CHAPTER 5

Statistical Estimation of Choice Probability Functions

5.1. Introduction

We shall now consider the procedures which can be used to estimate the unknown parameters in the probability models discussed in chapter. 4. Two basic techniques are available: (1) least squares regression analysis and (2) maximum likelihood methods. The form and applicability of these techniques depend on the structure of the probability functions whose unknown parameters are to be estimated.

The data available for the calibration will typically be a sample of I individuals, whom we can index $i = 1, \ldots, I$. For individual i one observes a vector s^i of individual characteristics, a list of available alternatives which we can index $j = 1, \ldots, J$, and a corresponding list of vectors x^{ji} of observed attributes of the alternatives. The observed choice of the individual can conveniently be denoted by defining

$$f_{ji} = \begin{cases} 1 & \text{if } j \text{ is chosen,} \\ 0 & \text{if } j \text{ is not chosen.} \end{cases}$$

In principle, the number of alternatives J_i available to individual i may vary with the individual, but in many applications it will be constant across individuals.

5.2. Estimation of the binary-choice model

We shall first discuss estimation of the binary-choice model, since this case has been thoroughly explored in the literature. We confine our attention to probability functions that are transformations into the zero–

one interval of linear-in-parameters functions, as this is the only case of real practical interest.

From eqs. (4.9) and (4.10), we have for individual i,

$$P_{1i} = G(V_{1i} - V_{2i}), \tag{5.1}$$

and

$$V_{ji} = V(x^{ji}, s^i) = \sum_{k=1}^{V} \beta_k Z^k(x^{ji}, s^i) = \beta' Z(x^{ji}, s^i), \tag{5.2}$$

where G is a cumulative distribution function mapping points on the real line into the unit interval, $\beta' = (\beta_1, \ldots, \beta_K)$ is a vector of unknown parameters, $z_k^{ji} = Z^k(x^{ji}, s^i)$ is a numerical function of x^{ji} and s^i, and $z^{ji\prime} = Z(x^{ji}, s^i)' = (z_1^{ji}, \ldots, z_K^{ji})$ is a vector of these numerical functions. The sample we have drawn then has the property that f_{1i} is an observation from a binomial distribution with probability

$$P_{1i} = G(\beta' z^{1i} - \beta' z^{2i}). \tag{5.3}$$

For the remainder of the discussion of estimation of the binary-choice model, we use the notation $z^i = z^{1i} - z^{2i}$, so that $P_{1i} = G(\beta' z^i)$. We assume that the available data also have the property that these observations are statistically independent across individuals. The statistical question is then how the parameters β can be estimated in a way which yields results that are "satisfactory" in the sense that the estimates lie close to the true parameter values. This question has been discussed in considerable detail in Maxwell (1961), McFadden (1973a), and Cox (1970); the last two references contain extensive bibliographies.

5.3. *The linear probability model*

The procedure for estimating the linear probability model is the simplest from a computational point of view. From eq. (4.11), this model can be written:

$$P_{li} = \begin{cases} 0 & \text{if } \beta' z^i < 0, & (5.4\text{a}) \\ \beta' z^i & \text{if } 0 \leqq \beta' z^i < 1, & (5.4\text{b}) \\ 1 & \text{if } 1 < \beta' z^i. & (5.4\text{c}) \end{cases}$$

From fig. 4.4, one sees that (5.4b) corresponds to the interval in the response curve in which the probability does not take on the extreme values zero or one. The conventional estimation procedure is to tacitly assume that all the data give responses in this interval, and apply ordinary least squares to the regression equation

$$f_{1i} = \beta' z^i + \varepsilon_i, \tag{5.5}$$

with $Ef_{1i} = P_{1i}$ implying $E\varepsilon_i = 0$. The estimates are then

$$\hat{\beta} = \left[\sum_{i=1}^{I} z^i z^{i'}\right]^{-1} \left[\sum_{i=1}^{I} z^i f_{1i}\right]. \tag{5.6}$$

It can be shown that these estimates are *unbiased*: the averages of the estimates calculated from repeated samples equal the true parameters. If the explanatory variables satisfy regularity conditions on their variability, the estimates are *consistent*: they converge to the true parameters with probability one as the sample size grows to infinity. One expects these regularity conditions to be met in cross-section data. Thus the estimates in eq. (5.6) are quite satisfactory when it is valid to assume that the response curve is linear and that all the observations lie in the range where the probabilities are between zero and one. Furthermore, these estimates can be obtained at low cost using conventional regression programs.

Two comments should be made about this estimation procedure. First, the error term ε_i in eq. (5.5) is heteroskedastic, with $E\varepsilon_i^2 = P_{1i}(1 - P_{1i})$. This suggests [see Goldberger (1964)] that it would be more efficient to obtain the ordinary least squares estimates in eq. (5.6), calculate consistent estimates $\hat{P}_{1i}$ of the choice probabilities, and then carry out a weighted least squares regression,

$$[\hat{P}_{1i}\hat{P}_{2i}]^{-1/2} f_{1i} = [\hat{P}_{1i}\hat{P}_{2i}]^{-1/2} z^{i'} \beta, \tag{5.7}$$

to obtain final estimates of β. One can show that in asymptotically large samples this procedure yields the best (i.e., most efficient) estimates possible. However, in small samples, this second procedure tends to place excessive weight on extreme observations, leading to more variable estimates than the one-pass ordinary least squares calculation. The two-pass procedure also aggravates the sensitivity to specification error of the linear probability model. For these reasons we do not recommend the use of the weighting procedure.

Our second comment on ordinary least squares estimates of the linear probability model concerns the treatment of observations for which

$\beta'(z^{1i} - z^{2i})$ lies outside the zero–one interval. Even when the specification of the model is valid, this phenomenon is quite likely to occur as a result of normal sampling effects. Fig. 5.1 illustrates a case with a single explanatory variable in which a series of observations that might well be obtained from the true model leads to a predicted "probability" which lies outside the zero–one interval for values of the explanatory variable at the extremes of the observed range. This outcome is inconsistent with the a priori restriction of values of the right-hand side of eq. (5.4) to the zero–one interval. We might ignore this inconsistency and continue to use the estimates obtained in eq. (5.6), setting the predicted P_i to the extreme value zero or one when $\beta' z^i$ is less than zero or greater than one, respectively. However, this may have the result that for some values of

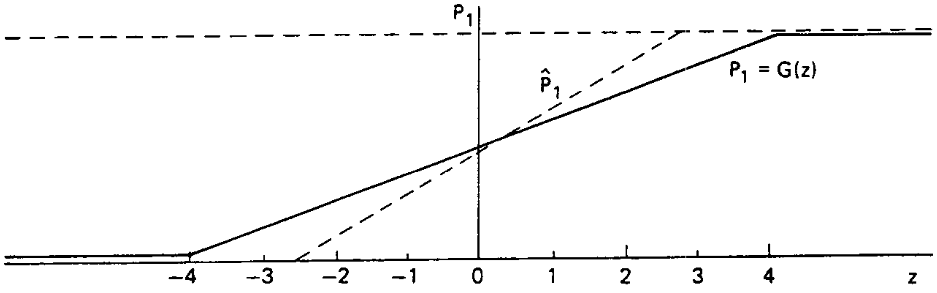

Fig. 5.1.

True model

$$P_{1i} = \begin{cases} 1 & \text{for } z^i \geqq 4 \\ 1/2 + z^i/8 & \text{for } |z^i| < 4 \\ 0 & \text{for } z^i \leqq -4 \end{cases}$$

Sample

i	1	2	3	4	5	6	7	8	9	10	11	12
z^i	−3	−2	−1	−1	−1	0	0	1	1	1	2	3
f_{1i}	0	0	1	0	0	1	0	1	1	0	1	1
P_{1i}	1/8	1/4	3/8	3/8	3/8	1/2	1/2	5/8	5/8	5/8	3/4	7/8

Unconstrained ordinary least squares estimate

$$\hat{P}_{1i} = \begin{cases} 1 & \text{for } z^i \geqq 2\tfrac{2}{3} \\ 1/2 + 3z^i/16 & \text{for } |z^i| < 2\tfrac{2}{3} \\ 0 & \text{for } z^i \leqq -2\tfrac{2}{3} \end{cases}$$

The linear expression $\frac{1}{2} + 3z^i/16$ exceeds one at the data point $z^{12} = 3$.

the explanatory variables we are required to predict that an alternative will be chosen with probability one, when in fact we observe that it is sometimes not chosen. Thus, even though the estimates of β are unbiased, the predicted probabilities for extreme observed values of the explanatory variables may be rather badly biased.

As an alternative to this procedure, we might estimate β in eq. (5.5) by least squares, subject to the inequality constraints

$$0 \leqq \beta' z^i \leqq 1. \tag{5.8}$$

This estimation procedure can be formulated as a quadratic programming problem and solved by a finite computational routine such as the Dantzig–Cottle algorithm. The resulting estimates of β will again be consistent provided the specification is correct, and in small samples they will tend to be distributed more tightly about the true parameter values even though they are no longer unbiased (see fig. 5.2). These properties suggest that the estimates obtained when the inequality constraints are imposed are preferable to those obtained using the simple ordinary least squares estimation method. On the other hand, this inequality-constrained estimation procedure is more costly, and as we shall see below, it is also more sensitive to specification error and does not eliminate the bias in extreme probabilities. Hence in demand analysis of transportation survey data where specification errors are likely, it seems preferable to reject the inequality-constrained least squares estimation procedure in favor of the simple ordinary least squares method.

We now turn to a discussion of the effects of specification errors on least squares estimation of the linear probability model. The first possibility we consider is that the linear response curve of the form indicated in fig. 5.1 is valid, but that some observations in our sample are drawn

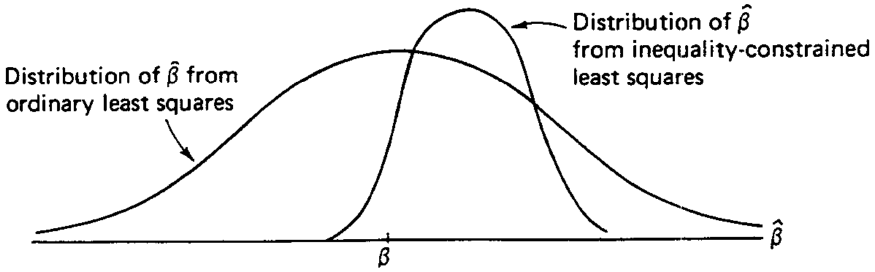

Fig. 5.2. Comparison of distribution of parameter estimates.

from the ranges of the explanatory variables where the probabilities take the extreme values. Fig. 5.3 illustrates such a sample and the corresponding fitted linear probability function obtained using the ordinary least squares estimates of eq. (5.6). In this case, the magnitudes of the parameter estimates are substantially biased below their true values. As a result, the linear probability model will tend to underestimate the elasticity of response with respect to explanatory variables for individuals in the intermediate probability range, and overestimate this elasticity in the extreme probability range. Thus the linear probability model would lead to forecasts of aggregate demand elasticities which are larger in magnitude than the true values in a transportation survey in which a large proportion of the observations correspond to "clear-cut" best choices with corresponding extreme probabilities.

A specification error can also occur in the case in which the true response curve is a smooth ogive, as in fig. 4.1. Fig. 5.4 gives an example

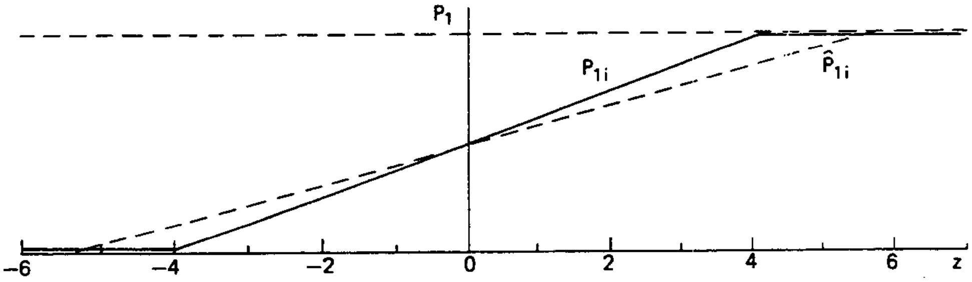

Fig. 5.3.

True model

$$P_{1i} = \begin{cases} 1 & \text{for} \quad z^i \geqq 4 \\ 1/2 + z^i/8 & \text{for} \quad |z^i| < 4 \\ 0 & \text{for} \quad z^i \leqq -4 \end{cases}$$

Sample

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
z^i	−7	−5	−3	−2	−1	−1	−1	0	0	1	1	1	2	3	5	7
f_{1i}	0	0	0	0	1	0	0	1	0	1	1	0	1	1	1	1

Ordinary least squares estimate

$$\hat{P}_{1i} = \begin{cases} 1 & \text{for} \quad z^i \geqq 5.28 \\ 1/2 + 9z^i/95 & \text{for} \quad |z^i| < 5.28 \\ 0 & \text{for} \quad z^i \leqq 5.28 \end{cases}$$

in which the true response function obeys the logistic law. As in the previous case, the fitted linear probability function underestimates the effect of a change in the explanatory variable in the intermediate probability range, overestimates the effect when the probabilities are near the extreme values, and predicts no effect when the extreme values are reached and the linear probability function is truncated. Using the linear probability model, the predicted increase in demand for alternative 1

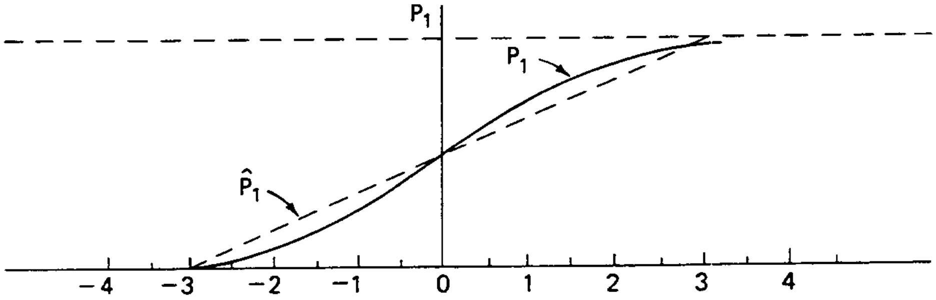

Fig. 5.4.

True model

$$P_{1i} = 1/(1 + e^{-z})$$

Sample

i	1-100	101-200	201-300	301-400	401-500	501-600	601-700
z^i	−3	−2	−1	0	1	2	3
no. times 1 chosen	5	12	27	50	73	88	95
P_{1i}	0.047	0.119	0.267	0.5	0.731	0.881	0.953

Ordinary least squares estimation of linear probability approximation to true model.

$$\hat{P}_i = \begin{cases} 1 & \text{for } z^i \geqq 3 \\ 1/2 + 0.167\, z^i & \text{for } |z^i| < 3 \\ 0 & \text{for } z^i \leqq -3 \end{cases}$$

Comparison of true and fitted probabilities

z	0	1	2	3	4
P_1	0.5	0.731	0.881	0.953	0.981
$\hat{P}_1$	0.5	0.67	0.83	1.0	1.0

Comparison of true and fitted effect on aggregate demand for alternative 1 of a one-unit increase in each X_i: true = 93, predicted = 100.

due to a one unit increase in the explanatory variable is 100 units; the true model yields an increase of 93 units. Thus, the linear probability model yields a sever percent overestimate of the true demand effect. The order of magnitude of this bias has been found to be relatively stable in limited Monte Carlo studies with data in range typically observed in transportation surveys. Further, a bias of this magnitude may not be serious when viewed against the pure statistical variability of the forecast. Thus, in many cases, the linear probability model may give satisfactory forecasts of aggregate demand elasticities.

5.4. Estimation of nonlinear response functions

We next consider estimation of the binary choice model in eq. (5.3) when G is a smooth ogive. The first estimation method we shall discuss is due to Berkson (1953) and is applicable when there are repeated observations for each value of the vector of explanatory variables. Changing notation slightly, let $i = 1, \ldots, I$ denote the levels (z^i, s^i) of the vectors of explanatory variables. Let R_i denote the number of observations at level i, and let r_i denote the number of times alternative 1 is chosen. Then r_i is binomially distributed, with $E(r_i/R_i) = P_{1i}$. Let g denote the inverse of the function G; e.g., for the logistic function $G(V) = 1/(1 + e^{-V})$, $g(P) = \log[P/(1 - P)]$.[1] By a Taylor's expansion,

$$g(r_i/R_i) = g(P_{1i}) + g'(\rho_i)\left[\frac{r_i}{R_i} - P_{1i}\right], \tag{5.9}$$

where ρ_i is a value between P_{1i} and r_i/R_i. Provided $R_iP_{1i} \gg 1$ and $R_i(1 - P_{1i}) \gg 1$, the second term in eq. (5.9) is to a close approximation a normally distributed random variable with zero mean. Hence we can rewrite this equation as

$$g(r_i/R_i) = \beta' z^i + \varepsilon_i, \tag{5.10}$$

[1] The specific form of the inverse function $g(P)$ can be derived by solving the original function, $G(V) = P = 1/(1 + e^{-V})$, for V. Considering the second equality and multiplying by $(1 + e^{-V})$ we obtain $P + Pe^{-V} = 1$, $e^{-V} = (1 - P)/P$, or $e^V = P/(1 - P)$. Taking logs of both sides, we derive the result $V = \log[P/(1 - P)] = g(P)$. For the arctan model, $G(V) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1} V$ has the inverse $V = \tan\left[\pi\left(P - \frac{1}{2}\right)\right]$; while for the probit model, $g(P)$ is the inverse cumulative standard normal curve.

where ε_i is a normally distributed error term, and apply ordinary least squares to estimate β. This procedure yields estimates that converge with probability one to the true values as the number of repetitions for each level i grows to infinity.

The case of the model (5.10) most commonly treated in the literature is the logistic distribution, yielding

$$\log\left[\frac{r_i}{R_i - r_i}\right] = \beta' z^i + \varepsilon_i. \tag{5.11}$$

Several modifications can be made in eq. (5.11) to improve the accuracy of the estimates. Cox (1970) shows that a slightly improved normal approximation and adjustment for heteroskedasticity can be attained by applying ordinary least squares to the model

$$\log\left[\frac{r_i + 1/2}{R_i - r_i + 1/2}\right] = \beta' z^i + \varepsilon_i, \tag{5.12}$$

using the resulting estimator $\bar{\beta}$ to calculate consistent estimates of the probabilities

$$P_i = 1/(1 + \mathrm{e}^{-\beta' z^i}),$$

and then applying least squares to the model

$$w_i \log\left[\frac{r_i + 1/2}{R_i - r_i + 1/2}\right] = \beta' z^i w_i + \varepsilon_i, \tag{5.13}$$

where

$$w_i = \sqrt{[R_i P_i (1 - P_i)]}.$$[2]

This model is known to give good estimates of the parameters, provided that a sufficient number of repetitions at each i value can be obtained; this is even true for samples of moderate size. [See Berkson (1955), Gart (1967), and Gilbert (1968).] However, application of the method to survey data involves two serious difficulties. First, the number of cells necessary to describe the possible configurations of explanatory variables

[2] It is possible to make a "one-pass" approximation to this procedure by applying least squares to eq. (5.13) with the empirical weights,

$$w_i = \sqrt{[R_i(r_i + 1)(R_i - r_i + 1)/(R_i + 1)(R_i + 2)]}.$$

However, as Cox (1970, p. 41) points out, the factors "$+\frac{1}{2}$" in eq. (5.13) and "$+1$" in the weight above may require modification to improve the small sample characteristics of the estimator in this case.

tends to increase with the power of the number of variables. For example, a model with K independent binary variables requires 2^K cells. Hence even for moderate K, the survey sample sizes necessary to obtain a few repetitions in every cell may be extremely large. This will be a particularly acute problem in transportation surveys, where many cells will have probabilities P_{1i} near zero or one, making it necessary to obtain a large number of repetitions to satisfy the conditions $R_iP_{1i} \gg 1$ and $R_i(1 - P_{1i}) \gg 1$.

The second difficulty in applying the method to survey data is that many explanatory variables are continuous, making a dichotomization necessary in order to define cells. Hence the Berkson regression, using cell means as values of the explanatory series, introduces an errors-in-variables effect which in general causes the magnitude of the estimates to be biased downward. In order to make the Berkson procedure statistically consistent when this problem is present, it is necessary to redefine the cells as the sample size increases so that the number of repetitions in every cell increases and the range of variation within each cell decreases. We report below on a Monte Carlo experiment for a very simple model with one explanatory variable which indicates that the Berkson procedure may provide desirable estimates even in the presence of the errors-in-variables effect. This suggests that the Berkson method should be used whenever it is feasible .However, despite its statistical advantages, this procedure is considerably less useful in analyzing survey data than it is in a laboratory setting because of the difficulty of defining cells for a large number of continuous explanatory variables.

The last estimation procedure we shall discuss is the method of maximum likelihood. This method does not require repetitions, and it can be adapted to a variety of estimation problems. Its primary disadvantage is that it involves much more costly computation than the preceding methods because the estimates must be obtained by numerical methods. Its primary advantage is that the estimates are consistent, and are the best possible estimates in very large samples. Limited Monte Carlo studies and analytic solutions [McFadden (1973a)] suggest that the maximum likelihood estimators are satisfactory in small samples, though not as desirable as the Berkson estimator when the latter procedure is feasible.

The maximum likelihood procedure could be applied to the probability function $P_i = G(\beta' z^i)$ for any distribution function G; however, we shall consider only the logit case $\log[P_{1i}/(1 - P_{1i})] = \beta' z^i$. We again let $i = 1, \ldots, I$ index individual observations. Since f_{1i} is binomially distributed, we can write the log of the probability of observing a given sample as

$$\begin{aligned} L &= \sum_{i=1}^{I} [f_{1i} \log P_{1i} + (1 - f_{1i}) \log(1 - P_{1i})] \\ &= \sum_{i=1}^{I} \log(1 + \exp(\beta' z^i)) + \sum_{i=1}^{I} f_{1i} \beta' z^i. \end{aligned} \tag{5.14}$$

This is termed the log likelihood function. Now suppose β is unknown. The method of maximum likelihood argues that the calculated probability of observing the given sample should be highest when the unknown β is near the true value, and hence that a satisfactory estimate of the parameters is the maximand of the log likelihood function, or, in other words, a value $\hat{\beta}$ which maximizes L.

The maximum can be found in the ordinary way by differentiating eq. (5.14) with respect to β and setting the derivatives equal to zero. The solution of the resulting set of equations yields the maximum likelihood estimates $\hat{\beta}$ for β. The first-order condition for a maximum is

$$\partial L / \partial \beta = \sum_{i=1}^{I} (f_{1i} - P_{1i}) z^i = 0. \tag{5.15}$$

The second-order condition for a maximum is

$$\frac{\partial^2 L}{\partial \beta \partial \beta'} = - \sum_{i=1}^{I} z^i P_{1i}(1 - P_{1i}) z^{i'} < 0. \tag{5.16}$$

The right hand side of eq. (5.16) is the negative of a weighted moment matrix. Hence provided that the data are not multicollinear, the matrix of second partial derivatives of L is negative definite, implying that L is strictly concave, the maximum likelihood estimate is unique, and the second order condition holds.

The mathematical properties of the likelihood function are quite useful in obtaining numerical solutions to the maximization problem. Provided that the explanatory variables are not multicollinear, the existence of the maximum is virtually certain in empirical samples of more than ten or twenty observations. Further, it is possible to use a finite quadratic programming algorithm, described in McFadden (1973a), to test for existence. The concavity of the log likelihood function allows the use of rapid iterative search procedures which are guaranteed to converge to the maximum. The empirical estimates given in chapter 7 were obtained using a mixed Newton–Raphson variable metric routine with a linear search method due to Davidon.[3] A general consequence of the theory is that in asymptotically large samples the covariance matrix of the maximum likelihood estimates, weighted by the square root of the sample size, equals the inverse of the negative of the expected value of $\partial^2 L/\partial\beta\partial\beta'$, evaluated at the true parameter vector. But from eq. (5.16),

$$\frac{-E\partial^2 L}{\partial\beta\partial\beta'} = \sum_{i=1}^{I} z^i P_{1i}(1 - P_{1i}) z^{i\prime}, \tag{5.17}$$

an expression ordinarily computed in the iterative search procedure. Evaluation of the inverse of the expression in eq. (5.17) at the maximum likelihood estimate provides a consistent estimator of the covariance matrix of the maximum likelihood estimator.

5.5. A Monte Carlo comparison of nonlinear estimators

We shall next describe a small Monte Carlo experiment which compares the maximum likelihood, Berkson, and linear probability model estimators in the case of the simplest possible binary logit model,

$$\log\frac{P_{1i}}{1 - P_{1i}} = \beta_1 z_1^i, \tag{5.18}$$

where the scalars z_1^i are drawn from a (continuous) logistic distribution. Table 5.1 gives selected results for the parameter value $\beta_1 = 1$, with the z_1^i distributed with mean zero and a semi-interquartile range of 1.10 in case A and 4.39 in case B. Case A yields selection probabilities between 0.2 and 0.8 for 88 percent of the observed values of z_1, while case B yields selection probabilities in this range for 33 percent of the observed values of z_1. For choice of mode, transportation surveys tend to yield a large number of "clear-cut" choices with extreme selection probabilities, and thus this case more closely resembles case B.

The Berkson estimator calculated here ranks the values of the independent variable, and then assigns observations to cells on the basis of

[3] This program was developed at the University of California at Berkeley by McFadden, Varian and Wills.

rank. The estimator is obtained by applying the least squares procedure of eq. (5.12) using cell means for the independent variable. If the linear probability model is normalized so that the parameter is comparable to the logit model parameter, it is just the Berkson estimator for cell size one.

Provided that systematic bias and standard errors are weighed equally, the mean square error is the best criterion for comparing the Berkson, maximum likelihood, and linear probability estimators. The mean square error is equal to the expected value of the squared deviation of the estimate from the true parameter value, and in the Monte Carlo study it is estimated by the average of this squared deviation over a series of randomly generated samples.

The statistical theory of the Berkson estimator suggests that in the absence of the effects of grouping, the expected mean square error of the estimator is minimized by taking maximum possible cell sizes and a small number of cells. When the continuous independent variable is dichotomized, one expects the optimal cell size to depend on the size of

Table 5.1

Small sample properties of the mȃximum likelihood estimator (MLE), Berkson estimator (BE), and linear probability model estimator (LPE).

	Case A						
Sample size	Bias			Mean square error			Berkson cell size
	MLE	BE	LPE	MLE	BE	LPE	
30	0.008	−0.243	−0.565	0.127	0.111	0.332	3
60	0.010	−0.103	−0.549	0.069	0.056	0.461	5
120	−0.062	−0.135	−0.550	0.071	0.079	0.313	6
240	0.096	−0.001	−0.508	0.027	0.011	0.259	6
	Case B						
30	0.098	−0.319	−0.742	0.181	0.138	0.551	10
60	−0.050	−0.295	−0.742	0.045	0.100	0.551	15
120	−0.005	−0.276	−0.776	0.034	0.083	0.603	20
240	−0.062	−0.214	−0.774	0.014	0.052	0.599	40

Note: The independent variable has a logistic distribution with mean zero and semi-interquartile range equal to 1.10 in case A and 4.39 in case B. The true parameter value is 1.0. Twenty Monte Carlo trials are calculated for each sample size. The Berkson estimator is reported for the cell size giving the minimum mean square error.

the sample and the variation of the independent variable. Table 5.2 compares the mean square errors of Berkson estimators with various cell sizes; the results for the cell size minimizing mean square error are those given in table 5.1.

We first consider the results of case A in table 5.1, where the independent variable results in selection probabilities near zero or one for only a small percentage of cases. The tabled figures represent the average of twenty Monte Carlo trials for each sample size, and because they are themselves subject to statistical variation, one should avoid attempting to draw more than general qualitative conclusions from the results. However, it is clear that the linear probability model estimator (LPE) has a large systematic bias and mean square error. While the maximum likelihood estimator (MLE) and Berkson estimator (BE) have comparable mean square errors at each sample size, the BE has a somewhat larger systematic bias, as a result of which it underestimates the magnitude of the parameter in every case. This is the expected outcome of the introduction of an errors in variables effect due to grouping, and it suggests that improvement in the BE might be obtained by introducing a correction for the grouping error. The statistical properties of the BE and the MLE are comparable; however, the BE is preferable because it is easier to calculate.

We next consider case B in table 5.1, where the independent variable yields many selection probabilities near zero or one. Here the effects of grouping on the BE are much more severe than in the previous case, and comparison of the mean square errors indicates that the MLE is clearly superior for sample sizes above 60. However, as before, the MLE and the BE are different because the former has a somewhat higher variance while the latter has a substantial systematic bias, and it is again possible that a correction for grouping would make the BE comparable to the MLE.

In both cases A and B the LPE is inferior to the MLE and the BE. Its performance is relatively worse, however, where there are a high proportion of extreme probabilities.

The results of varying the cell size in computing the Berkson estimator given in table 5.2 show that increasing the cell size from one to more than one results in a substantial improvement, while beyond this range mean square error is a relatively flat function of cell size. Comparison of cases A and B indicates that a high proportion of extreme selection

Table 5.2

Variation of the mean square error of the Berkson estimator with changing cell size.

Case A: Semi-interquartile range of $z^i = 1.10$

Sample size	Number of cells	Cell size	Mean square error
30	30	1	0.33
	15	2	0.16
	10	3	0.11
	6	5	0.11
	5	6	0.13
	3	10	0.17
60	60	1	0.30
	30	2	0.13
	20	3	0.08
	15	4	0.07
	12	5	0.06
	10	6	0.08
	6	10	0.10
120	120	1	0.31
	60	2	0.14
	40	3	0.09
	30	4	0.07
	24	5	0.08
	20	6	0.08
	15	8	0.08
	12	10	0.08
	10	12	0.10
	6	20	0.11
240	240	1	0.26
	120	2	0.08
	80	3	0.03
	60	4	0.02
	48	5	0.01
	40	6	0.01
	30	8	0.01
	24	10	0.02
	20	12	0.02
	12	20	0.02
	6	40	0.02
	4	60	0.02

Table 5.2 (continued)

Case B: Semi-interquartile range of $z^i = 4.39$

Sample size	Number of cells	Cell size	Mean square error
30	30	1	0.55
	15	2	0.38
	10	3	0.28
	6	5	0.19
	5	6	0.17
	3	10	0.14
60	60	1	0.55
	15	4	0.22
	10	6	0.16
	6	10	0.10
	5	12	0.10
	4	15	0.11
	3	20	0.08
120	120	1	0.60
	30	4	0.27
	20	6	0.20
	12	10	0.13
	10	12	0.11
	8	15	0.10
	6	20	0.08
	5	24	0.09
	4	30	0.10
	3	40	0.12
240	240	1	0.60
	60	4	0.27
	40	6	0.19
	24	10	0.12
	20	12	0.11
	16	15	0.08
	12	20	0.07
	10	24	0.07
	8	30	0.06
	6	40	0.05
	5	48	0.07
	4	60	0.07
	3	80	0.09

probabilities in the sample require larger cell sizes in order to provide satisfactory estimates of cell relative frequencies, and the use of these larger cell sizes introduces a larger bias due to grouping.

5.6. *Estimation of the multiple-choice model*

We will now consider estimation of the multiple-choice model with $i = 1, \dots, I$ observations and J_i alternatives at observation i. We start from the "strict utility" probability model described in eqs. (4.39) and (4.40), and examine multiple-choice generalizations of the linear probability model and the logit model. These models can both be derived from eq. (4.39) by appropriate specification of the functional form of the "representative" component of utility $V(x, s)$. To simplify notation, we shall at this point assume that the attributes determined by x and s can be summarized in a vector of numerical functions $z = Z(x, s) = (Z^1(x, s), \dots, Z^K(x, s))$. Further, we suppress any alternative-specific parameters; for example, we write $V(x^i, s) = \beta' Z(x^i, s)$ rather than $V(x^i, s) = \beta_i' Z(x^i, s)$. This does not imply any restriction on the generality of the model since alternative-specific effects can be introduced by defining each variable to be zero on all except one alternative.

We give several examples to illustrate the generality of this formulation, and to indicate how "ranked" or "unranked" alternatives are treated in the model.

The simplest example is the case in which the number of alternatives J_i available for each individual $i = 1, \dots, I$ is constant and the alternatives of all the individuals are ranked (for example, $j = 1$ is always "no-trip", alternative $j = 2$ is "auto trip", alternative $j = J_i = 3$ is "bus trip"). Consider an explanatory variable z_k^{ji}, where k indexes the variable, j indexes the alternative, and i indexes the individual. In this model, z_k^{ji} may be a "generic" variable, such as "trip cost", which will be zero for no-trip, the auto out-of-pocket and parking charges for an auto trip, and the bus fare for a bus trip. "Trip cost deflated by wage rate", a variable describing the interaction between socioeconomic characteristics of the individual and attributes of the alternative, would be another such generic variable. Alternately, z_k^{ji} may be a "mode-specific" variable, such as a variable which is one for the auto trip alternative and zero otherwise and indicates a "pure auto demand preference" shift effect. A variable which equals auto trip cost for the auto trip alternative and is zero otherwise and which indicates a "mode-specific" demand curve is another example of a variable of this kind. It is essential to have ranked alternatives in order to include mode-specific variables. For example, a variable which is one for the first mode and zero otherwise is meaningless unless the first mode had some specific identification, such as the "no-trip" alternative. Furthermore, as we pointed out earlier, it is possible to predict the effect of the introduction of new modes only if all the variables in the model are generic; otherwise the new mode will also have mode-specific effects which cannot be forecast without direct observation on choices including this mode. Thus it is desirable to confine the analysis to generic variables whenever possible, in order to make the model useful for policy purposes. Whether mode-specific effects are present is an empirical question. Tests of the significance of mode-specific variables can be carried out in the case of ranked alternatives, and if these effects are not significant, an empirical basis is established for the use of a generic-variables model.

The second example we consider is the case in which the alternatives available to each individual are ranked and the number of alternatives may vary from individual to individual. For example, the alternatives might be "no-trip" ($j = 1$), "auto-trip" ($j = 2$), "local bus trip" ($j = 3$), and "express-bus trip" ($j = 4$), ranked always in this order, and the fourth alternative might not be available to all the individuals sampled. In this model both generic and mode-specific variables can be given a meaningful interpretation. For example, a variable which is one for mode 4 and zero otherwise could be used to forecast an "express-bus" pure shift effect. This effect could then be included in forecasting the effects of an extension of express-bus service to individuals to whom it is not currently available. In a similar fashion, one could pool samples of individuals who have the "no-trip" option with those who do not.

The last example we consider is the case in which the alternatives available to each individual are unranked. In this case there may be either an equal or an unequal number of alternatives for each individual, and there is no natural correspondence between the "first" alternatives of different individuals. For example, the alternatives may be different destinations of shopping trips, these alternatives being described by generic variables such as time and cost of trip to the destination, attributes of the destination such as flexibility for multipurpose shopping trips, ease of parking, etc. Individuals sampled from dispersed geographical areas will face different lists of alternative shopping areas, and there will normally be no meaningful way to "pair" alternatives from one individual to another. Hence in this model it is not meaningful to introduce "mode-specific" variables, as they would not reflect real behavior, and the model must be of the generic-variable form. However, there can be mixed models in which, for example, the first two alternatives are ranked, (for example alternative $j = 1$ is always "no-trip" and alternative $j = 2$ is always "central business district trip"), while the remaining alternatives represent local shopping trips and are unranked. In such a mixed model, the ranked alternatives may have "mode-specific" variables, but the unranked alternatives can depend only on generic variables. For example, in the illustration above, the explanatory variables could be a "no-trip" shift effect which is one for this alternative and zero otherwise, a "CBD" shift effect, and generic variables measuring inclusive cost of trip and attractiveness of destination.

We will now discuss estimation of multiple-choice selection probabilities by means of the linear probability model and the logit model.

A multiple-choice linear probability model can be obtained as in eqs. (4.65) and (4.66), by setting $V(z^i) = \log \beta' z^i$. The sum over the alternatives for each individual of the resulting linear probability model must equal one. For this condition to hold either z^{ji} for one alternative must be defined as the necessary residual, or all the z^{ji} must be normalized by a weight w_i. However, both these procedures imply that the "representative" utility of one alternative depends on the attributes of all available alternatives, and this is contrary to the assumptions about the independence of tastes and opportunities that are usually made. Furthermore, the weighting procedure in the second case destroys the simple linear regression structure, and imposition of inequality constraints imposes an additional computational nonlinearity. In view of these drawbacks and the additional disadvantage that the linear model is sensitive to specification errors, we conclude that the multinomial linear probability model formulated in eqs. (4.65) and (4.66) does not yield a practical estimator with satisfactory statistical properties. However, as noted below, under some conditions it is possible to use a linear approximation to a nonlinear model to obtain relatively satisfactory parameter estimates in small samples.

5.7. Multinomial logit analysis

A multiple-choice logit model is obtained in eq. (4.68) by taking $V(x, s)$ to have the linear-in-parameters form $V(x^i, s^i) = \beta' z^i$. Then the selection probabilities satisfy

$$\log (P_{ji}/P_{1i}) = \beta'(z^{ji} - z^{1i}), \tag{5.19}$$

or

$$P_{ji} = \frac{1}{\sum_{k=1}^{J_i} \exp(\beta'(z^{ki} - z^{ji}))}. \tag{5.20}$$

When there are repetitions at each level of the vector of explanatory variables, eq. (5.19) can be adapted to a Berkson-type analysis, with β estimated by least squares applied to the equation

$$\log\left[\frac{r_{ji} + 1/2}{r_{1i} + 1/2}\right] = \beta'(z^{ji} - z^{1i}) + \varepsilon_{ji}, \tag{5.21}$$

for $i = 1, \ldots, I$ and $j = 2, \ldots, J_i$, and where r_{ji} is the number of times alternative j is chosen at level i. Efficient estimation requires a two-pass least squares procedure which takes account of heteroskedasticity and covariance of the "within-i" dependent variables. This procedure is discussed in Theil (1970), and the question of the appropriate choice of weights is discussed in McFadden (1973a). The advantages and disadvantages of this procedure are the same as those that were mentioned in the discussion of the binary choice case.

When individual decisions are observed and repetitions are not available, the multiple-choice logit model can be estimated by the maximum likelihood procedure. The log likelihood function in this case is given by

$$L = -\sum_{i=1}^{I} \sum_{j=1}^{J_i} f_{ji} \log\left[\sum_{k=1}^{J_i} \exp(\beta'(z^{ki} - z^{ji}))\right], \tag{5.22}$$

where, as before, $f_{ji} = 1$ if alternative j is chosen and $f_{ji} = 0$ otherwise. McFadden (1973a) has shown that this function is concave in the parameter vector β, implying that there is a unique maximum likelihood estimator whenever a maximum exists. As in the binary choice case, the derivatives of the log likelihood function are readily computed, with

$$\partial L/\partial\beta = \sum_{i=1}^{I}\left[\sum_{j=1}^{J_i}(f_{ji} - P_{ji})z^{ji}\right], \tag{5.23}$$

and

$$\partial^2 L/\partial\beta\partial\beta' = -\sum_{i=1}^{I}\sum_{j=1}^{J_i}(z^{ji} - \bar{z}^i)P_{ji}(z^{ji} - \bar{z}^i)', \tag{5.24}$$

where

$$\bar{z}^i = \sum_{j=1}^{J_i} z^{ji}P_{ji}, \tag{5.25}$$

and

$$P_{ji} = \frac{\exp(\beta' z^{ji})}{\sum_{k=1}^{J_i}\exp(\beta' z^{ki})}. \tag{5.26}$$

Provided the data z^{ji} are not multicollinear, they will normally satisfy a full-rank, or non-degeneracy, condition which guarantees that the Hessian matrix in eq. (5.24) is negative definite. Then L is strictly concave and any vector $\hat{\beta}$ satisfying $\partial L/\partial\beta = 0$ is a unique maximizer for the likelihood function; hence there is a unique maximum likelihood estimator. McFadden (1973a) has given conditions for the existence of the maximum likelihood estimator; he has also given a finite algorithm to test for existence and a demonstration that existence is virtually certain in samples of reasonable size. The maximization of L, which is equivalent to the solution of the system of equations

$$\sum_{i=1}^{I}\sum_{j=1}^{J_i}(f_{ji} - P_{ji})z^{ji} = 0, \tag{5.27}$$

can be carried out by a variety of standard iterative procedures such as the gradient, Newton–Raphson, Fletcher–Powell, and Davidon methods. The procedure used in this study is a mixed Newton–Raphson variable metric routine with a linear search method due to Davidon.[4]

[4] A general purpose statistical program for analysis of qualitative data, QUAIL (for quantitative, intermittent, and limited dependent variable statistical analysis program), written by Wills, Glanville, and McFadden of the University of California, Berkeley, is available for this analysis. The program also allows transgeneration and storage of data, and selection of variables, alternatives, and cases.

A typical Newton–Raphson iteration, starting from a candidate parameter vector $\bar{\beta}$ and associated probabilities $\bar{P}_{ji}$ from eq. (5.26), has the form

$$\hat{\beta} = \bar{\beta} + \left[\sum_{i=1}^{I} \sum_{j=1}^{J_i} (z^{ji} - \bar{z}^i)\bar{P}_{ji}(z^{ji} - \bar{z}^i)'\right]^{-1} \cdot \left[\sum_{i=1}^{I} \sum_{j=1}^{J_i} (z^{ji} - \bar{z}^i)(f_{ji} - \bar{P}_{ji})\right]. \tag{5.28}$$

Note that $\hat{\beta}$ can be interpreted as the ordinary least squares estimator in the linear model

$$\sqrt{(P_{ji})} \cdot (f_{ji} - \bar{P}_{ji}) = \sqrt{(P_{ji})} \cdot \beta'(z^{ji} - \bar{z}^i) + \varepsilon_{ji}, \tag{5.29}$$

and that for a fixed initial $\bar{\beta}$, such as $\bar{\beta} = 0$, eq. (5.29) can be interpreted as a linear probability model. The estimates obtained from eq. (5.29) by a single Newton–Raphson iteration are not consistent. However, we know from experience that rough estimates obtained in this way usually agree in sign and magnitude with the full maximum likelihood estimates. This is particularly likely when the frequency of extreme selection probabilities in the sample is low, and for small samples the statistical properties of these estimates are often as good as those of the maximum likelihood estimator.

5.8. *Measures of goodness of fit*

A goodness of fit measure is a summary statistic indicating the accuracy with which a model approximates the observed data. In the case of qualitative response models, accuracy may be judged either in terms of the fit between calculated probabilities and observed response frequencies, or in terms of the ability of the model to forecast observed responses.

Measures of the first type are based on the observed frequencies f_{ji} and corresponding probabilities P_{ji} calculated from the estimated model. By analogy to the measure of fit used in regression analysis, we could define a sum of squared residuals $\sum_{i=1}^{I} \sum_{j=1}^{J_i} (f_{ji} - P_{ji})^2$. Because the terms entering this sum are heteroskedastic (i.e., have differing variances), the analogy with regression analysis suggests a correction to achieve heteroskedasticity. Define a sum of squared adjusted residuals,

$$S(\beta) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (f_{ji} - P_{ji}(\beta))^2/P^*_{ji}, \tag{5.30}$$

where β is the parameter vector at which the calculated probabilities $P_{ji}(\beta)$ are being computed and P^*_{ji} are the true selection probabilities. McFadden (1973a) has shown that this measure has the same statistical properties in large samples (where the f_{ji} are relative frequencies with repetitions) as does the sum of squared residuals measure in regression analysis. This suggests as a goodness of fit measure the analogue of the multiple correlation coefficient,

$$R^2 = 1 - S(\hat{\beta})/S(\bar{\beta}), \tag{5.31}$$

where $\hat{\beta}$ is the maximum likelihood estimator, $\bar{\beta}$ is zero or is zero except for coefficients of alternative dummies, and P^*_{ji} is replaced by its consistent estimator $P_{ji}(\hat{\beta})$. Values of this index are roughly comparable to multiple correlation coefficients obtained in ordinary least squares. However, the index lacks desirable statistical properties in small samples, and is very sensitive to model specification error at extreme probabilities, these specification errors biasing the index toward one.

A much more satisfactory measure of goodness of fit can be obtained from the log likelihood function,

$$L(\beta) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} f_{ji} \log P_{ji}(\beta). \tag{5.32}$$

A term f_{ji} log $P_{ji}(\beta)$ is near zero if alternative j is chosen and the calculated probability P_{ji} of this outcome is near one, and is large negative if the probability of this outcome is near zero. The log likelihood function has a convenient statistical distribution in large samples, and can be given an intuitive interpretation using information theory; e.g., Theil (1969, 1970). We can transform the log likelihood function into an index analogous to the multiple correlation coefficient by defining

$$\rho^2 = 1 - L(\hat{\beta})/L(\bar{\beta}), \tag{5.33}$$

where $\hat{\beta}$ is the maximum likelihood estimator and $\bar{\beta}$ is zero or is zero except for coefficients of alternative dummies. Suppose $\bar{\beta}$ contains $\bar{k} \geqq 0$ parameters and $\hat{\beta}$ contains $\hat{k}$ parameters, including the parameters that appear in $\bar{\beta}$. Then, in large samples, $[\hat{k}/(\hat{k} - \bar{k})]$ $[\rho^2/(1 - \rho^2)]$ is distributed approximately $F(\hat{k} - \bar{k}, \hat{k})$; this distribution can be used to test the hypothesis $\beta = \bar{\beta}$. The ρ^2 and R^2 indices both vary in the unit interval (except when some coefficients in $\bar{\beta}$ are excluded from β, in which case a poor fit may yield ρ^2 or R^2 negative); the graph below summarizes schematically a relatively stable empirical relationship between the indices.

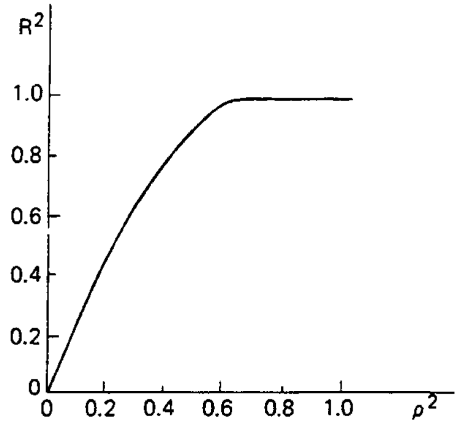

Fig. 5.5.

In terms of consistency and statistical properties, the ρ^2 index appears to provide a practical and theoretically sound index of goodness of fit.

The second type of measure of fit is based on the accuracy of the model in forecasting observed responses. In order to define these measures, we must first discuss briefly the problem of classification and forecasting. Suppose the selection probabilities P_{ji} are known. Assume that if an individual chooses an alternative j and any other alternative is forecast, a cost c_j of misclassification is incurred.[5] The forecasting rule is a function $\delta_j = \delta_{ji}(P_{1i}, \ldots, P_{J_i i})$ with $\delta_j = 1$ if choice j is forecast and $\delta_j = 0$ otherwise. Noting that the expression $1 - \delta_{ji}(P_{1i}, \ldots, P_{J_i i})$ is zero if j is forecast and one otherwise, the total cost of misclassification is

$$C = \sum_{i=1}^{I} \sum_{j=1}^{J_i} f_{ji} c_j (1 - \delta_{ji}). \tag{5.34}$$

Then, the expected cost of misclassification,

$$EC = \sum_{i=1}^{I} \sum_{j=1}^{J_i} P_{ji} c_j (1 - \delta_{ji}), \tag{5.35}$$

[5] This formulation implicitly ranks the alternatives; one could more generally specify a cost c_{ji}^k of forecasting choice k for individual i when the actual choice is j. This requires a more cumbersome analysis, but the conclusions are similar.

is minimized if the forecasting rule selects an alternative j maximizing $P_{ji}c_j$; i.e., $\delta_{ji} = 1$ implies $c_jP_{ji} \geqq c_kP_{ki}$. We adopt this forecasting rule.[6]

A variety of indices of goodness of fit could be based on the cost statistic C valued at the optimal decision rule and using calculated values of the P_{ji} from maximum likelihood estimation or a similar statistical procedure. For example, one might take as an index the reduction in cost as a proportion of the cost of a random classification. However, we shall confine our attention to the case of all $c_j = 1$, where cost is proportional to the total number of individuals misclassified, and the case of $c_j = 1/\bar{P}_j$, where $\bar{P}_j = (1/I)\sum_{i=1}^{I} P_{ji}$ is the sample average probability of choosing j. In the first of these cases, $c_j = 1$, an appealing index of goodness of fit is the proportion of successful forecasts $1 - C/I$. We note that the use of maximum likelihood estimators in the decision rule will tend to maximize this proportion only in asymptotically large samples. Hence, in small samples the values of this percentage may be somewhat erratic. Manski (1974) has developed an alternative family of estimators which maximize the proportion of successful forecasts and have desirable statistical properties.

An implication of the optimal forecasting rule for $c_j = 1$ is that alternatives which have a low average probability $\bar{P}_j$ will be forecast relatively infrequently; intuitively this is because the numbers of misclassifications resulting from directing forecasts away from low probability alternatives is low. Consequently, the individual forecasting rule will tend to underestimate the frequency of aggregate choice of less-used alternatives. A second measure which corrects this bias is based on the weights $c_j = 1/\bar{P}_j$. Since $c_jP_{ji} = P_{ji}/\bar{P}_j$ is on average equal for various j, the aggregate frequencies of forecasts resulting from the optimal decision rule will tend to cluster around the observed aggregate frequencies, as desired. A weighted proportion of successful forecasts,

$$\lambda = 1 - \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i} f_{ji}(1 - \delta_{ji})/\bar{P}_j}{\sum_{i=1}^{I}\sum_{j=1}^{J_i} f_{ji}(1 - 1/J_i)/\bar{P}_j}, \tag{5.36}$$

should provide a satisfactory goodness of fit measure in this case.

[6] Anderson (1958) shows that this decision rule has optimal statistical properties.