INTERPRETATION AND DIAGNOSTICS OF THE MULTINOMIAL LOGISTIC REGRESSION

Danny Steinberg, Economics, University of California, San Diego

ABSTRACT

Multinomial logistic regression (MNL) is a powerful tool for the analysis of problems with a categorical dependent variable having more than two values. It has had applications in evaluation of clinical trials, market research, transportation research, and has been applied in virtually every social science and management analysis context. Nonetheless, most potential users are unfamiliar with the principles underlying the model, are not aware that there are two radically distinct versions of the MNL model, cannot interpret the typically confusing output correctly, and do not know how to assess a model for goodness of fit. This paper seeks to address this problem by laying out the basic theory underlying the model, shows how the coefficients should be interpreted, and derives some useful diagnostics. We show how the results of previously published SAS macros, or of PROC MLOGIT (Steinberg (1986)) can be processed further to yield tables of elasticities of outcome probabilities, and a prediction success table suitable for diagnosing model failure. We also show how to use and interpret White's Information Matrix test and how to test for the Independence of Irrelevant alternatives.

1. Introduction

The multinomial logit (MNL) model has roots in biometric, psychometric and econometrics and has been applied with some regularity in the fields of medicine and transportation research over the last two decades. The core application is the regression modelling of a categorical dependent variable such as a consumer choice or an outcome of a drug therapy; the purpose is to identify factors that are useful in predicting a particular outcome and to quantify that importance with point estimates for regression coefficients. In spite of the fact that a great many processes are better modelled as discrete rather than as continuous random variables, the MNL model has largely remained a fringe technique. Standard textbooks offer little coverage of this rather involved methodology, and only a handful of specialised texts have emerged over the last decade. McCullagh and Nelder (1983) are an exception among general statistics texts; specialised volumes include Ben Akiva and Lerman (1985), Wrigley (1985), Maddala (1983) and Train (1986) and the review articles of McFadden (1979, 1982, 1984). It is still possible for a graduate student specializing in methodology to complete his studies with no training in this area. Although software to ease the burden of computing these models has emerged to render the estimation activity painless, the lack of familiarity with MNL has left most practitioners unsure of when to use the model and what to make of the output when they do use it. Although the standard SAS release has a procedure that can be coaxed to perform polytomous MNL (PROC CATMOD) it does not provide nearly enough diagnostic output and it does not permit estimation of McFadden’s conditional multinomial logit model.

The purpose of this paper is to take a step in the direction of remedying this problem, by explaining in simple terms what the MNL model is, how to interpret the coefficients, and how to reduce the output from the model to a readily understandable form. It is possible, with effort, to compute the diagnostics we discuss in the standard release of SAS, but there is no reasonable way to estimate the conditional MNL model. However, the diagnostics and both versions of logit are fully automated and efficiently programmed in PROC MLOGIT, an after-market SAS procedure.

The layout of this paper is as follows. In section 2, we outline some of the formulas underlying the MNL model, and develop the polychotomous (or polytomous) logit model. In section 3, we discuss the interpretation of the coefficients, and show how tables of elasticities and prediction success can be calculated. Section 4 introduces the conditional logit model and contrasts it with the polytomous (or universal) logit. In section 5 we develop a set of diagnostic tests to assess the goodness of fit of the model. Throughout, we illustrate our points with reference to the output of some SAS jobs running PROC MLOGIT.

2. Multinomial Logit Model: Fundamentals

A classic application of MNL is to discrete choice problems, for example, when a voter must cast a ballot in a primary election for one of the candidates, say Hart, Glenn or Mondale. This three options example is simple enough to grasp, yet contains all the elements necessary to illustrate the most complex applications. Our goal may be to determine which voters are attracted to which candidate, so as to forecast how the candidates would perform in each state in the nation. Our regressors might very well include variables such as age, sex, race, education, income, marital status, and so forth. Although tables produced by PROC FREQ will yield a broad stroke picture, the fine detail will be available only when we control for many demographics as well as possible interactions. It is evident, that if we code the dependent variable as 1, 2 and 3 for the candidates in the order listed above, ordinary regression will yield no useful information, as the coding of the dependent variable is arbitrary.

The theoretical apparatus for the MNL model is best approached by first looking at the basic probability formulas. For MNL, we can think of the list of regressors as generating a score for each candidate. The score is determined by the usual predicted value formulas

\[ n_j = x_j\beta_1 + x_j\beta_2 + x_j\beta_3 + \cdots + x_j\beta_k \]

(1)

In spite of the fact that a great many processes are better modelled as discrete rather than as continuous random variables, the MNL model has largely remained a fringe technique. Standard textbooks offer little coverage of this rather involved methodology, and only a handful of specialised texts have emerged over the last decade. McCullagh and Nelder (1983) are an exception among general statistics texts; specialised volumes include Ben Akiva and Lerman (1985), Wrigley (1985), Maddala (1983) and Train (1986) and the review articles of McFadden (1979, 1982, 1984). It is still possible for a graduate student specializing in methodology to complete his studies with no training in this area. Although software to ease the burden of computing these models has emerged to render the estimation activity painless, the lack of familiarity with MNL has left most practitioners unsure of when to use the model and what to make of the output when they do use it. Although the standard SAS release has a procedure that can be coaxed to perform polytomous MNL (PROC CATMOD) it does not provide nearly enough diagnostic output and it does not permit estimation of McFadden's conditional multinomial logit model.

The purpose of this paper is to take a step in the direction of remedying this problem, by explaining in simple terms what the MNL model is, how to interpret the coefficients, and how to reduce the output from the model to a readily understandable form. It is possible, with effort, to compute the diagnostics we discuss in the standard release of SAS, but there is no reasonable way to estimate the conditional MNL model. However, the diagnostics and both versions of logit are fully automated and efficiently programmed in PROC MLOGIT, an after-market SAS procedure.

The layout of this paper is as follows. In section 2, we outline some of the formulas underlying the MNL model, and develop the polychotomous (or polytomous) logit model. In section 3, we discuss the interpretation of the coefficients, and show how tables of elasticities and prediction success can be calculated. Section 4 introduces the conditional logit model and contrasts it with the polytomous (or universal) logit. In section 5 we develop a set of diagnostic tests to assess the goodness of fit of the model. Throughout, we illustrate our points with reference to the output of some SAS jobs running PROC MLOGIT.

2. Multinomial Logit Model: Fundamentals

A classic application of MNL is to discrete choice problems, for example, when a voter must cast a ballot in a primary election for one of the candidates, say Hart, Glenn or Mondale. This three options example is simple enough to grasp, yet contains all the elements necessary to illustrate the most complex applications. Our goal may be to determine which voters are attracted to which candidate, so as to forecast how the candidates would perform in each state in the nation. Our regressors might very well include variables such as age, sex, race, education, income, marital status, and so forth. Although tables produced by PROC FREQ will yield a broad stroke picture, the fine detail will be available only when we control for many demographics as well as possible interactions. It is evident, that if we code the dependent variable as 1, 2 and 3 for the candidates in the order listed above, ordinary regression will yield no useful information, as the coding of the dependent variable is arbitrary.

The theoretical apparatus for the MNL model is best approached by first looking at the basic probability formulas. For MNL, we can think of the list of regressors as generating a score for each candidate. The score is determined by the usual predicted value formulas

\[ n_j = x_j\beta_1 + x_j\beta_2 + x_j\beta_3 + \cdots + x_j\beta_k \]

(1)
\[ y_i = X_i \beta_1 + X_i \beta_2 + X_i \beta_3 + \cdots + X_i \beta_k \]

where the \( X_i \)'s are the regressors and the \( \beta \)'s are the coefficients to be estimated. Equation (1) shows how the scores would be calculated for a single observation. The input data consists of the value of the dependent variable \( y_i \), which would be 1, 2 or 3, and the variables \( X_1 \cdots X_k \). The MNL model will provide us with several sets of coefficients so that a separate score can be computed for each potential value of the dependent variable. In our example, we have three candidates, so three scores would need to be computed. As one of the scores can be normalized to zero, implying that one of the sets of coefficients can be normalized to zero, only two sets of coefficients need be estimated. In PROC MLOGIT, the score for the highest value of the dependent variable is always normalized to zero; therefore in our example, the score for Mondale will always be zero. Of course this could be changed by rearranging the coding of the data, for example, letting Mondale = 1, Hart = 2 and Glenn = 3.

For a given observation the probability formulas are then:

(2) \[ \text{prob}(y=1) = \frac{e^y_1}{e^y_1 + e^y_2 + e^y_3} \]

and in general

(3) \[ \text{prob}(y=i) = \frac{e^y_i}{\sum_j e^y_j} \]

where \( J \) is the number of alternatives in the choice model.

Predicted probabilities are easy to derive; we just exponentiate the scores and form ratios.

3. Estimation and Interpretation

To estimate our model in PROC MLOGIT we would write, say

PROC MLOGIT DATA=PRIMARY NCAT=3;
MODEL CHOICE = AGE SEX RACE INCOME;

The initial output would look something like this:

Multinomial Logit Analysis

<table>
<thead>
<tr>
<th>Dependent Variable: CANDIDATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Of Records Processed: 38</td>
</tr>
<tr>
<td>Number Of Choices in Each Category:</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

### independent variable means

<table>
<thead>
<tr>
<th>Variable</th>
<th>D=1</th>
<th>D=2</th>
<th>D=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>AGE</td>
<td>28.25</td>
<td>26.13</td>
<td>28.02</td>
</tr>
<tr>
<td>BLACK</td>
<td>.0000</td>
<td>.0000</td>
<td>.7692E-01</td>
</tr>
<tr>
<td>INCOME</td>
<td>4786.</td>
<td>4951.</td>
<td>5558.</td>
</tr>
</tbody>
</table>

log likelihood at iteration 7 is -26.332768
convergence achieved.

### Results of Estimation

Log Likelihood: -26.33276

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>-4.8536</td>
<td>6.6030</td>
<td>-0.7350</td>
</tr>
<tr>
<td>AGE</td>
<td>.3565</td>
<td>.2208</td>
<td>1.6144</td>
</tr>
<tr>
<td>BLACK</td>
<td>-10.5395</td>
<td>30.0032</td>
<td>-0.3512</td>
</tr>
<tr>
<td>INCOME</td>
<td>-1.259E-02</td>
<td>7.463E-03</td>
<td>-1.0171</td>
</tr>
</tbody>
</table>

-2 times log likelihood ratio (chi squared): 10.008572
with 6 degrees of freedom

### WALD TESTS ON COEFFICIENTS ACROSS ALL CHOICES

<table>
<thead>
<tr>
<th>Source</th>
<th>Chi-Square</th>
<th>D.F.</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>0.92733</td>
<td>2</td>
<td>0.62898</td>
</tr>
<tr>
<td>AGE</td>
<td>2.60638</td>
<td>2</td>
<td>0.27166</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.22735</td>
<td>2</td>
<td>0.89255</td>
</tr>
<tr>
<td>INCOME</td>
<td>4.31721</td>
<td>2</td>
<td>0.11549</td>
</tr>
</tbody>
</table>

These coefficients are not straightforward to interpret; we have two coefficients for each regressor. Some are significant in one place and not in another, others switch signs. What do we conclude from this output?

It turns out that each set of coefficients has an interpretation something akin to dummy variables in ordinary regression. Each set of coefficients represents essentially the outcome of a binary logit model in which a two choice problem has been analyzed; in each case, the pair of "choices" consists of the reference choice, which is indicated by the highest value of the dependent variable, and one of the other choices. In our example, the first set of coefficients represents how the regressors influence the choice between Hart and Mondale; the second set of coefficients represents how the regressors influence the choice between Glenn and Mondale. It is not difficult to imagine higher income persons favoring Hart over Mondale in that two man race giving INCOME a positive coefficient in the first vector, yet higher income persons favoring Mondale over Glenn, giving INCOME a negative coefficient in the second vector.

Because the coefficients are always measuring movement between a pair of choices, one of which is a fixed reference choice, the printed results will change if we recode the dependent variable. This is no more surprising than the analogous situation for regression with dummy variables. Thus, the left out dummy constitutes the "reference group", and all coefficient estimates represent shifts of the regression line relative to the reference group. In the MNL
model, we are measuring shifts in the probability of an outcome relative to the reference group. We obtain information on contrasts between each choice and the reference choice, but we do not obtain any direct information on the contrast between any two non-reference choices. In our example, we do not obtain direct information on how the regressors affect the choice between Hart and Glenn.

If the results of the MNL model are really conglomerations of pairwise choices, and are interpretable only in a pairwise fashion anyway, why bother with this model at all? Why not estimate each pair separately with simpler software? (For further discussion of pair-wise estimation see Begg and Gray(1984)). There are three chief reasons for using the MNL model with all choices. First, the the full model is fully efficient relative to the pooled pairwise submodels that could be estimated by the binary logit specification. Second, the full model permits hypothesis tests of restrictions across several choices. Third, if we are really interested in the larger problem, the MNL results can be packaged in an interpretable way.

One use of the information available only from the joint estimation appears in the table of Wald tests beneath the main regression results. These tests are of the joint hypotheses that all the coefficients associated with a particular variable are zero. For example, the test statistic for the hypothesis that both AGE coefficients are zero is a Chi-squared variable with two degrees of freedom, and is not significant.

The information presented so far provides only a partial picture of the model results and performance, restricted to a perspective of single or multiple pairwise comparisons against a single reference group. All of the printed results (except the log likelihood) would change if we reordered the coding of the dependent variable. A method for providing coefficient reports for all possible choices of the reference group has been suggested by Molyneaux and Stone(1985), and PROC CATMOD permits alternative selection of the default group without recoding. However, to understand the model fully we need information that is invariant with respect to the reference group and that provides an intuitive picture of the results.

In the example problem above, the critical information we want is how the regressors affect the absolute distribution of votes across the three candidates. Essentially, this is captured by a table of derivatives giving the change in probability of each choice with respect to a change in each regressor. Recalling equations (1) and (2) we note that each regressor appears in the formula for the score for each choice; in consequence, changing AGE for example influences not only $s_1$, but also $s_2$. In the multinomial case therefore, we may not be able to easily assess the overall impact of a regressor on the probability of a choice. The formula for the derivatives can be written as:

$$ \frac{\partial P_j}{\partial x_k} = P_i (\beta_k - \sum_{i=1}^{J-1} P_i \beta_k) $$

The influence of the $k$th regressor on the $j$th probability can be be decomposed into two parts. The first is a direct effect which is given by the coefficient $\beta_k$ and is proportional to the probability of the $j$th choice. The second component is the negative of a weighted average of the other coefficients $\beta_k$ -- the coefficients for the regressor $x_k$ in the other scores. Even though the direct effect $\beta_k$ might be positive the overall impact might still be to reduce the probability of the $j$th choice. The reason -- the same regressor may have an even bigger effect on some other choices.

This information is most conveniently presented in the form of a table which is produced by PROC MLOGIT. The table shows us how a small change in the regressors will reallocate probability.

<table>
<thead>
<tr>
<th>Independent Variable Derivatives</th>
<th>VARIABLE</th>
<th>D-1</th>
<th>D-2</th>
<th>D=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CONSTANT</td>
<td>-1.010</td>
<td>1.905</td>
<td>-3.854E-01</td>
<td></td>
</tr>
<tr>
<td>2 AGE</td>
<td>1.10E-01</td>
<td>-5.80E-02</td>
<td>-1.690E-01</td>
<td></td>
</tr>
<tr>
<td>3 BLACK</td>
<td>-3.57E-01</td>
<td>-1.67E-01</td>
<td>3.231</td>
<td></td>
</tr>
<tr>
<td>4 INCOME</td>
<td>-2.682E-04</td>
<td>-8.57E-04</td>
<td>1.30E-03</td>
<td></td>
</tr>
</tbody>
</table>

Here we see that increasing AGE increases the probability for candidates one and two, at the expense of candidate three, whereas BLACKs are drawn to candidate three drawing primarily from candidate two, but also from candidate one. Similarly higher income persons are more likely to prefer candidate three, at the expense of candidates one and two.

Note that the sum of all changes across a row is always zero; if probability is increased in one category, there must be offsetting corresponding decreases probability in another category or categories. Also, the derivative table is a nonlinear function of the data and the estimated coefficients; a separate table could be constructed for each individual. PROC MLOGIT gives the user the option of evaluating the derivatives at the sample average of the $X$ variables, although this practice is not recommended. By default, the table is computed for each observation in the data and then averaged.

This derivative table represents a complete summary of the model predictions, based on all the coefficients estimated. Using it can supplant working with the coefficients once standard errors have been computed. The standard errors of the derivatives are computed by a one term Taylor series approximation, but where not reported above.
4. Discrete Choice

The MNL model as developed above can be thought of as a classification tool - a way of assigning observations to responses - rather than a vehicle for the analysis of behavior. A variant of the model, introduced by McFadden, is designed to allow the analysis of discrete choice, and is known as the conditional logit model. For this variant of MNL, the roles of the independent variables and estimated coefficients are reversed from that of the polytomous logit model. Letting \( \beta \) be the vector of parameters to be estimated and \( X_i \) the vector of data for the \( i \)'th alternative, then for a given observation the probability that the \( i \)'th outcome is chosen is:

\[
\text{prob}(i) = \frac{e^{\beta^T X_i}}{\sum_j e^{\beta^T X_j}}
\]

Note that here a single set coefficients \( \beta \) are estimated, and it is the \( X \)'s that vary between possible outcomes. (Indeed they must vary for the model to be estimated.) An interpretation of the model is that the \( X \)'s are measurements of characteristics of the alternatives among which an individual can choose, and the \( \beta \)'s represent value weights which individuals place on these characteristics. For example, if the choice between alternative travel modes to work is sensitive to the price and total travel time of each mode, then the attractiveness of an alternative is modelled as

\[
U_j = \beta_1 \text{time}_i + \beta_2 \text{price}_i + \text{error}
\]

where time, and price, are the characteristics of the \( i \)'th choice. The characteristics of the other choices do not appear in equation (6). The probability that a person chooses the \( i \)'th alternative is the probability that it is the most attractive option available, which is the familiar MNL formula when the error term is extreme value.

A crucial difference between the polytomous and conditional logit models is that in the latter the number of coefficients estimated does not increase with the number of alternatives analysed. The \( \beta \) values we need increases instead. To analyse the choices among say the options of walking, riding the bus, riding the train and driving to work in Manhattan, we would need the travel time and money cost of each alternative for each data point in the sample.

A second difference between the two models is that the number of alternatives open to each person in the study sample may vary - some may have four options, others may only have two or three.

As a simple example, we turn to a different problem, involving the prediction of the type of fish a fishing boat will gear itself for, on the basis of the expected revenues available from each option. For each boat we have data on four alternatives available, stored in the SAS variables REV1-REV4. The syntax for this model is a little different than the usual model statement, and is given in the following statements:

```sas
PROC MLOGIT NCAT=4 NOCON; WEIGHT RATE;
MODEL DEPVAR = REV (REV1 REV2 REV3 REV4);
```

Here we indicate that there is only one coefficient to be estimated, namely REV (which is a label and not a SAS variable), and that the relevant data for the four options are in the SAS variables appearing in parentheses. The PROC statement indicates that we will not be estimating a constant, and that there are four possible values for the dependent variable.

The resulting output is:

### ESTIMATION OF THE CONDITIONAL LOGIT MODEL

**OBSERVATIONS WILL BE WEIGHTED BY RATE**

**SUM OF WEIGHTS** = 106.01391

**NUMBER OF CASES AND SUM OF WEIGHTS**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>6.632234</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>11.76016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>28.27845</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>60.15341</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DEPENDENT VARIABLE MEANS**

<table>
<thead>
<tr>
<th>VAR</th>
<th>D-1</th>
<th>D-2</th>
<th>D-3</th>
<th>D-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>REV</td>
<td>378.97</td>
<td>392.93</td>
<td>415.57</td>
<td>541.86</td>
</tr>
</tbody>
</table>

**CONVERGENCE ACHIEVED.**

<table>
<thead>
<tr>
<th>NUMBER OF ITERATIONS</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOG LIKELIHOOD</td>
<td>-135.14978</td>
</tr>
</tbody>
</table>
| LOG LIKELIHOOD
| CONSTANTS ONLY MODEL | -147.78257 |

**VARIABLE PARAMETER STD ERR T-STAT**

| REV | .379D-03 | 5.851D-01 | 4.6058 |

Note that if this model had been estimated by the polytomous logit method we would have had 4*(4-1)/2 = 12 coefficients estimated instead.

For further details on the conditional logit see the references cited in the introduction.

5. Diagnosis

How do we tell if we have a good fit? Several methods have been proposed in the literature; some are informative, and others are useless, but no definitive metric has yet been accepted. We will discuss some of these diagnostic here.

It is instructive to begin with one of the obvious measures: prediction success. A predicted probability for each candidate is implied for each data point; if we add up all these predicted probabilities over the entire data set, we obtain the aggregated model predictions. These are total expected votes for each option (candidate). We could compare actual and expected votes for each value of the dependent variable. Unfortunately, this diagnostic is of limited value because any model which includes a constant term will show a perfect fit under this criterion; equating the summation of predicted probabilities for each choice and the actual number of observations selecting each option is one of the first order conditions defining the maximum likelihood estimator. Because of this feature of the model
McFadden has introduced the somewhat more involved prediction success table, reproduced below.

<table>
<thead>
<tr>
<th>Choice</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>4.00</td>
<td>8.00</td>
<td>26.00</td>
<td>38.00</td>
</tr>
</tbody>
</table>

% successfully predicted: 0.262, 0.281, 0.744

The counts down by looking at the number of votes for each choice, this table breaks success into three parts: the proportion of successful predictions. If this model also predicts only 10.5% of those voting for candidate one, then it would successfully choose candidate three. It would get 0.68*26 = 17.68 persons, or 0.68 percent correctly predicted. This model with no explanatory power (a constants only model) appears to generate a rather large proportion of successful predictions. If this model also predicted that 4/38 = 10.5% of the sample would choose candidate one, then it would successfully predict only 10.5% of those voting for candidate one, and so would appear to be poor on this criterion.

The success index helps to put the first model into perspective. It tells us that this model does not do all that much better than chance in predicting candidates two or three; its main strength is in predicting more accurately than chance for candidate one. At the present time, we do not have a test statistic for transforming these results into a criterion for judging model adequacy. Methods for formalizing this assessment of goodness of fit have been proposed by Costanza et al. (1982) and Hubert and Golledge (1982).

A second element of diagnosis is White's Information Matrix test. For correctly specified models $A(\beta) = B(\beta)$ where

$$A(\beta) = E \left[ \frac{\partial \ln L}{\partial \beta} \left( \frac{\partial \ln L}{\partial \beta} \right)^T \right], \quad B(\beta) = E \left[ \frac{\partial \ln L}{\partial \beta} \left( \frac{\partial \ln L}{\partial \beta} \right)^T \right]$$

where $\ln L$ is the log likelihood of the model and $\beta$ is the vector of coefficients to be estimated. The equality states that the negative of the expected value of the matrix of second derivatives of the log likelihood with respect to the estimated parameters is equal to the matrix formed by summing the outer product of the score (gradient) vector across all observations. Often this equality is exploited for computational purposes as when the right hand side is used as an approximation for the left side for parameter updates and calculation of standard errors at convergence. White has shown how to test this equality; the test is based on a quadratic form on a vector of the differences between the elements of the two estimates of the information matrix. Letting $d$ be a vector of element by element differences between these two matrices, the test statistic is:

$$d'V^{-1}d$$

where $V$ is an appropriate variance estimator for $d$. This is asymptotically distributed as a Chi-squared statistic with degrees of freedom equal to the number of elements in $d$. The test need not be conducted for each distinct element of the symmetric matrices. As the Information matrix test requires calculation of third derivatives of the likelihood function, simplifications have been sought. One approximation has been suggested by Lancaster (1984) and is widely used in econometrics. However, Spady (1986) has discovered that this approximation has very poor finite sample properties, suggesting that the test be computed exactly whenever possible.

Several specification tests specific to the conditional logit model have also been proposed. The simplest to conduct is to compare the results to a polytomous specification in which all the variables appearing in the conditional model statement are entered. A likelihood ratio test can then be conducted.

As a final remark, it is important to be aware of the sensitivity of the logit model to outliers, and methods for detecting influential observations have been introduced by Pregibon (1981) and Cook and Weisberg (1982).

Danny Steinberg, 3672 Caminito Cielo del mar
San Diego, CA 92130 (619) 491-8599

References


Begg, Colin B. and Robert Gray (1984): Calculation of polytomous logistic regression parameters...
using individualized regressions, Biometrika 71, 11-18.


