of the husband (I) we have the simple differential equation:

$$\frac{dP/dI}{P} = b(1 - P) \tag{13}$$

whose solution is:

$$\log_e \frac{P}{1-P} - bI + k = 0 \tag{14}$$

where k represents the constant of integration. Rearranging and changing the notation slightly equation (14) reduces to:

$$P = \frac{1}{1 + e^{-(a+bI)}} \tag{15}$$

The curve described by this formula has been called the logistic curve and Berkson, who first used the linearising transformation given by the first term in equation (14), coined the word logit to describe the expression:

$$\log_e \frac{P}{1-P}$$

Before looking at the use of this logit transformation in the models examined in this study, some points of interest arise out of the nature of the logistic curve. From equation (15) it can be seen that as the variable I tends to become a large positive number, the denominator, $1 + \exp\{-(a+bI)\}$, tends to one so that P also tends to one, while as I tends to become a large negative number the denominator becomes increasingly large and P tends to zero. The logit transformation will therefore ensure that estimates of P will always lie between 0 and 1. Moreover, like the probit, the logit transformation turns a bounded variable into an unbounded one which is linear in form. Thus from equation (15):

$$\log_e \frac{P}{1-P} = a + bI$$

There is one other similarity between the logit and the probit. It can be shown that the rate of change in P reaches a maximum when P is equal to 0.5 and that it is symmetrical. The logistic curve is therefore very similar to the cumulative normal. They are both asymptotic to zero and one, and are symmetrically s-shaped with a maximum gradient when P = 0.5.

As in the probit model described earlier, the expected shape of the logistic curve, when the income of the husband is the independent variable on a continuous scale, is that of a reverse s-shape since the labour force participation rates of married women are expected to decline as the income of the husband rises. This makes no difference to the foregoing discussion since the curve is symmetrical. However, to meet the condition the actual form of the curve which was estimated is:

$$P = 1 - \frac{1}{1 + e^{-(a+bI)}}$$

which reduces to:

$$P = \frac{1}{1 + e^{(a+bI)}} \tag{16}$$

The logit transformation is therefore unchanged since the only effect of reversing the shape of the curve is that the signs of the parameters are also reversed.

The regression equations of the logits of the labour force participation rates took the same three forms as those for the simple additive models and those based on the probit.

The coefficients obtained from the results of these three regression models are given in Table 11. The proportions of the total variation explained were all larger than those obtained in the corresponding probit models but in no case was the difference more than one half of one percentage point. When the "explained" variation using the logit transformation is calculated on the original units, the differences between the logit and probit methods are in favour of the probit in two out of three models and in favour of the logit in the other one. However, the differences were again so small as to be negligible.

The last comparison to be made, then, is between the estimates themselves. It is just conceivable that the allocation of the sum of squares could be nearly the same, using the two transformations, but with the predicted values for factor/level combinations noticeably different. Here again, however, there is no evidence to show that this is the case and it must be concluded that the probit and the logit transformations do essentially the same job. To illustrate this point, the estimated labour force participation rates obtained using the probit and logit transformations are given in Table 12 for the first six and last six observations.

TABLE 11. Regression Equations of Logits of Labour Force Participation Rates of Married Women in Urban Ontario

	Coefficients of					
Constant	Child status		Education		Husband's income	
Model I B						
					($'000)	
$R^2 = 0.9067$	– 0.954	Children under 6	– 0.772	Elementary	– 0.377	10 +	– 1.163
N = 54		No children under 6	+ 0.204	Secondary	+ 0.029	7 - 10	– 0.687
		No children	+ 0.568	University	+ 0.349	5 - 7	+ 0.058
						3 - 5	+ 0.466
						1 - 3	+ 0.686
						0 - 1	+ 0.639
(Standard error of coefficients)			(0.112)		(0.112)		(0.159)
Model II B							
$R^2 = 0.8685$	– 0.144	Children under 6	– 0.772	Elementary	– 0.377	Continuous variable	
N = 54		No children under 6	+ 0.204	Secondary	+ 0.029	$'000 natural seale	– 0.131
		No children	+ 0.568	University	+ 0.349		
(Standard error of coefficients)			(0.127)		(0.127)		(0.010)
Model III B							
$R^2 = 0.7596$	– 0.246	Children under 6	– 0.772	Elementary	– 0.377	Continuous variable	
N = 54		No children under 6	+ 0.204	Secondary	+ 0.029	$'000 on Log_2 scale	– 0.364
		No children	+ 0.568	University	+ 0.349		
(Standard error of coefficients)			(0.172)		(0.172)		(0.044)

TABLE 12. Probit and Logit Estimates Compared

Child status, education and income of husband	Estimated participation rates	
	Probit	Logit
	per cent	
Children under 6:		
Elementary:		
$10,000 and over	2.8	3.7
7,000 - $9,999	5.1	5.8
5,000 - 6,999	11.3	11.5
3,000 - 4,999	16.5	16.3
1,000 - 2,999	20.0	19.5
Under $1,000	19.2	19.8
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
No children:		
University:		
$10,000 and over	24.6	23.2
7,000 - $9,999	33.8	32.6
5,000 - 6,999	50.3	50.5
3,000 - 4,999	59.7	60.6
1,000 - 2,999	64.7	65.7
Under $1,000	63.6	64.6

The question now arises, given that such a transformation is called for, which of two alternative transformations to use when it is known that the results and their immediate interpretation will be so little different. But besides these purely practical considerations there is still the fact that the assumptions on which the two transformations are based are fundamentally different, and in the last analysis it is in terms of these assumptions that the results must be interpreted. This is particularly so when the assumptions purport to describe some causal relationship whether concerned with physical or biological processes or, as is the case in this Study, with the physiological or psychological makeup of human beings. In his paper, "Why I Prefer Logits to Probits" Berkson had this to say:

> "If it is seriously believed that there is some physical property more or less stably characterising each organism, which determines whether or not it succumbs, then it is justifiable to advance the hypothesis of a distribution of tolerances. In that case one should be prepared to suggest the nature of this characteristic so that the hypothesis may be capable of corroboration by independent experiments. If on the other hand the formulation is only that of a "mathematical model", to guide the method of calculation, then it would seem more objective and heuristically sounder not to create any hypothetical tolerances, but merely to postulate that the proportion of organisms affected follows the integrated normal function. I am interested in the slope of the dosage mortality line as a "rate", of the objectively observed increase of mortality with increase of dosage, not as a standard deviation of hypothetical tolerances of the animals. I should of course be very much interested in the last, if tolerance of the animals is what I was observing and studying. But we are not dealing with measured tolerances, we are dealing with a dosage mortality curve, and when my probitistic friends present a standard deviation of tolerances, they may be asserting a substantial quantity for the variability of something that in fact does not exist at all. I once had a teacher of philosophy who employed the Socratic method in class. When a student gave a simple and especially plausible explanation of a very complex social phenomenon the professor said, "Please, sir, do not make up history." I should like to ask mathematical statisticians when they formulate mathematical models to please not make up physics."

It is for this reason that the present author prefers the logit transformation based on the logistic curve. Put quite simply it attempts to describe the relationship between the effect of some stimulus and the stimulus itself without attempting to describe the causal process.

V. GENERALISED LOGISTIC CURVE

In the last section it was seen that the use of either of two transformations, the probit and the logit, generally improved the predictive power of the models examined, bearing in mind, of course, that the simple additive model had already explained 86 per cent of the total variation in the participation rates. It was also seen that the probit and logit transformations were so little different from each other in practice if not in concept. However, as has been suggested above, the logit transformation is based on assumptions which are, perhaps, more appropriate to the type of analysis being undertaken, and it is for this reason only that the logit, or rather the logistic curve has been used for the further developments introduced in this section.

Consider again the participation rate P, of a given population. It seems reasonable to assume that, instead of having limits of 0 and 1, P will in practice have lower and upper limits, L and U, such that $0 \leqslant L < P < U \leqslant 0$.

If it is now further assumed that the rate of change in P, with respect to I, relative to its position measured from its lower limit, P – L, is proportional to the remaining fraction of its total range, then the differential equation which describes the relationship is:

$$\frac{dP/dI}{P-L} = b\,\frac{(U-P)}{(U-L)} \qquad (17)$$

This differential equation can now be solved to show that:

$$P = L + \frac{U-L}{1+e^{-(a+bI)}} \qquad (18)$$

where a is a constant.

However, since in the example used in this Study, the shape of the curve is the reverse of the more common pattern – i.e. the participation rate declines as the income rises so that the upper asymptote is associated with a low income – the form of equation required is that of (18) but with L and U interchanged, or:

$$P = U - \frac{U-L}{1+e^{-(a+bI)}} \qquad (19)$$

Now consider what happens to equation (19) when I is assumed to take certain values. As the exponent becomes increasingly large the denominator of the last term on the right hand side of equation (19) approaches 1 so that P approaches L, the lower limit. But when the exponent becomes a large negative number the same denominator becomes very large and the last term in the equation approaches zero so that P then approaches U, the upper limit. Equation (19) then, which defines a generalised version of the logistic curve, has a lower limit L and an upper limit U. It can of course

be readily seen that if U is set equal to one and L to zero, equation (19) is identical to equation (16).

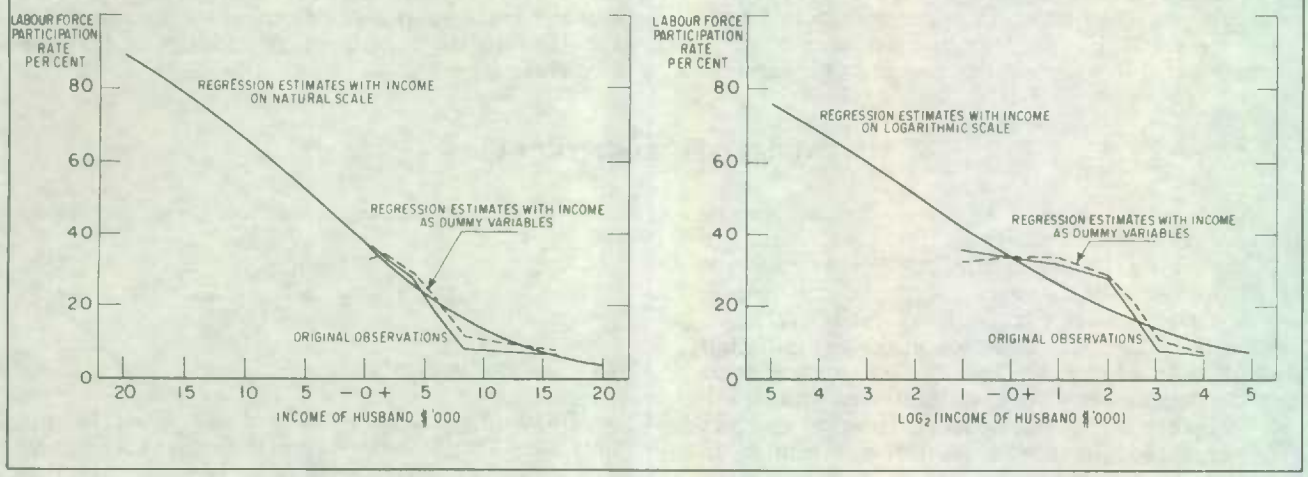
It can further be shown that the curve is again symmetrically s-shaped with a maximum slope when:

$$P = (L + U)/2$$

The reasoning behind the foregoing can best be illustrated by examining the two graphs in Chart 3, below, on each of which are plotted the original six observations (one for each income level) for the group of married women with children under six years of age who have had a university education. One graph is drawn with the income of the husband on its natural scale and the other with income on a logarithmic scale. Also plotted on each graph are the estimates obtained from the regression model using the logit transformation in which income was represented as a dummy variable. For estimates obtained from the other two models—with income on its natural scale and with the logarithm of the income—the fitted curves have been plotted on the appropriate graph but drawn so as to go outside the range of the original observations. This illustrates the way in which they are forced to seek asymptotes at 0 and 1 when a continuous scale is used. This, of course, does not happen when income is allowed to take a free form and is represented in the equation by a set of dummy variables. It can also be seen on this chart why the use of the logarithm of the income variable did not improve the fit of the equation even though, as may be observed, the curve of the original data and "dummy" coefficients are clearly not symmetrical on the arithmetic scale. Yet the same original data plotted against the logarithm of the husband's income indicates that the best fit is still likely to be obtained with income included in this way, and with a curve which is theoretically symmetrical and s-shaped, but with asymptotes, particularly the upper asymptote, lying well inside the theoretic bounds of zero and one.

CHART-3

ACTUAL PARTICIPATION RATES AND ESTIMATES OBTAINED FROM LOGIT REGRESSION MODELS FOR WOMEN WITH CHILDREN UNDER 6 YEARS OF AGE AND WITH UNIVERSITY EDUCATION BY THE INCOME OF THE HUSBAND



When more than one population group is under investigation the assumption, implicit in equation (19), that U and L are constants is unrealistic. Rather it must be assumed that they are functions of the factors which define the population groups, e.g. education of the wife, and child status. With three levels of the child status factor and three educational attainment levels there are therefore nine pairs of asymptotes to be estimated. But with this additional assumption is it also reasonable to assume that "a" and "b" in equation (19) are also constant between population groups? The effect of changing "a" is to change the location of the mid-point of the curve along the stimulus (income scale) while the effect of changing "b" is to change both the mid-point of the curve and its slope.

The assumption made in this Study is that wives with different social characteristics will not respond in the same way to the same level of the husband's income, i.e., it is expected that each curve will be located at a different point along the income scale so that "a" must be allowed to vary. But around this point it is assumed that the effect on the participation rate of the same proportionate change in income will be the same for all population groups, i.e., the slope of the curve at the mid-point will be the same and therefore "b" can be left as a constant. It is implicit in the second assumption that the logarithm of the husband's income will be used in the model.

In defining the form that the asymptotic values take it would be possible to go back to the original assumptions made at the beginning of this Study and assume that changes in the asymptotes, as the factor levels are changed, are additive. This would be the simplest thing to do but unless interaction effects were included it would in theory permit the asymptotes to go outside the range of 0 and 1. Instead, the form of the model employed is one which constrains both the upper and lower set of asymptotes to keep within the bounds of 0 and 1 by assuming that they also are defined by two multivariable logistic curves of the form:

$$A_L = \frac{1}{1 + e^{-[b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4]}} \quad (20)$$

$$A_U = \frac{1}{1 + e^{-[b_5 + b_6 x_1 + b_7 x_2 + b_8 x_3 + b_9 x_4]}} \quad (21)$$

Where A_L and A_U are the set of lower and upper asymptotes respectively and where

$x_1 = 1, x_2 = 0$ when there are no children in the family

$x_1 = 0, x_2 = 1$ when there are children under six in the family

$x_1 = -1, x_2 = -1$ when there are children, but none under six, in the family

$x_3 = 1, x_4 = 0$ for completed elementary schooling or less

$x_3 = 0, x_4 = 1$ for some high school or completed high school

$x_3 = -1, x_4 = -1$ for some university education or with degree

The reason for the special form in which the dummy variables for child status and education are expressed is that it was also desirable from the point of view of interpretation to be able to view the coefficients for each factor as deviations from the overall mean. But because the method of estimation which had to be used was not that of "non-linear least squares", the dummies themselves were constrained to produce results which could be directly interpreted in this way. Thus it is required that three coefficients, say, a_1, a_2 and a_3 are to be equated to zero, or:

$$-(a_1 + a_2) = a_3$$

then the expression incorporating three normal dummy variables, say:

$$a_1 Z_1 + a_2 Z_2 + a_3 Z_3$$

becomes:

$$a_1 Z_1 + a_2 Z_2 - (a_1 + a_2) Z_3$$

or:

$$a_1 (Z_1 - Z_3) + a_2 (Z_2 - Z_3)$$

so that the three original dummy variables have been reduced to:

$$x_1 = (Z_1 - Z_3)$$

$$x_2 = (Z_2 - Z_3)$$

which will take the values indicated above.

Since no constraints need be placed on the expanded form of the constant a in equation (19) and if b in that equation is replaced by b_{15} then equation (19) is:

$$P = A_U - \frac{A_U - A_L}{K} \quad (22)$$

where A_U and A_L are as defined in equations (20) and (21) and

$$K = 1 + e^{-[b_{10} + b_{11} x_1 + b_{12} x_2 + b_{13} x_3 + b_{14} x_4 + b_{15} I]} \quad (23)$$

in which I is the logarithm of the income of the husband.

Equation (22), with an error term added, is the model whose parameters were estimated from the data used in this Study.

The set of normal equations obtained from equation (22) will contain non-linear terms so that the classical least squares method cannot be employed to provide estimates of the parameters. Moreover, no linearising or other simple transformation is available to assist in the estimation of the parameters. However, techniques are available for the estimation of parameters of models whose "normal equations" contain non-linear terms and one of these[22] has been employed to do this. The technique is one which requires an initial set of estimates to be provided and which then uses an iterative process to successively approximate to the least squares solution. It is, however, well beyond the scope of this Study to go into any detail of the method.

Before proceeding to the results obtained using this model it should be mentioned that one very real problem was encountered in the estimation of the parameters. This is a problem which is not uncommon in the estimation of non-linear models in general and one which has been previously met in an examination of methods used to estimate the logistic function in particular.[23] With a complex model the surface described by the function of the residual sum of squares – which is to be minimised – may be so pitted with craters that any set of initial estimates will cause the iterative process to converge on a local minimum rather than the absolute mini-

---

[22] See D.W. Marguardt, *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*, Journal of the Society of Industrial Applied Mathematics, Vol. II, No. 2, June, 1963.

[23] See F.R. Oliver, *Methods of Estimating the Logistic Growth Function*, Journal of the Royal Statistical Society, Series C, Applied Statistics, Vol. 13, No. 2.

mum. Indeed one set of quite plausible initial estimates of the parameters of equation (22) gave rise to a "solution" which had a "least squares" larger than the sum of squares about the mean.

However, this situation can be dealt with, even though it does introduce an element of subjectivity to the method. The final estimates to equation (22), which will be discussed later, were obtained by first providing an initial set of estimates which were then modified by the iterative process employed by the method until a "least squares" solution was reached. At the same time the computer was programmed to print out not only the standard output of the original values, the predicted values and the differences, but also the upper and lower asumptotes and a measure of the rate of decline, as the income rose, from the upper to the lower asymptote. From an inspection of this information it was then not too difficult to ascertain whether the given "least squares" solution was associated with a local minimum (which could be improved upon by selecting a new set of initial estimates) or with a value at or close to the true minimum which could be accepted. This procedure was repeated until a "satisfactory" solution was found. This is obviously where subjectivity enters into the process. For it does not necessarily follow that values of the residual sum of squares which are very close to each other are always obtained from estimates of the parameters which are themselves sufficiently close as to give rise to the same interpretation. Fortunately this latter problem was not encountered in this study, at least not in the region of the very low values of the residual sum of squares.

What now follows then is a brief description and discussion of the results which were obtained based on the least "least squares" solution found.

As was to be expected, a very high proportion of the variation in the participation rates was explained – 93.85 per cent – which, despite the loss in the number of degrees of freedom because of the increase in the number of parameters in the model, gave a standard error term lower than that obtained in any other model.

Turning now to the actual estimate of the parameters these are given in Table 13 together with their estimated standard errors. It is important now to remember that the parameters other than the constant term associated with upper and lower asymptotes have been constrained to measure the deviations from the average for the given factor level combination. From this Table it can be seen that while the estimates of the parameters associated with the upper asymptote are all significant[24], supporting the view that the upper limit of the participation rates will vary from one socio-economic

[24] Because of possible correlations between the parameter estimates the standard t test may give rise to an overstatement of the significance of the estimates.

TABLE 13. Parameter Estimates and Standard Errors in Non-linear Regression Model

	Parameter	Estimate	Standard error of estimate	t-value
Upper asymptote	b_0	0.328	0.0535	6.131
	b_1	0.689	0.0754	9.138
	h_2	– 0.200	0.0681	– 2.937
	b_3	0.477	0.0740	6.446
	b_4	– 0.154	0.0671	– 2.295
Lower asymptote	b_5	2.455	0.3559	6.898
	b_6	0.592	0.4368	1.355
	b_7	– 0.237	0.3074	– 0.771
	b_8	– 0.200	0.2885	– 0.693
	b_9	– 0.046	0.3044	– 0.151
Path between upper and lower asymptotes	b_{10}	8.142	2.1985	3.703
	b_{11}	– 0.992	0.4083	– 2.430
	b_{12}	0.074	0.4252	0.174
	b_{13}	– 1.465	0.4637	– 3.159
	b_{14}	0.094	0.3595	0.261
	b_{15}	– 3.030	0.8172	– 3.708

group to another, there is no such support for a similar view that the lower limit is also subject to variation. It is beyond the scope of this Study to discuss the empirical analysis in detail, but when it is remembered that it was assumed the average income for the $10,000 and over income group was $16,000 and that this average value was assumed to obtain for each of the nine groups (not, on reflection, a very plausible assumption) it would take a much more detailed study to establish whether significant variation in the lower asymptote was present, or should be assumed to be present.

The last group of parameters given in Table 13 determine the "path between the upper and lower asymptote". Not all the estimates of these parameters can be judged to be significant, but the same qualification made above with regard to the estimates associated with the upper and lower asymptotes applies: non significance again means not significantly different from the average. It is perhaps interesting to note, at this stage, just what the coefficients b_{11}, b_{12}, b_{13} and b_{14} do to the path to be followed between the asymptotes. The income at which the participation rate is exactly half way between the upper and lower asymptote is, from equation (19), equal to -a/b. But it was later argued that when more than one group was being considered, with an assumption that their asymptotes would vary, that at the same time it should be assumed that the hitherto constant term "a" should also be allowed to vary. It was for this reason that four dummy variables were introduced into this part of the expression, to represent the nine sub groups, with associated coefficients b_{11} to b_{14}. The income levels at which the participation rate is now half way between the upper and lower asymptote for each of the groups can now be obtained directly from the expression:

$$\text{Log}_2 I_{(U+L)/2} = \frac{-(b_{10} + b_{11}x_1 + b_{12}x_2 + b_{13}x_3 + b_{14}x_4)}{b_{15}}$$

These income levels, $I_{(U+L)/2}$, and the associated upper and lower asymptotes for each of the nine child status/education groups, are given in Table 14 and the full set of results are graphically illustrated in Chart 4. In Chart 4, as an aid to interpretation, the participation rates have been plotted against the natural income scale; the symmetry in the curve which would have been present if the logarithm of the income had been used has therefore been lost.

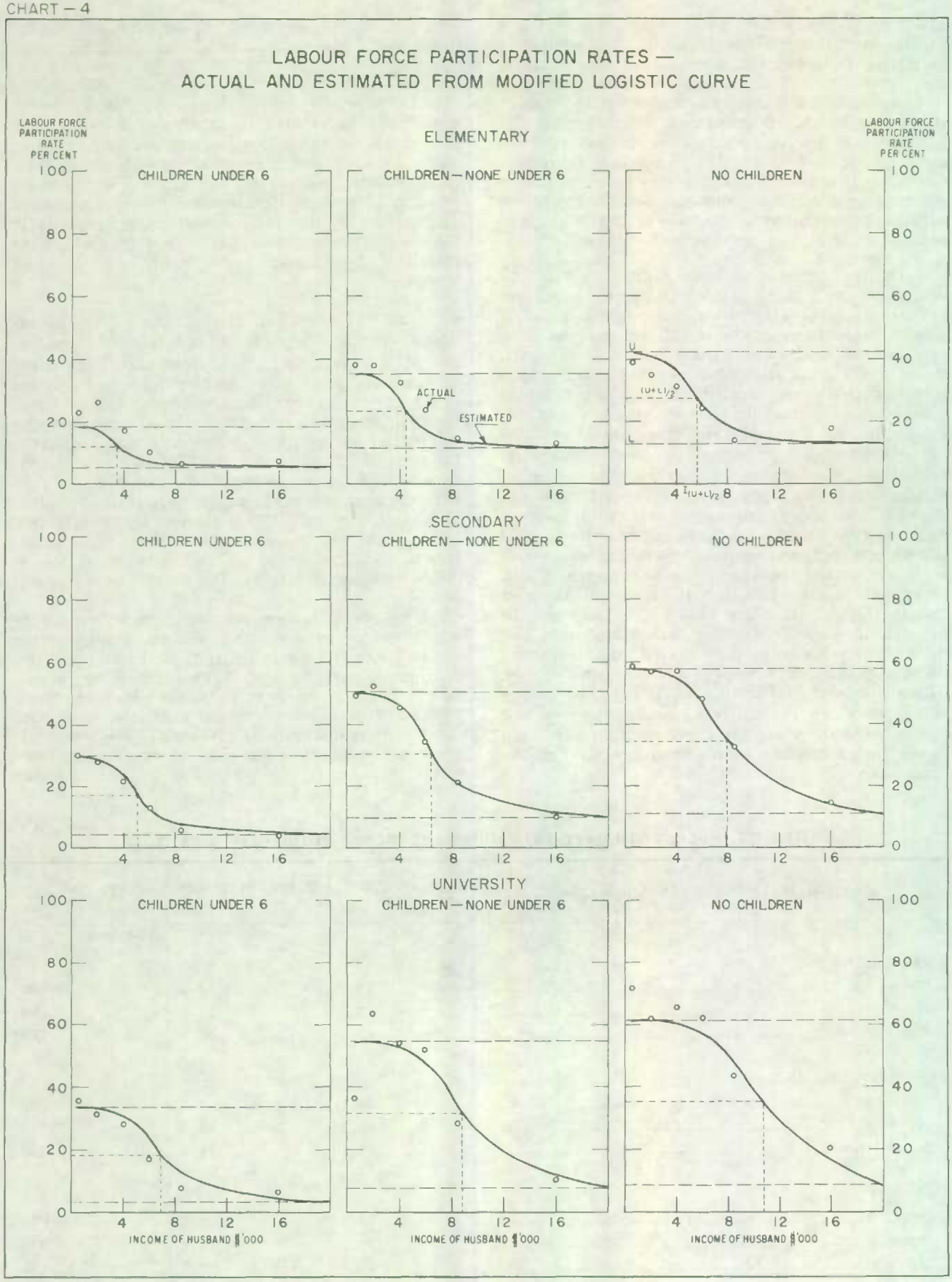
Table 14 and Chart 4 illustrate how the operation of the coefficients b_{11} to b_{14}, referred to above, is that of a shift mechanism which places the curve at some point along the income scale independent of the upper and lower limits. Thus it can be seen, for example, that although the upper asymptote for the participation rate of wives with some children, but none under six, and with secondary school education (50.6 per cent) is higher than that of wives with children under six but with university education (33.3 per cent) it is estimated that the income at which the wife's participation rate is most sensitive to a change in the husband's income is lowest ($6,670 compared with $7,000) for the former group.

The question therefore arises as to what economic, as opposed to mathematical, interpretation can be placed on the position of the curve along the continuous variable scale? The critical income of $6,670 is different, even if by not very much, from $7,000, but what does it mean? Similarly, the labour force participation rate of wives who have a secondary school education, but have no children, is estimated to be most sensitive to change, as a result

TABLE 14. Upper and Lower Asymptotes and Income of Husband at (U+L)/2

Factor/level, education and child status	Upper asymptotes U	Lower asymptotes L	$I_{(U+L)/2}$
			$
Elementary:			
Some < 6	18.3	5.5	3,660
None < 6	35.3	11.7	4,600
No children	42.2	13.0	5,660
Secondary:			
Some < 6	29.7	4.7	5,220
None < 6	50.6	10.2	6,670
No children	57.8	11.4	8,090
University:			
Some < 6	33.3	3.6	7,000
None < 6	54.8	7.8	8,930
No children	61.8	8.7	12,690

CHART – 4


LABOUR FORCE PARTICIPATION RATES —
ACTUAL AND ESTIMATED FROM MODIFIED LOGISTIC CURVE
LABOUR FORCE PARTICIPATION RATE PER CENT
ELEMENTARY
CHILDREN UNDER 6
CHILDREN—NONE UNDER 6
NO CHILDREN
ACTUAL
ESTIMATED
U
(U+L)/2
L
SECONDARY
UNIVERSITY
100
80
60
40
20
0
4
8
12
16
INCOME OF HUSBAND $'000

of a change in her husband's income, when that income is about $8,100 a year. If there are children under six in the family, the corresponding income level is estimated to be $5,200. It is not sufficient to say that the wife with no children is better able to combine going out to work with the normal duties of running a house, and will therefore still choose to do so even when the husband's income is at a fairly high level. For what is being considered is the income effect, either direct or indirect: the direct effect of being more easily available for work outside the home has already been taken care of in the positioning of the upper and lower asymptotes.

Similarly, the reason why the income of the husband at which the wife's labour force participation rate is most sensitive to change increases as her standard of education increases cannot be said to be due directly to the fact that the higher her education the greater the job opportunities open to her. this effect is again reflected in the upper and lower limits placed on her participation rates. And moreover, as has been shown above, the ranking of the asymptotes does not necessarily agree with the ranking of this critical income value.

It could further be shown that while a change in the position of the curve along the income scale will obviously change the participation rate associated with a given income level it will not change the average labour force participation rate for that particular child status/educational attainment group. And yet the position of the curve **is** determined by the levels of the child status and educational attainment factors. It can, therefore, be seen that this "shift mechanism" reflects the indirect effect which these characteristics of the wives have on their response, in terms of labour force attachments, to a change in their husband's income. In this sense it is operating in a way similar to that of an interaction effect. But, in addition, it would seem reasonable to suggest that the $I_{(U+L)/2}$ income value which locates the curve along the income scale can also be thought of as measure, or index, of the wives **expectation from employment**, defined to represent some vague notion of the expectation which the wife has in terms of both job and/or personal satisfaction and the family's economic standard of living. So that it can be said that the higher this income value the greater is the expectation from employment for that group of wives, regardless of the proportion of wives in that group who do go out to work.

This, then concludes the section on the model based on the generalised logistic curve. What now follows, in the final section, is a brief summary of the Study as a whole with some tentative conclusions on the merits and demerits of the methods examined.

VI. SUMMARY

What, then, has been found out in this Study? On page 9 in the Introduction it was stated that the purpose of the study was to consider ways in which the labour force participation rates of the population, cross-classified by a set of demographic, social or economic factors, can be related functionally to the different levels of those factors. The data used in the example were obtained from the 1961 Census for 54 groups of married women in urban Ontario who were categorised (1) by a child status factor which represents both the presence or absence of children in the family and the age of any children, (2) by the level of their educational attainment and (3) by the income of their husband. Ten models were examined: three which were termed simple additive models: three using the probit transformation: three based on the logit transformation and one which was a variant of the generalised logistic curves. In addition, Section III examined two analytical tools, the analysis of variance and dummy variance regression analysis, which are used in the examination of the first nine models. Section II briefly examined the nature of the labour force participation rate as a statistical variable and discussed some of the problems encountered in its analysis.

Table 15 now summarises the results with regard to the proportion of the variation in the participation rates explained by the 10 models which were examined. The first point to note is that in all models the explained variance was very high and give clear evidence of the significance of the effect which the factors examined have on the labour force participation rates. This will generally be the case in such studies since the subject is now well documented and the important factors have been identified. The main interest in any study will therefore be centred on the values attached to the coefficients and in the interpretation which can be placed on the results.

In each case where it applies (all except the generalised logistic curve) the proportion of the explained variation given in the table is that obtained without taking note of any interaction effect, either by incorporating variables into the model to specifically take care of known interactions (see page 19) or by partitioning the model. For the simple additive type of model without transformation and the probit and logit models, it is clear that including income as a continuous variable, either on a natural or logarithmic scale, causes a significant, even if small, reduction in the explained variation – particularly so when the logarithm of the income was used. Since the generalised logistic curve with income on a continuous (logarithmic) scale produced the best result in terms of this criterion it would seem that the choice of which model to use in practice is between:

(1) A simple additive type of model with no transformation of the independent variable and with all factor/levels represented by dummy variables, but with, in certain circumstances, the means to test for, and incorporate, "interaction" effects,

(2) A model similar to (1) but with the independent variable transformed into "probits" (see page 21),

(3) As for (2) but with a "logit" transformation,

(4) A model based on the generalised logistic curve with income on a continuous scale.

The first of the four models is undoubtedly the simplest to calculate and also the simplest to interpret (see page 19). The last is by far the most difficult to calculate but not necessarily, despite the complex form of the equation, the most difficult to interpret. It certainly has to be interpreted in a different and, perhaps, unusual way by reference, for example, to the change in the asymptotes caused by a change from one factor level to another; but given these constraints it is relatively easy to understand.

The models using the probit and logit transformation fall between these two extremes in terms of the ease with which they can be used, although of the two the logit would have to be judged the simpler since this transformation can be more easily "written in" as an option in standard computer programmes. However, providing that the number of observations is not unduly large, use can be made of the tables of the probit transformation which partially overcomes the problem. The interpretation of the estimated coefficients based on the probit and logit transformation, at least in terms of their immediate effect on the estimated participation rates, is virtually identical.

Turning now to the assumptions made in formulating the models it is not surprising to find that the ease with which the results can be interpreted is directly related to the simplicity of the assumptions behind the model. Certainly this is so in the case of the first model of the four now under consideration. Results based on the assumptions that the change in the labour force participation rate in response to a change in one of the factor levels is (a) independent of the level of the participation rate and (b) independent of the level of the other factor are very simple to interpret. But they may not always be very meaningful since the assumptions are not very realistic. (See page 16.) In particular, it was shown in Section III that the coefficients obtained from analysis of the simple additive model could easily result in estimates of labour force participation rates, for certain factor/level combinations, which lie outside the range of 0 and 100 per cent. However, it was indicated that if the number of factors being examined were few, perhaps no more than three or four, then this problem may not be so serious as to invalidate the approach completely. But this would, in part, depend on the extent of any significant interaction between the effects of the different factors. For this reason it was advised

TABLE 15. Proportion of Variation in Participation Rates Explained by Ten Models

Model	Explained variation
	per cent
Simple additive models:	
All dummy variables	86.2
With income on natural scale	82.6
With income on logarithmic scale	73.7
Probit transformation:	
All dummy variables	90.2 (90.4)[1]
With income on natural scale	86.4 (88.2)[1]
With income on logarithmic scale	75.9 (73.3)[1]
Logit transformation:	
All dummy variables	90.7 (99.3)[1]
With income on natural scale	86.9 (88.4)[1]
With income on logarithmic scale	76.0 (72.1)[1]
Generalised logistic curve:	
Income on logarithmic scale	93.9

[1] Figures in brackets denote the proportion of explained variation after the transformed estimates have been converted back to participation rates.

that prior analysis of the data by analysis of variance would indicate whether significant interaction effects were present which could then be allowed for.

It was seen that although the assumptions made in formulating the probit and logit models are totally different they happen in practice to yield essentially the same results. The practical advantage of either of these two transformations is that they constrain the estimates to lie between 0 and 100 per cent and in doing so allow for the fact that the change in participation rate arising from a change in the level of one of the factors is not independent of the participation rate. The difference between the two sets of assumptions behind these transformations has already been noted in Section IV on page 27, and need not be explored in detail here. The use of the logit requires no assumption to be made about the underlying causal mechanism, if one exists, between the wife's participation in the labour force and the socio-economic characteristics of the family. The probit transformation, on the other hand, is based on the assumption of a specific causal process.

The reason for modifying the logistic curve to provide the equation used in the last model was explained in some detail in Section V. It was argued that, when one of the variables was a continuous variable, it was reasonable to assume that the upper and lower limits of the participation rates for a particular population group, may not be 100 and zero per cent. In every other respect the assumptions incorporated in the generalised logistic curve are essentially the same as those for the logit model.

If, for reasons mentioned at the end of Section IV, the logit is to be preferred to the probit which of the three remaining models should be used in practice?

When all the variables are represented by dummies there is clearly no case for using the model based on the modified logistic curve. But when a continuous variable is included both the additive model and that using the logit transformation are liable to give poor results unless only a few factors are being examined. Similarly there is reason to believe that a simple additive model, with only dummy variables but with more than, say, three factors, will yield inferior estimates unless interaction effects are specifically allowed for. In this case the logit transformation may, on *a priori* reasoning, be expected to yield significantly better results, but the data used in this study did not allow this view to be tested.

Which model to use will, therefore, have to be determined by the nature of the data and the expected ability of each model to represent that data. But the decision need not be entirely one of judgement. The use of variance analysis will indicate whether interaction effects are significant and have to be allowed for. And alternative regression models can often be fitted in the same computer run so that a direct comparison can be made of their "explanatory power".

This Study has simply attempted to compare alternative lines of approach which have been found useful in practice and which may be of use to others.

APPENDIX

BIBLIOGRAPHY OF METHODOLOGICAL AND RELATED STUDIES

Aitchison, J. and Brown, J.A.C. *The Lognormal Distribution*, Cambridge, 1957.

Allingham, J.D. and Spencer, B. *Women Who Work: Part II*, Special Labour Force Studies No. 2 Series B, Dominion Bureau of Statistics, Ottawa, 1968.

Bain, A.D. *The Growth of Demand for New Commodities*, Journal of the Royal Statistical Society, Series A, Vol. 126, Pt. 2, 1963.

Berkson, J. *Application of the Logistic Function to Bio-assay*, American Statistical Association Journal, No. 39, 1944.

Berkson, J. *Relative Precision of Minimum Chi-square and Maximum Likelihood Estimates of Regression Coefficients*, Second Berkeley Symposium on Mathematical Statistics and Probability, 1951.

Berkson, J. *Why I Prefer Logits to Probits*, Biometrics, Vol. 7, 1951.

Berkson, J. *A Statistically Precise and Relatively Simple Method of Estimating the Bio-assay with Quantal Response, Based on the Logistic Function*, American Statistical Association Journal, No. 48, 1953.

Bowen, W.G. and Finnegan, T.A. *Educational Attainment and Labor Force Participation*, The American Economic Review, May, 1966.

Claringhold, P.J., Biggers, J.D. and Emmens, C.W. *The Angular Transformation in Quantal Analysis*, Biometrics, Vol. 9, No. 4, 1953.

Cochran, W.G. *The Analysis of Variance When Experimental Errors Follow the Poisson or Binomial Laws*, Annals of Mathematical Statistics, No. 11, 1940.

Cornfield, J. and Mantel, N. *Some New Aspects of the Application of Maximum Likelihood to the Calculation of the Dosage Response Curve*, American Statistical Association Journal, No. 45, 1950.

Croxton, F.E. and Crowden, D.J. *Applied General Statistics*, London, 1951.

Davies, O.L. (ed.) *The Design and Analysis of Industrial Experiments*, London, 1954.

Dyke, J.V. and Patterson, H.D. *Analysis of Factorial Arrangements When the Data are Proportions*, Biometrics, March, 1952.

Farrell, M.J. *The Demand for Motor-Cars in the United States*, Journal of the Royal Statistical Society, Series A, Vol. 117, Pt. 2, 1954.

Feldstein, M.S. *A Binary Variable Multiple Regression Method of Analysing Factors Affecting Perinatal Mortality and Other Outcomes of Pregnancy*, Journal of the Royal Statistical Society, Vol. 129, Pt. 1, 1966.

Finney, D.J. *Probit Analysis. A Statistical Treatment of the Sigmoid Response Curve*, Cambridge, 1952.

Finney, D.J. *The Principles of Biological Assay*, Journal of the Royal Statistical Society, Vol. 110, 1947.

Finney, D.J. *Statistical Method in Biological Assay*, London, 1964.

Fisher, R.A. *The Analysis of Variance with Various Binomial Transformations*, Biometrics, Vol. 10, March, 1954.

Fisher, R.A. and Yates, F. *Statistical Tables for Biological, Agricultural and Medical Research*, London, 1957.

Fisher, R.A. *The Design of Experiments*, London, 1951.

Fisher, R.A. *Statistical Methods for Research Workers*, New York, 1958.

Goldberger, A.S. *Econometric Theory*, New York, 1964.

Johnston, J. *Econometric Methods*, New York, 1963.

Kendall, M.G. *Natural Law in the Social Sciences*, Journal of the Royal Statistical Society, Series A, Vol. 124, Pt. 1, 1961.

Klein, L.R. *Text book of Econometrics*, 1953.

Malinvaud, E. *Statistical Methods of Econometrics*, Chicago, 1966.

Mansfield E. and Hensley, C. *The Logistic Process: Tables of the Stochastic Epidemic Curve and Applications*, Journal of the Royal Statistical Society, Series B, Vol. 22, No. 2, 1960.

Marguardt, D.W. *An Algorith for Least-Squares Estimation of Non-linear Parameters*, Journal of the Society of Industrial Applied Mathematics, Vol. 11, No. 2, June 1963.

Melichar, E. *Least Squares Analysis of Economic Survey Data*, Board of Governors of the Federal Reserve System, 13 pp. mimeo.

Morgan, J.N. and Sonquist, J.A. *Problems in the Analysis of Survey Data, and a Proposal*, American Statistical Society Journal, Vol. 58, 1963.

Oliver, F.R. *Methods of Estimating the Logistic Growth Function*, Journal of the Royal Statistical Society, Series C, Applied Statistics, Vol. 13, No. 2, 1964.

Ostry, S. *Provincial Differences in Labour Force Participation Rates*. One of a series of Labour Force Studies in the 1961 Census Monograph Programme, Ottawa, 1968.

Ostry, S. *The Female Worker in Canada*. One of a series of Labour Force Studies in the 1961 Census Monograph Programme, Ottawa, 1968.

Plackett, R.L. *Models in the Analysis of Variance*, Journal of the Royal Statistical Society, Series B, Vol. 22, No. 2, 1960.

Rosenbluth, G. *The Structure of Academic Salaries in Canada*, Canadian Association of University Teachers Bulletin, Vol. 16, No. 4, 1967.

Scott, J.T. *Factor Analysis and Regression*, Econometrica, Vol. 34, No. 3, July, 1966.

Sedransk, J. *Designing Some Multi-factor Analytical Studies*, American Statistical Association Journal, December, 1967.

Suits, D.B. *Use of Dummy Variables in Regression Equations*, American Statistical Association Journal, Vol. 52, December, 1957.

Tobin, J. *Estimation of Relationships of Limited Dependent Variables*, Econometrica, Vol. 26, No. 1, January, 1958.

Tobin J. *The Application of Multivariate Probit Analysis to Economic Survey Data*, Cowles Foundation Discussion, Paper No. 1, December, 1955.

Warner, S.L. *Multivariate Regression of Dummy Variates under Normality Assumptions*, American Statistical Association Journal, Vol. 58, December, 1963.