A Multivariate Approach to Social Inequality

M. Ribe

SCB
R & D Report
Statistics Sweden
Research - Methods - Development
U/STM - 33

INLEDNING

TILL

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Föregångare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

Efterföljare:

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report, U/STM 1987:33. A multivariate approach to social inequality / Martin Ribe. Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-RnD-USTM-1987-33

A Multivariate Approach to Social Inequality

M. Ribe

SCB
R & D Report
Statistics Sweden
Research - Methods - Development
U/STM - 33

Från trycket	Maj 1987
Producent	Statistiska centralbyrån, Enheten för statistiska metoder
Ansvarig utgivare	Staffan Wahlström
Förfrågningar	Martin Ribe, tel. 08 7834854

ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1987

A Multivariate Approach to Social Inequality

By M. Ribe

CONTENTS

SCB C 122 Arkiv. 2 1980 10 000 ex

Abstract

The object is to study the structure of social inequality with respect to welfare indicators, such as car ownership. Statistics are to be given which reflect to what extent such inequalities depend on contrasts between specific groups, such as manual workers and non-manual employees. An approach based on a logit model is presented. The approach involves a presentation of the results in the form of comparable index numbers for the inequality between groups. The method has been conveniently implemente in SAS, by PROC CATMOD, PROC MATRIX, and DATA-step applications.

1 INTRODUCTION

Statistics on welfare in Sweden are obtained by the Survey on Living Conditions, which is regularly carried out by Statistics Sweden. Several aspects of welfare are covered, such as consumption, employment, housing, health, and leisure. Reports from the Survey also give statistics on the social inequality in welfare between population groups; cf. Statistics Sweden (1981, 1987a, 1987b, 1987c), Nordic Council (1984).

The purpose here is to give a method for measuring welfare inequality as comparable index numbers. Consider a specific welfare indicator of yes/no type, such as "owning/not owning a car", "employed/unemployed", or "living/not living in a dwelling of acceptable standard". The purpose is to measure to what extent the social inequalities in this indicator depend on the respective background factors:

- Family situation
- Sex
- Socio-economic group
- Nationality (immigrants/Swedes)
- Geographical region

The factor "family situation" is a summary classification based on age, cohabitation, and age of youngest child.

We want to be able to make comparisons concerning the relative contributions of these factors to the inequality. In particular comparisons over time are of interest. For instance one may ask: "Has socio-economic group over the years become a more important or a less important factor for the inequality in the studied indicator?" We thus want to find some comparable index numbers which express this "degree of importance". This is the aim of the method presented here.

The index numbers are to be presented in some publications on Living Conditions from Statistics Sweden (1987a, b). There it is

to serve as a complement to extensive usual frequency tables.

Plan of the paper. The aims and specifications for the work are discussed in Sections 2-3. Section 4 gives the formal definition of the inequality index. Then in Sections 5-7 some motivations for choosing that definition are presented. Sections 8-11 take up some theoretical points, and Section 12 finally deals with the computational implemention in SAS.

Acknowledgements. The present work was carried out as a development task at Statistics Sweden, for the Survey on Living Conditions. The specifications of the aims essentially grew out by very useful talks with Dr. Joachim Vogel, who is Head of the Section for Living Conditions. The work also had great benefit of ideas and comments from Prof. Jan Hoem at the University of Stockholm, Prof. Bengt Rosén, Dr. Claes Cassel, Dr. Jan Hagberg, Dr. Harry Lütjohann and others at Statistics Sweden, and further Dr. Rolf Aaberge at Statistisk Sentralbyrå, Oslo, and Dr. Jan Selén at the Swedish Institute of Social Research, Stockholm.

2 PROPERTIES OF AN INEQUALITY INDEX

To find a suitable definition of the inequality index aimed at, let us state some properties which we would like the index to have. Concerning income inequality, measures of inequality have been extensively treated in the litterature; cf. Nygård and Sandström (1981), Sen (1976), Sen (1986). The present problem is somewhat different. First, we have to deal with yes/no-type indicators, like car ownership, rather than a continuous variable like income. We shall not give a rigorous set of mathematical axioms, but rather postulate some essential aims in more loose terms.

In this Section some general properties are listed, and in the next the properties are specified further. After that the inequality index is defined in Section 4.

(1) Basic aim. The inequality index shall quantify the extent to which the existing inequalities in some given welfare indicator, such as car ownership, depend on some given background factor, such as family situation or socio-economic group.

(2) Comparability over time. A primary purpose of the inequality index is to show the development over time in the impact of various background factors. The index numbers must thus be comparable over time and not subject to drift by irrelevant circumstances.

(3) Separation of background factors. The index numbers should reflect the impact of each background factor separately. An index pertaining to the impact of (e.g.) socio-economic group should not be affected by mere implications of the fact that different socio-economic groups have different age structures or different geographical distribution. It should reflect the pure effect of socio-economic group, "everything else alike". The demand for this property is partly a consequence of the demand for comparability over time, but also an end in itself.

(4) Comparability beween background factors. The index numbers should have such a normalized scale as to yield comparison of the importance of different background factors. For instance, one may like to compare the level and time-trend in inequality due to socio-economic group, vs. that due to family situation.

(5) Independence of size of background groups considered. An inequality index may sometimes pertain to the inequality between two groups of very different sizes, such as manual workers vs. entrepreneurs. The inequality index must fully recognize the inequality in such cases. It would be useless if an index for inequality between manual workers and entrepreneurs must always be small, just because the entrepreneurs are so few compared to the manual workers. The index should thus not be affected by the sizes of the background groups considered. This demand is partly a consequence of the demand for comparability between background factors (cf. previous paragraph).

(6) Comparability between indicators. The index numbers should allow comparison between different indicators, such as "car ownership" vs. "minimum housing standard". In particular this entails:

(7) Comparability between different percentage levels of indicators. Different indicators may have rather different percentage levels. For instance, there may be a couple of tenths of the Swedish population who do not have a car, but only very few per cent who do not live in a dwelling of acceptable standard. But irrespective of the percentage levels, there may be little or great inequality between groups. If one group has one per cent living in a substandard dwelling and another group has four per cent, there is a substantial inequality between the groups. Even though both percentages are low, their ratio takes a value which is far from 1. The index must recognize that kind of inequality. It must not automatically become small when the percentages are small.

Of course an index cannot tell everything. By nature it is a summary statistic, and to go further into detail one has to supplement it with other statistics, such as extensive frequency tables. The comparability properties are however an advantage of an index.

3 FURTHER SPECIFICATIONS

Let us now state a little more specifically what the inequality index is to look like.

Postulate 1. The inequality index will always pertain to the inequality between two groups. For instance, when considering socio-economic inequality, we may have one index for the inequality between manual workers and non-manual employees, one for that between manual workers and entrepreneurs, and so on. Likewise, concerning regional inequality, we may have one index for

the Stockholm conurbation versus other major cities, one for the Stockholm conurbation versus rural areas, and so on.

Postulate 2. The index shall work as if the two groups differed only with respect to the one background factor under concern. The index shall thus pertain purely to a difference in socio-economic group, or geographical region, or one other factor. Specifically the index shall not reflect differences in size or in the remaining background factors, but it shall work as if the two groups were alike in these respects.

Postulate 3. The inequality index shall range between -100 and 100. The value 0 means perfect equality, i.e., that the percentage (e.g., of car-owners) is equal in both groups. The values - 100 and 100 mean total inequality, i.e., that the percentage is 0 in the first group and 100 in the second, resp. 100 in the first and 0 in the second.

Postulate 4. If the two groups are interchanged, the index shall always change by merely reversing its sign. Likewise, if the indicator is replaced by its negation (e.g., if percentages of non car-owners are used, instead of those of car-owners), the index shall change by merely reversing its sign.

Comment: We thus restrict ourselves to consider inequality between groups within pairs of groups. This restriction considerably simplifies the comparability issues of the preceding Section; cf. particularly statements (4) and (5) there. Without the restriction it would be rather problematic to find an index for, e.g., socio-economic or regional inequality in general. The outcome may depend very much on the subdivisions used, especially on the number of groups distinguished between. The pairs of groups are much less ambiguous, and they allow comparison between index numbers for different pairs.

Also notice the connection between statement (3) in the preceding Section, and Postulate 2 here.

4 DEFINITION OF THE INEQUALITY INDEX

We are now ready to define the inequality index in terms of its actual computation. Consider a welfare indicator Y with the two possible values 0 and 1 (meaning, e.g., "not car-owner" resp "car-owner"). An inequality index I shall be computed pertaining to the inequality in Y between two groups X = 0 and 1, say. This is done in a three-step procedure:

Step 1. Let p denote the probability that Y = 1 for a given person. Assume a logit model to describe how p depends on the background factors. The variable X is defined in terms of one of those background factors. Estimate the parameters of the logit model from actual observed data, for each year in the time-series of interest.

Step 2. For all the background factors except the one used to define X, obtain statistics for a standard reference population. For instance, this could be the most recent year's population, or an average for a few of the most recent years. Plug those statistics into the logit model, with estimated parameter values from Step 1. For each year in the series, the model thus yields "predicted" values p_0 and p_1 for the probability that Y = 1 for a person in the standard population, given that X = 0 resp. 1.

Step 3. The values p_0 and p_1 found in Step 2 finally yield the inequality index,

$$I = \frac{p_0 - p_1}{(p_0 + p_1)\,(2 - p_0 - p_1)} \cdot 100.$$

Technical details on Steps 1 and 2 will be given later (Sections 8, 9 and 11).

The intuitive idea of the procedure may be explained as follows. By means of a multivariate method (a logit model) we obtain adjusted values p_0 and p_1 of the proportions of persons for which Y = 1 in the two groups. Those values are adjusted so as to allow comparison with respect to the defining distinction between the two groups, "everything else alike" (cf. Postulate 2). The index I measures a kind of relative difference between p_0 and p_1. We shall shortly recognize I as a regression coefficient (Section 6).

5 MOTIVATIONS IN GENERAL

Let us somewhat discuss the motivations for the definition just given. The use of a logit model hardly needs a very particular explanation. This is now a standard technique of multivariate analysis, almost like regression analysis is; cf., e.g., Breslow and Day (1980), and Koch and Edwards (1985). It is often used to achieve a "separation of background factors", as demanded in statement (3) of Section 2. However the relevance of logit models will be somewhat discussed later (Section 11).

By the use of a standard population in Step 2, we can get the model-predicted values p_0 and p_1. The estimated model-parameters alone do not give p_0 and p_1 themselves, but only the odds-ratio

$$\frac{p_1 \,/\, (1 - p_1)}{p_0 \,/\, (1 - p_0)} \quad .$$

This can be obtained as the exponentiated value of the coefficient for X in the model.

Would it then be possible to base an inequality index on the odds ratio (suitably transformed), and skip the standard population? It would actually not. Notice that when $p_0 = 0$,

than the odds ratio is always 0, irrespective of the value of p_1. This means that the odds ratio partly fails to reflect even unequivocal differences in inequality. An inequality index based on the odds ratio must violate Postulate 3 of Section 3, in failing to distinguish total inequality. So Step 2, with its reference to a standard population, is essential. Nevertheless it appears that the index I is rather robust with respect to the choice of the standard population.

Remark: The logit model actually can never give exactly $p_0 = 0$, for mathematical reasons. But still the reference just made to the case $p_0 = 0$ is apparently quite relevant. For $p_0 = 0$ is a natural limit case and an approximate possibility.

6 MOTIVATIONS FOR THE INDEX FORMULA

The formula for I in Step 3 needs some discussion. By formalizing Postulate 2 of Section 3 we can construct a simple model setting, for the inequality index. Consider a hypothetical population which is made up of two groups, with X = 0 and X = 1 respectively. In the hypothetical population those two groups have the same size and the same structure with respect to the remaining background factors (those not used to define X). In the two groups the proportion of those persons, for which Y = 1, is equal to p_0 and p_1 respectively. For a person picked at random in this population, the joint probability distribution of X and Y is as given in Table 1.

Table 1 A simple model setting

	X = 0	X = 1	Total
Y = 0	$\frac{1 - p_0}{2}$	$\frac{1 - p_1}{2}$	$1 - \frac{p_0 + p_1}{2}$
1	$p_0/2$	$p_1/2$	$\frac{p_0 + p_1}{2}$
Total	1/2	1/2	1

The consideration just made is really nothing but a formalization of a rather natural and simple idea. That idea was informally stated in Postulate 2 of Section 3. The model of Table 1 thus sets the stage for the inequality index, which should have a natural meaning in this model. And indeed the index I has such a meaning. Namely, let us consider simple linear regression, with Y as independent and X as dependent variable, and let $\beta_{X.y}$ be the regression coefficient. Then we have

$$I = - \beta_{X.y} \cdot 100.$$

This fact can also be stated more non-technically: In the hypothetical population, consider the difference between the percentage of persons with X = 1 among those with Y = 0, and the percentage of persons with X = 1 among those with Y = 1. Then I is equal to that difference.

And the use of $\beta_{X.y}$ is indeed a logical choice. Let us again look at statement (1) of Section 2. As stated there we should start out from the existing inequality in Y, the welfare indicator. We should then have the index tell to what extent this inequality means an inequality in terms of X, the background factor at study. As an expression of the dependency of X upon Y, the regression coefficient $\beta_{X.y}$ appears to be fit for this role.

Let us now list some easily verified mathematical properties of the index I.

Proposition. The inequality index I enjoys the following properties:

(i) $-100 \leq I \leq 100$.

(ii) $I = 0$ precisely when $p_0 = p_1$.

(iii) $I = -100$ precisely when $p_0 = 0$ and $p_1 = 1$; and $I = 100$ precisely when $p_0 = 1$ and $p_1 = 0$.

(iv) If X is replaced by 1 - X, then I changes to -I. If Y is replaced by 1 - Y, then I changes to -I.

(v) I is strictly monotonous in p_0 and p_1.

(vi) $|I| \geq |p_0 - p_1| \cdot 100$, and I and $p_0 - p_1$ are always of equal sign.

(vii) As p_0, $p_1 \to 0.5$, then $I/(p_0 - p_1) \to 100$.

(viii) As p_0, $p_1 \to 0$ so that $p_0/p_1 \to 0$, then $I \to -50$.

Comments: Properties (i)-(iv) here mean that I satisfies Postulates 3 and 4 of Section 3. Property (v) is naturally essential in view of the intuitive notion of inequality. As (vi) and (vii) show, there is further a natural relationship between I and the percentage difference $(p_0 - p_1) \cdot 100$. For percentages not too far from 50, the index I crudely approximates that difference. When the percentages get closer to 0 or 100, then I becomes more and more adjusted upwards compared to the percentage difference. The latter fact is finally reflected in (viii), which is also essential in view of statement (7) of Section 2.

Property (viii) can also be understood intuitively. Consider a situation where p_0 and p_1 are both small, but with p_0 very much smaller than p_1, so that p_0/p_1 is small. Then if one knows that Y = 1 for a person, then that person is likely to be in the group with X = 1, for most such persons are. For a person with Y = 1, one may thus predict an X-value near 1. On the other hand, knowing that Y = 0 for a person gives practically no clue whether X = 0 or 1 for the person; anyhow most persons have Y = 0. So the predicted X-value should be near 0.5. Thus the difference between the predicted X-values for Y = 1 and Y = 0 becomes nearly 0.5, and then $-\beta_{X.y} \cdot 100 = I$ becomes near to -50.

7 RULED-OUT ALTERNATIVES

As shown in the preceding Section, the reasons for using $\beta_{x.y}$ in the formula for the index should be logical enough. Still, one might ask what would happen if one tried to use some other measure of the association between X and Y, instead of $\beta_{x.y}$. Let us thus consider such measures for the simple model in Table 1.

A self-evident candidate is the usual correlation coefficient ρ_{xy}. One might thus consider a possible inequality index defined as

$$\rho_{xy} \cdot 100 = \frac{p_0 - p_1}{\sqrt{(p_0 + p_1)(2 - p_0 - p_1)}} \cdot 100.$$

Actually this formula is very similar to that for I, differing only by the square root in the denominator.

Yet another alternative might be the squared correlation coefficient ρ_{xy}^2. This might seem attractive in view of the connection with variance decomposition. As is well known one can write

$$\sigma_y^2 = \rho_{xy}^2 \sigma_y^2 + (1 - \rho_{xy}^2) \sigma_y^2 = \sigma_b^2 + \sigma_w^2,$$

where σ_b^2 and σ_w^2 are the variance in Y between resp. within the two groups X = 0 and X = 1. Thus ρ_{xy}^2 is the "between-groups proportion" of the total variance in Y.

However, ρ_{xy} and ρ_{xy}^2 have the property that they are always small when p_0 and p_1 are both small (or both close to 1). This means an unacceptable violation of statement (7) in Section 2. The correlation coefficient is too insensitive for inequality at extreme percentage levels.

To make the index sensitive to inequality at low percentage levels, one might think of using the simple formula

$$\frac{p_0 - p_1}{p_0 + p_1} \cdot 100$$

for an index. In terms of the model in Table 1 this can actually be interpreted as a coefficient of variation of a conditional probability:

$$CV(P(Y = 1 \mid X)) \cdot 100.$$

That interpretation appears somewhat artificial, and the alternative is indeed unacceptable for at least two reasons. Like the odds ratio it fails to distinguish total inequality; cf. Section 5. It also fails on the second part of Postulate 4. The modulus of such an index could actually take a most different value if Y was replaced by 1 - Y. This cannot be allowed, since the modulus of the index should express the strength of inequality, for which the choice between Y and 1 - Y should be a nonconsequential trivial matter of notation.

8 SAMPLING CONSIDERATIONS

As mentioned in Section 1 the source data in the present application are obtained from the Swedish Survey on Living Conditions. This is a sample survey, and thus the results will be subject to sampling errors. Chiefly due to these sampling errors, the estimated parameters in the logit model are influenced by some uncertainty. There would, however, be some uncertainty even if the data comprized the whole population and not just a sample, since data in practice deviate more or less from a model. The uncertainty is quantified by the model as estimated variances and covariances of the parameter estimates. Using linear approximations and an assumption of normal distribution, one can straightforwardly compute a confidence interval for the index.

The random uncertainty of the index is particularly noticeable when p_0 and p_1 are both near 0 or 1. It may be desirable to stabilize the index by smoothing, e.g., by three-year moving averages.

The Survey on Living Conditions uses a stratified sample with varying sampling probabilities. This means that one has to consider whether or not to weight the observations, with their inverse sampling probabilities as weights, in the logit analysis. The use resp. nonuse of such weighting correspond to what is known as "design-based" resp. "model-based" inference; cf. Särndal (1985). The parameter estimates become unbiased only with weighting. On the other hand the estimation of variances and covariances in the usual model works only without weighting. Besides it is likely that the variances of the parameter estimates often become somewhat smaller without weighting; and trouble with outlying weights is eliminated.

Here one has to take a pragmatic view. In surveys with a very drastic stratification, weighting may be the only choice, to avoid a totally disturbing bias. However, the Survey on Living Conditions is not of that kind, and experience confirms that weighting and nonweighting mostly tend to give very similar results in analyses. It is thus feasible not to weight, so in view of the advantages that alternative was chosen in the present application.

9 TECHNICAL POINTS

Let us for clarity state the relevant formulas for Steps 1 and 2 of Section 4.

It is convenient to represent each of the background factors in terms of one or more dummy variables, i.e., variables with 0 and 1 as the only possible values. For each background factor the population is subdivided into two or more groups. A dummy vari-

able is introduced for each such group except one (for each factor), the "reference group". For instance the factor "socio-economic-group" may be described by the dummy variables $x_1,\ldots,x_5$, so that

$(x_1,\ldots,x_5) =$

(0,0,0,0,0) denotes manual workers (reference group)
(1,0,0,0,0) denotes assistant nonmanual employees
(0,1,0,0,0) denotes intermediate/higher nonmanual employees
(0,0,1,0,0) denotes entrepreneurs
(0,0,0,1,0) denotes farmers
(0,0,0,0,1) denotes others.

Let $x_1,\ldots,x_k$ be all the dummy variables for all the background factors. Given the values of $x_1,\ldots,x_k$ for a person, the logit model predicts a probability $p = P(Y=1)$ about the welfare indicator Y for that person. It does so by the formula

$$p = \left(1 + \exp\left(-b_0 - \Sigma_{i=1}^{k} b_i x_i\right)\right)^{-1}.$$

Here $b_0,\ldots,b_k$ are the model parameters which are estimated in Step 1 and used in Step 2 (cf. Section 4). The estimation is done by the Maximum Likelihood method: Let $y_{(0)},\ldots,y_{(n)}$ be the observed values of Y for the individuals 1,...,n in the sample. Let $p_{(0)},\ldots,p_{(n)}$ be the corresponding predicted values of p, considered as functions of the same parameters $b_0,\ldots,b_k$. In the estimation the values of the parameters are then so determined that the product

$$\Pi_{j=1}^{n} \left(1-p_{(j)}\right)^{1-y_{(j)}} p_{(j)}^{\,y_{(j)}}$$

becomes as large as possible.

Suppose that $\bar{x}_1^s,\ldots,\bar{x}_k^s$ are the mean values of the dummy variables in a standard reference population. Thus $\bar{x}_i^s$ is simply the proportion of persons such that $x_i = 1$ in the reference population. These statistics are used, together with the estimated values of $b_0,\ldots,b_k$, in Step 2 (cf. Section 4). This gives predicted values of p for various groups. For instance, if $x_1,\ldots,x_5$ describe socio-economic group in the way just stated, then the predicted value of p for "entrepreneurs" is

$$(1 + \exp(- b_0 - b_3 - \Sigma_{i=6}^{k} b_i \bar{x}_i^s))^{-1}.$$

10 WHY NOT LEAST SQUARES

Least Squares estimation is an alternative to Maximum Likelihood estimation in a logit model, using a somewhat different criterion to determine the parameter values $b_0,\ldots,b_k$.

The two estimation methods can be described as follows. Let

$$\underset{\sim}{x} = (x_1,\ldots,x_k)$$

denote the vector with components $x_1,\ldots,x_k$. For each posible value of $\underset{\sim}{x}$ let $n(\underset{\sim}{x})$ be the number of observations (persons in the sample) in that "cell" $\underset{\sim}{x}$. Also suppose that $\bar{y}(\underset{\sim}{x})$ is the proportion of persons for which Y = 1 in that cell, and that $p(\bar{x})$ is the corresponding value of P(Y=1) predicted by the model. Then the Maximum Likelihood method minimizes

$$-\Sigma_{\underset{\sim}{x}}\, n(\underset{\sim}{x})((1-\bar{y}(\underset{\sim}{x}))\, \log(1-p(\underset{\sim}{x})) + \bar{y}(\underset{\sim}{x})\, \log\, p(\underset{\sim}{x}))\ ,$$

while the Least Squares method minimizes

$$\Sigma_{\underset{\sim}{x}}\, n(\underset{\sim}{x})(\log\, \bar{y}(\underset{\sim}{x}) - \log\, p(\underset{\sim}{x}) - \log(1- \bar{y}(\underset{\sim}{x})) + \log(1-p(\underset{\sim}{x})))^2,$$

(where 0 log 0 = 0). In both formulas the summation extends over all cells $\underline{x}$.

There is a similarity between these two formulas, but sometimes the distinction may not be inessential. For cells where $\bar{y}(\underline{x})$ and $p(\underline{x})$ are close to 0 or 1, the latter sum is seen to have an exaggerated sensitivity to small changes in $\bar{y}(\underline{x})$. So the Least Squares method is less robust in this respect. Now, in the present application $\bar{y}(\underline{x})$ may sometimes vary heavily between cells, and a decent precision is required in most (if not all) parameters. Thus Maximum Likelihood is preferred.

11 SOME BACKGROUND RELATED TO LOGIT MODELS

Though the use of logit models is now an established technique of multivariate analysis, it may be in order to recapitulate some background for the relevance of such models. These aspects are well known but perhaps not so often explicitly discussed in the literature.

The most clear motivation for logit models appears in connection with evolution processes; see Montroll (1987). Consider for instance the process of introducing a new technical facility, such as the dishwasher, in a population. Let $Q(t)$ denote the proportion of the population which has a dishwasher at time t. Typically $Q(t)$ will increase over time as more and more people get themselves a dishwasher. And the more prevalent dishwashers have become, the more likely it is that anyone who is still without a dishwasher will acquire one soon.

Indeed, as a simple idealized model, it may be assumed that for a person, who is still without a dishwasher at time t, the probability of getting a dishwasher before time $t + dt$ is proportional to $Q(t)dt$, for dt small. Since the proportion of persons without a dishwasher at time t is $1 - Q(t)$, it is then seen that $Q(t)$ may be expected to follow a solution to the differential equation

$$dQ(t)/dt = cQ(t)(1 - Q(t)),$$

where c is a constant. (Random fluctuations in Q(t) due to the finiteness of the population are disregarded here.)

The general solution (such that $0 < Q(t) < 1$) to this differential equation is given by

$$Q(t) = (1 + e^{-c(t-t^0)})^{-1},$$

where t^0 is a constant. This function follows an S-shaped curve. For early times t the curve lies steadily on a low level, only slowly increasing. Later it gradually increases faster and faster until it reaches the level $Q = 0.5$, and after that the increase gradually slows down as the curve approaches the level $Q = 1$.

Now what is interesting in our context is that the evolution process may be differently lagged in different population groups. The development may come earlier in some groups and later in others. To obtain a model for this phenomenon one may let t^0 depend on various background variables. Assuming a linear model for this dependence, one thus gets the model

$$Q(t) = (1 + \exp(-ct - b_0 - \textstyle\sum_{i=1}^{k} b_i x_i))^{-1}.$$

As before the x_i are background variables expressing socio-economic group, geographical region, etc.

This model is of course idealized in various respects. And in the present application we are not interested in using it to describe evolution processes. What is essential to us is that the reasoning provides a basis for the study of inequality between groups with respect to welfare indicators. And indeed the considerations may be supposed to have a wider applicability than evolution processes, even though the model is most easily motivated for such settings. Thus the variable t need not be thought of as time in a literal sense, but rather as some more general kind of "promoting" variable for the welfare indicator in question.

In our application we certainly do not want to model time by means of that variable t. We actually estimate the logit model for each year separately, as explained in Section 4. To us the variable t has no interest in itself, and thus the term ct in the last formula is absorbed into the constant b_0. The model stated in Section 9 is thus obtained.

This reasoning is of course rather loose and may not be relied too heavily upon. It is then also important that the logit model appears to be fairly robust, working sensibly in different situations.

12 IMPLEMENTATION IN SAS

The software system SAS is quite convenient for the implementation of the computation of the inequality index. The main computational step is the parameter estimation in the logit model. Here the procedure CATMOD is used. The parametrization in CATMOD differs a little from the formulas in Section 9. Letting $b_0', \ldots, b_k'$ denote the parameters in CATMOD, we get the logit model in the form

$$1 - p = (1 + \exp(-b_0' + \Sigma_{i=1}^{k} b_i' (2x_i - 1))^{-1},$$

if only dummy variables (0-1-variables) are used. See further the discussion about design matrix, response function, etc., in the Chapter on CATMOD in the Manual; SAS (1985b).

The parameter estimates are retrieved in a SAS data set, requested by OUTEST = ... in a RESPONSE statement under PROC CATMOD. This data set contains both the parameter estimates and their estimated variance-covariance matrix. The data set is then taken as input to processing in procedure MATRIX and/ or DATA-steps. There the inequality index itself, together with its confidence interval, is computed.

The results are feasibly tabulated by procedure PRINT; cf. SAS (1985a, c). This is a quick and simple way to obtain easily readable tables. An example is shown in Table 2.

In the present application all of the computations were carried out on an IBM/MVS mainframe. Table 2 was printed from a PC/AT, after DOWNLOADing of the output data from the mainframe. In future applications more of the work may be done on the PC/AT. Running CATMOD however takes considerable CPU-time and should still be done on the mainframe. But the OUTEST datasets from CATMOD could be DOWNLOADed to the PC/AT, where the remaining computations, tabulations etc. could be taken care of. This kind of distributed processing may enhance the flexibility for modifications regarding choice of variables, smoothing over time, modes of output and editing, etc.

Table 2

Inequality Indexes (Example) 16
14:53 Thursday, March 19, 1987
- Car Ownership -

CASE=A) 16-24 YRS OLD VS. RETIRED

YEAR	P0	P1	P0 - P1	INDEX I	95 % CONFIDENCE MARGIN +/-	3-YR MO-VING AVE OF I	95 % CONFIDENCE MARGIN +/-
75	0.801102	0.322742	0.478360	48.6	3.5	.	.
76	0.847099	0.320868	0.526231	54.2	3.4	50.2	2.1
77	0.776662	0.299543	0.477119	48.0	3.6	46.5	2.3
78	0.751334	0.383680	0.367654	37.4	4.5	44.3	2.3
79	0.792738	0.324140	0.468598	47.5	3.9	41.7	2.6
80	0.774976	0.383129	0.391847	40.2	4.9	43.0	2.6
81	0.746370	0.336661	0.409709	41.3	4.7	37.4	2.8
82	0.729186	0.428514	0.300672	30.8	4.9	35.8	2.9
83	0.756874	0.414631	0.342243	35.3	5.1	29.4	3.0
84	0.722890	0.514467	0.208423	22.1	5.3	26.8	3.1
85	0.713438	0.492310	0.221127	23.1	5.7	.	.

CASE=B) ENTREPRENEURS VS. MANUAL WORKERS

YEAR	P0	P1	P0 - P1	INDEX I	95 % CONFIDENCE MARGIN +/-	3-YR MO-VING AVE OF I	95 % CONFIDENCE MARGIN +/-
75	0.947004	0.771309	0.175695	36.3	2.6	.	.
76	0.931028	0.814636	0.116392	26.2	3.1	28.9	1.7
77	0.912654	0.790618	0.122036	24.1	3.3	28.4	1.8
78	0.958574	0.821873	0.136701	35.0	3.1	26.3	2.1
79	0.916572	0.829389	0.087183	19.7	4.4	30.9	2.3
80	0.973227	0.851420	0.121807	38.1	4.1	29.4	2.5
81	0.952963	0.839315	0.113649	30.5	4.3	34.3	2.4
82	0.965496	0.852209	0.113286	34.2	3.9	34.3	2.4
83	0.978155	0.872213	0.105941	38.3	4.3	35.0	2.5
84	0.965424	0.863712	0.101712	32.5	4.6	30.5	3.0
85	0.950682	0.889575	0.061106	20.8	6.2	.	.

CASE=C) WOMEN VS. MEN

YEAR	P0	P1	P0 - P1	INDEX I	95 % CONFIDENCE MARGIN +/-	3-YR MO-VING AVE OF I	95 % CONFIDENCE MARGIN +/-
75	0.812163	0.841280	-.029117	-5.0	1.5	.	.
76	0.830157	0.861772	-.031616	-6.0	1.5	-6.0	0.9
77	0.818986	0.858178	-.039192	-7.1	1.5	-6.7	0.9
78	0.844866	0.878492	-.033626	-7.0	1.7	-7.1	1.0
79	0.836761	0.873404	-.036642	-7.3	1.7	-9.4	1.1
80	0.838608	0.901510	-.062902	-13.8	2.1	-11.3	1.1
81	0.837017	0.896661	-.059644	-12.8	2.0	-13.2	1.2
82	0.844899	0.902699	-.057800	-13.0	1.9	-13.0	1.2
83	0.855635	0.910502	-.054867	-13.2	2.1	-13.6	1.2
84	0.858115	0.916816	-.058701	-14.6	2.0	-15.4	1.2
85	0.864534	0.931785	-.067251	-18.3	2.2	.	.

12 REFERENCES

N.E. Breslow and N.E. Day (1980), Statistical Methods for Cancer Research, Vol. 1, International Agency for Research on Cancer, Lyon.

G.G. Koch and S. Edwards (1985), Logistic regression, Encyclopedia of Statistical Sciences, (ed. by S. Kotz and N.L. Johnson), John Wiley, New York, Vol. 5, pp. 128-133.

E.W. Montroll (1987), On the dynamics of some sociotechnical systems, Bulletin of the American Mathematical Society, 16, 1-46.

Nordic Council (1984), Level of Living and Inequality in the Nordic Countries, Nordic Council/Nordic Statistical Secretariat Stockholm.

F. Nygård and A. Sandström (1981), Measuring Income Inequality, Stockholm University.

C.-E. Särndal (1985), How survey methodologists communicate, Journal of Official Statistics, 1, 49-64.

SAS (1985a), User's Guide: Basics, Version 5 Edition, SAS Institute, Cary.

SAS (1985b), User's Guide: Statistics, Version 5 Edition, SAS Institute, Cary.

SAS (1985c), Procedures Guide for Personal Computers, Version 6 Edition, SAS Institute, Cary.

A. Sen (1976), Poverty: an ordinal approach to measurement, Econometrica, 44, 219-231.

P.K. Sen (1986), The Gini coefficient and inequality indexes: some reconcilations, Journal of the American Statistical Association, 81, 1050-1057.

Statistics Sweden (1981), Social Report on Inequality in Sweden, Living Conditions Report No. 27, SCB/Liber.

Statistics Sweden (1987a), Det Svenska Klassamhället, Levnadsförhållanden 1975-85, Rapport 50, SCB (in Swedish)

Statistics Sweden (1987b), Ojämlikheten i Sverige, Levnadsförhållanden 1975-85, Rapport 51, SCB (in Swedish; an English edition is forthcoming).

Statistics Sweden (1987c), Perspektiv på Välfärden 1987, Serie Levnadsförhållanden, SCB (in Swedish with English notes)

Tidigare nummer av Promemorior från U/STM:

NR

1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)

4 Handledning i AID-analys. Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)

9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)

10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)

11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)

12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)

13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)

14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)

15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)

16 Variance estimators of the Gini coefficient - simple random sampling. Arne Sandström, Jan Wretman och Bertil Waldén (Memo, Februari 1985)

17 Variance estimators of the Gini coefficient - probability sampling. Arne Sandström, Jan Wretman och Bertil Waldén (1985-07-05)

18 Reconciling tables and margins using least-squares. Harry Lütjohann (1985-08-01)

19 Ersättningens och uppgiftslämnarbördans betydelse för kvaliteten i undersökningarna om hushållens utgifter. Håkan L. Lindström (1985-11-29)

20 A general view of estimation for two phases of selection. Carl-Erik Särndal och Bengt Swensson (1985-12-05)

21 On the use of automated coding at Statistics Sweden. Lars Lyberg (1986-01-16)

22 Quality Control of Coding Operations at Statistics Sweden. Lars Lyberg (1986-03-20)

23 A General View of Nonresponse Bias in Some Sample Surveys of the Swedish Population. Håkan L Lindström (1986-05-16)

24 Nonresponse rates in 1970 - 1985 in surveys of Individuals and Households. Håkan L. Lindström och Pat Dean (1986-06-06)

25 Two Evaluation Studies of Small Area Estimation Methods: The Case of Estimating Population Characteristics in Swedish Municipalities for the Intercensal Period. Sixten Lundström (1986-10-14)

26 A Survey Practitioner's Notion of Nonresponse. Richard Platek (1986-10-20)

27 Factors to be Considered in Developing a Reinterview Program and Interviewer Debriefings at SCB. Dawn D. Nelson (1986-10-24)

28 Utredning kring bortfallet i samband med den s k Metropolit-debatten. Sixten Lundström (1986-11-10)

29 Om HINK-undersökningens design- och allokeringsproblematik. Bengt Rosén (1986-10-06)

30 Om HINK-undersökningens estimationsförfarande. Bengt Rosén (1987-01-28)

31 Program för minskat bortfall i SCBs individ- och hushållsundersökningar - förslag från aktionsgruppen för uppgiftslämnarfrågor. Lars Lundgren (1987-02-20)

32 Practical estimators of a population total which reduce the impact of large observations. Jörgen Dalén (87-03-03)

Kvarvarande exemplar av ovanstående promemorior kan rekvireras från Elseliv Lindfors, U/STM, SCB, 115 81 Stockholm, eller per telefon 08 7834178

GRUPP 03
LUNDIN STURE
A BF

STOCKHOLM