A list of Working Papers on the last pages

No. 128, 1984

INFORMATION CRITERION AND ESTIMATION OF MISSPECIFIED QUALITATIVE CHOICE MODELS

by

David Brownstone

August, 1984

August, 1984

INFORMATION CRITERION AND ESTIMATION OF MISSPECIFIED QUALITATIVE CHOICE MODELS*

by David Brownstone**

University of California, Irvine

and Industrial Institute for Economic

and Social Research, Stockholm

Abstract

This paper investigates misspecified estimation and model selection criteria derived from the "Information Criterion (see Akaike (1973))" for qualitative choice models. Four estimators for the "Information Criterion" are derived for general qualitative choice models. Two of these estimators were previously derived by Akaike (1973) and Chow (1981) for arbitrary likelihood functions. The new estimators are derived by taking analytic expectations of the log likelihood function. A number of Monte Carlo experiments are performed using binominal logit models to investigate the behavior of the Information Criterion estimators with realistic sample sizes. The new analytic estimators are more accurate than the more general estimators, but they do not always perform as well in minimizing prediction or estimation error. Monte Carlo results also show that the usual asymptotic distribution properties of the maximum likelihood estimator are poor approximations for sample sizes as large as 1,000 observations with only two variables.

* This is a heavily revised version of a paper presented at the Econometric Society meetings in Dublin, September 1982. Gregory Chow, Thomas Cooley and Ken Small provided many useful comments, but I take full responsibility for all errors. Financial support was provided by the Industrial Institute for Economic and Social Research and the National Science Foundation through Grants Nos. SES-8002010 and SES-8012582.

** Mailing Address: School of Social Science
University of California
Irvine, California 92717

1 Introduction

This paper examines the application of Akaike's "Information Criterion (Akaike (1973))" for selection and estimation of misspecified qualitative choice models. In addition to their common axiomatic basis, these topics are related by a simple practical consideration. If there is only one "true" model, then consideration of more than one implies estimation of at least one misspecified model. Throughout this paper, the superscript "^" denotes Maximum Likelihood Estimate (MLE). Subscripts on expectation operators denote the distribution over which the expectation is taken. Thus $E_{\hat{\theta}}$ denotes the expectation with respect to the sampling distribution of the MLE of θ.

The Information Criterion attempts to select models which, on average, yield better future predictions. "Better" means a smaller expected sum of the log-likelihood ratios of the true density of future observations to the one specified by the model. If $g(\cdot)$ is the true density of each of n independent future observations $(\tilde{y}, \ldots, \tilde{y}_N) = \tilde{Y}$ and $f(\cdot|\theta)$ is the density specified by a possible model, then better predictions by $f(\cdot|\theta)$ means a smaller expectation

$$I_N(g;\ f(\cdot|\theta)) = E \sum_{i=1}^{N} \log(g(\tilde{y}_i)/f(\tilde{y}_i|\theta)) \geq 0, \qquad (1.1)$$

where the expectation is evaluated with respect to the true density $g(\cdot)$.

(1.1) is just the Kullback-Leibler (1951) information measure, or the expected log-likelihood ratio, to discriminate between the two models. Bayesians have criticized this Information Criterion on the grounds that it is inadmissable. However, Chow (1981) argues that its use can be justified (from a Bayesian standpoint) in large samples and that the Bayesians have not yet produced a better alternative. Eq. (1.1) also defines a metric on the space of all probability functions, so the Information Criterion can be interpreted as selecting the approximate model whose density is closest to the truth.

If θ is estimated by $\hat{\theta}$, then the model selection criterion (1.1) is estimated by:

$$E_{\hat{\theta}} I_N(g;\ f(\cdot|\hat{\theta})) \equiv E_{\hat{\theta}}\{E_{\tilde{Y}} \sum_{i=1}^{N} (\log g(\tilde{y}_i) - \log f(\tilde{y}_i|\hat{\theta}))\}. \qquad (1.2)$$

Taking the expectation with respect to $\hat{\theta}$ helps reduce the variability across researchers using independent calibration data sets.

If two approximate models, $f_1(\cdot|\theta)$ and $f_2(\cdot|\theta)$, are being compared, then it is only necessary to compute $E_{\hat{\theta}}E_{\tilde{Y}}(\log f(\tilde{y}|\hat{\theta}))$ since $E_{\hat{\theta}}E_{\tilde{Y}}(\log g)$ is constant for all approximate models. Therefore, all that is needed to implement the model selection criterion (1.2) is a practical method for estimating

$$E_{\hat{\theta}}E_{\tilde{Y}}(\log f(\tilde{Y}|\hat{\theta})). \tag{1.3}$$

Estimators for (1.3) have been derived by Akaike (1973) and Chow (1981) both for general likelihood functions and the linear model.

These estimators are specialized to qualitative choice models in the second section, and two new Information Criterion estimators for qualitative choice models are also derived. The new estimators are derived by first evaluating expectations with respect to $\tilde{Y}$ in Equation (1.3), so they are called Evaluated Information Criterion (EIC) estimators. Computational formulas are given for general qualitative choice models and the Multinominal Logit model.

Section 3 gives the results of a Monte Carlo study on the sampling properties of the various information criterion estimators. This study also investigates the sampling distribution of the maximum likelihood estimator (MLE) for correct and misspecified models. The most striking result is that the usual asymptotic normal approximation for the MLE is quite poor. For the designs considered in this study, these approximation errors overwhelm the standard error corrections for misspecified models given in Chow (1981) and White (1982). Unfortunately, these approximation errors also cloud the comparison between the information criterion estimators. The EIC estimators seem to be more accurate, and all of the information criterion selection rules are more likely to reject smaller models than likelihood ratio tests at the 5 or 10 percent level. The Monte Carlo experiments discussed in this paper used binominal logit models because of their low computational cost and popularity. Experiments were also run using binominal probit models to ensure that the results are not sensitive to choice of functional form.

The Monte Carlo results for the small sample behavior of the MLE confirm and extend the study in Davidson and MacKinnon (1984). Preliminary work also indicates that the small sample behavior of the MLE deteriorates further as the number of independent variables and/or the number of discrete alternatives increases. Therefore it would appear that many of the published estimates and test statistics for qualitative choice models should be used cautiously. Much more work is needed to delineate where the asymptotic approximations are valid. One promising alternative to the likelihood-based techniques is the Bayesian approach given by Zellner and Rossi (1984). If the numerical integrations they use are feasible, then the Bayesian approach avoids small-sample problems.

2 Derivation of Information Criterion Estimators

As in Chow (1981), the true density will be a member of a finite dimensional parametric family of densities denoted $f(y|\Theta_0)$. Any approximate model is derived by imposing linear restrictions on Θ:

$$H\Theta + b = 0 \tag{2.1}$$

Although this is not the most general framework where the Information Criterion can be applied, it does encompass most of the situations which arise in applied econometric work. According to the Information Criterion, the best approximate model achieves the highest expected information defined by Equation (1.3). Define Θ_* as the parameter of the best approximate model satisfying (2.1), i.e.:

$$\Theta_* = \max_{\Theta} E_{\tilde{Y}} L(\tilde{Y},\Theta) \text{ such that } H\Theta + b = 0 \tag{2.2}$$

where $L(\cdot)$ is the log-likelihood function corresponding to $f(\cdot)$.

Akaike (1973) and White (1982) argue that the natural estimator of Θ_* is the maximum likelihood estimator:

$$\hat{\Theta}_* = \max_{\Theta} L(Y,\Theta) \text{ such that } H\Theta + b = 0. \tag{2.3}$$

White (1982) shows that $\hat{\Theta}_*$ is consistent (for Θ_*) and asymptotically normally distributed. The estimator White provides for the asymptotic covariance of $\hat{\Theta}_*$ is wrong (see Chow (1982) and White (1983)), and the correct estimator requires an estimate of the true parameter Θ_0. Although these ideas are relatively new in the econometrics literature, the reference lists in Akaike (1973) and White (1982) are close to a list of "Who's who" in modern mathematical statistics. Unlike the classical approach, the Information Criterion provides a unifying theoretical framework for modelling the decision processes required in applied work. The main competition is the Bayesian framework, which is frequently hampered by computational problems and the need to specify prior distributions.

One issue which has not been discussed in the literature is modelling the correlation between parameter estimates for two different approximate models based on the same data set. Modelling this correlation should yield more accurate estimates, and the Monte Carlo studies reported in the next section confirm this intuition. However, if maximum likelihood estimators are used, then model selection procedures which account for the correlation between the estimators will always choose the larger of the two nested models. This proposition will be proved for multinominal logit models later in this section. Currently there is no formal proof for the general case, but an intuitive sketch is quite easy:

> Suppose Θ_{**} is derived by imposing additional linear constraints on Θ_*. From information theory (see Kullback (1959))

$$E_{\tilde{Y}} L(\tilde{Y},\Theta_*) \geq E_{\tilde{Y}} L(\tilde{Y},\Theta_{**})$$

> If "^" denotes maximum likelihood estimates from a fixed sample Y, then also $L(Y,\hat{\Theta}_*) \geq L(Y,\hat{\Theta}_{**})$, and this inequality holds for small neighborhoods around $\hat{\Theta}_*$ and $\hat{\Theta}_{**}$. When the difference between the log-likelihoods at $\hat{\Theta}_*$ and $\hat{\Theta}_{**}$ is small, then the correlation between $\hat{\Theta}_*$ and $\hat{\Theta}_{**}$ becomes high. Therefore taking expectations yields

$$E_{\hat{\Theta}_*,\hat{\Theta}_{**}}(L(Y,\hat{\Theta}_*) - L(Y,\hat{\Theta}_{**})) \geqslant 0 \tag{2.4}$$

Finally, if expectations with respect to $\tilde{Y}$ are evaluated at maximum likelihood estimates from the same sample Y, then Kullback's (1959) results imply:

$$E_{\tilde{Y}}E_{\hat{\Theta}_*,\hat{\Theta}_{**}}(L(\tilde{Y},\hat{\Theta}_*) - L(\tilde{Y},\hat{\Theta}_{**})) \geqslant 0$$

One solution to this problem is to treat, as Chow (1981) implicitly does, $\hat{\Theta}_*$ and $\hat{\Theta}_{**}$ as if they are independent. This can be justified by the need to predict how the selected model will perform using new independent data sets. Imposing independence yields reasonable selection criteria for nested models at the cost of reduced accuracy in estimating the basic information measure (Eq. 1.3). It may therefore be better to use estimators which account for correlation between approximate model parameter estimates when comparing approximate models with the same dimensionality.

The remainder of this section presents derivations of the information criterion estimators used in the Monte Carlo study. Formulas for general qualitative choice models are given first, followed by the much simpler versions of these estimators for the popular Multinominal Logit Model. The derivation of Chow's (1981) estimator will be reviewed briefly since it is similar to the derivation of the EIC estimators.

Chow begins by expanding $L(\tilde{Y},\hat{\Theta}_*)$ in a Taylor series expansion around Θ_* yielding

$$L(\tilde{Y},\hat{\Theta}_*) = L(\tilde{Y},\Theta_*) + \frac{\partial L(\tilde{Y},\Theta_*)}{\partial\Theta}(\hat{\Theta}_*-\Theta_*) +$$

$$+ \frac{1}{2}(\hat{\Theta}_*-\Theta_*)' \frac{\Theta^2 L(Y,\Theta_*)}{\partial\Theta\partial\Theta'}(\hat{\Theta}_*-\Theta_*) \qquad (2.5)$$

Taking expectations of each term in (2.5) together with the first order conditions for Eqs. (2.2) and (2.3), Chow derives:

$$E_{\hat{\Theta}_*} E_{\tilde{Y}} L(\tilde{Y},\hat{\Theta}_*) \approx L(Y,\hat{\Theta}_*) -$$

$$- \frac{1}{2}\,\mathrm{tr}\big(J(\Theta_*,\Theta_o)E(\hat{\Theta}_*-\Theta_*)(\hat{\Theta}_*-\Theta_*)'\big) \qquad (2.6)$$

where

$$J(\Theta_*,\Theta_o) \equiv - E_{\tilde{Y}} \frac{\partial^2 L(\tilde{Y},\Theta_*)}{\partial\Theta\partial\Theta'} \qquad (2.7)$$

and Θ_o defines the distribution of Y.

All that is needed to compute Chow's estimator are estimators of $J(\Theta_*,\Theta_o)$ and the covariance of $\hat{\Theta}_*$. Due to a typographical error, the "1/2" factor was omitted from Eq. (2.6) in Chow's paper. Akaike's (1973) estimator is:

$$E_{\hat{\Theta}_*} E_{\tilde{Y}} L(\tilde{Y},\hat{\Theta}_*) \approx L(Y,\hat{\Theta}_*) - k \qquad (2.8)$$

where k is the dimension of Θ_*. This estimator is derived using a different expansion, and, as Chow (1981) points out, incorrectly assumes that:

$$k = \mathrm{tr}\left(J(\Theta_*,\Theta_o)E(\hat{\Theta}_*-\hat{\Theta}_*)(\hat{\Theta}_*-\Theta_o)'\right)$$

This equality only holds if $\Theta_* = \Theta_o$.

The log-likelihood function of the true model for the general qualitative choice model is given by:

$$L = \sum_{n=1}^{N} \sum_{i=1}^{J_n} Y_{in} \log P_{in}(Z_{in},\Theta_o) \qquad (2.9)$$

where Y_{in} is the number of times case n chose discrete alternative i ($i = 1, \ldots, J_n$), Z_{in} is a vector of exogenous attributes, and Θ_o is a vector of fixed parameters. P_{in} is the conditional probability that case n will choose alternative i given values for Θ and Z_{in}. A popular choice for P_{in} is

$$P_{in} = \frac{\exp(Z_{in}\,\Theta)}{\sum_{j=1}^{J_n} \exp(Z_{jn}\,\Theta)} \tag{2.10}$$

which generates the Multinominal Logit (ML) model.

For the general qualitative choice model given in Eq. (2.9)

$$\begin{aligned} J(\Theta_*,\Theta_o) &= -\,E_{\tilde{Y}} \sum_n \sum_i \tilde{Y}_{in} \frac{\partial^2 \log P_{in}(Z_{in},\Theta_*)}{\partial\Theta\partial\Theta'} = \\ &= -\sum_n \left(\sum_i \tilde{Y}_{in}\right) \sum_i P_{in}(Z_{in},\Theta_o) \frac{\partial^2 \log P_{in}(Z_{in},\Theta_*)}{\partial\Theta\partial\Theta'} \end{aligned} \tag{2.11}$$

which can be consistently estimated by replacing Θ_o and Θ_* with any consistent estimators. An estimator for the covariance matrix of $\hat{\Theta}_*$ is derived in the Appendix.

The EIC estimators are derived by first taking expectations with respect to $\tilde{Y}$ in Eq. (1.3) yielding:

$$E_{\hat{\Theta}_*} E_{\tilde{Y}} \sum_n \sum_i \tilde{Y}_{in} \log P_{in}(\hat{\Theta}_*) =$$

$$= E_{\hat{\Theta}_*} \sum_n \sum_i R_n \; P_{in}(\Theta_o) \log P_{in}(\hat{\Theta}_*) \qquad (2.12)$$

where $R \; = \sum_i \tilde{Y}_{in}$

Unfortunately Θ_o is unknown, but it can be replaced by its maximum likelihood estimator, $\hat{\Theta}$, yielding

$$E_{\hat{\Theta}} E_{\hat{\Theta}_*} \sum_n \sum_i R_n \; P_{in}(\hat{\Theta}) \log P_{in}(\hat{\Theta}_*) \equiv E_{\hat{\beta}} \; Q(\hat{\beta}) \qquad (2.13)$$

where $\hat{\beta} = \binom{\hat{\Theta}}{\hat{\Theta}_*}$. As in the derivation of Chow's estimator, the expectation with respect to $\hat{\beta}$ can be approximated by expanding $Q(\hat{\beta})$ around the true values β_o yielding:

$$E_{\hat{\beta}} \; Q(\hat{\beta}) \approx Q(\hat{\beta}) + \frac{1}{2} \, \mathrm{tr} \left[\left(\frac{\partial^2 Q(\hat{\beta})}{\partial\beta\partial\beta'} \right) E(\hat{\beta}-\beta_o)(\hat{\beta}-\beta_o)' \right] \quad (2.14)$$

In the Appendix it is shown that $N^{1/2}(\hat{\beta}-\beta_o)$ has an asymptotic Normal distribution with mean 0 and covariance matrix

$$\lim_{N\to\infty} \begin{bmatrix} NJ^{-1}(\Theta_o,\Theta_o) & J^{-1}(\Theta_o,\Theta_o)KS' \\ SK'J^{-1}(\Theta_o,\Theta_o) & N^{-1}SVS' \end{bmatrix} \qquad (2.15a)$$

where $K_{ij} = \text{Cov}\,[\frac{\partial L(\Theta_o)}{\partial \Theta_i}, \frac{\partial L(\Theta_*)}{\partial \Theta_j}]$,

$V = \text{Cov}\,[\frac{\partial L(\Theta_*)}{\partial \Theta}]$, and

$$\begin{bmatrix} S & Q \\ Q' & R \end{bmatrix} = \begin{bmatrix} N^{-1}J(\Theta_*,\Theta_o) & -H \\ -H' & 0 \end{bmatrix}^{-1} = \qquad (2.15b)$$

$$= \begin{bmatrix} NJ^{-1}(\Theta_*,\Theta_o)-NJ^{-1}(\Theta_*,\Theta_o)H(H'J^{-1}(\Theta_*,\Theta_o)H)^{-1}H'J^{-1}(\Theta_*,\Theta_o) & -J^{-1}(\Theta_*,\Theta_o)H(H'J(\Theta_*,\Theta_o)H)^{-1} \\ -(H'J^{-1}(\Theta_*,\Theta_o)H)^{-1}H'J^{-1}(\Theta_*,\Theta_o) & -N^{-1}(H'J^{-1}(\Theta_*,\Theta_o)H)^{-1} \end{bmatrix}$$

Chow (1981) shows that if the restrictions are exclusion restrictions on a subset of parameters, then

$\Theta_* = (\Theta_*^1, 0) \quad H' = (0 \;\; I)$ and:

$$J(\Theta_*,\Theta_o) = \begin{bmatrix} J_{11}(\Theta_*,\Theta_o) & J_{12}(\Theta_*,\Theta_o) \\ J_{21}(\Theta_*,\Theta_o) & J_{22}(\Theta_*,\Theta_o) \end{bmatrix} \tag{2.16}$$

The matrix S in Eq. (2.15b) then becomes:

$$S = \begin{bmatrix} NJ_{11}^{-1}(\Theta_*,\Theta_o) & 0 \\ 0 & 0 \end{bmatrix}$$

The first and second derivatives of the choice probabilities, P_{in}, are computed as part of the computation of the maximum likelihood estimates $\hat{\Theta}$ and $\hat{\Theta}_*$. Therefore, all of the Information Criterion estimators given in this section require only small additional computational costs. Finally, all of the formulas derived in this section remain valid for differentiable nonlinear restrictions $h(\Theta) = 0$. The only required changes are to replace $H'\Theta + b$ by $h(\Theta)$ and interpret H as the matrix of first partial derivatives evaluated at Θ_* or $\hat{\Theta}_*$.

The EIC estimator defined by Equations (2.14) and (2.15) allows for non-zero correlation between $\hat{\Theta}$ and $\hat{\Theta}_*$. This estimator will be called the dependent EIC, or DEIC, estimator in the remainder of this paper. If $K \equiv 0$, then Equations (2.14) and

(2.15) define the independent EIC, or IEIC, estimator. As discussed earlier in this section, the DEIC estimator is expected to be more accurate than the IEIC. The DEIC is more appropriate for comparing approximate models of the same dimensionality.

Although the EIC estimators derived above are not very expensive to compute, they are not as simple as Akaike's estimator (Eq. 2.8) or the standard likelihood ratio test at fixed confidence levels. Fortunately, these formulas can be simplified for the most popular qualitative choice model, the Multinominal Logit (ML) model.

The ML model is defined by Eqs. (2.9) and (2.10) and the necessary derivatives can be calculated from:

$$\frac{\partial P_{in}(\Theta)}{\partial \Theta} = P_{in}(\Theta)(Z_{in}-\bar{Z}_n(\Theta)) \qquad (2.17)$$

where $\bar{Z}_n = \sum_i P_{in}(\Theta) Z_{in}$

From (2.17) we get

$$\frac{\partial L(\Theta_*)}{\partial\Theta} = \sum_n \sum_i (Y_{in} - R_n P_{in}(\Theta_*)) Z_{in} \tag{2.18}$$

and

$$\frac{L(\Theta_*)}{\partial\Theta} - E\,\frac{L(\Theta_*)}{\partial\Theta} = \sum_n \sum_j (Y_{in} - R_n P_{in}(\Theta_o)) Z_{in} \tag{2.19}$$

which is independent of Θ_* and identical to $\partial L(\Theta_o)/\partial\Theta$! Therefore

$$K = V(\Theta_*) = V(\Theta_o) = J(\Theta_o, \Theta_o) = \tag{2.20}$$

$$= \sum_n R_n \sum_i (Z_{in} - \bar{Z}_n(\Theta_o)) P_{in}(\Theta_o)(Z_{in} - \bar{Z}_n(\Theta_o))'$$

Since $V(\hat{\Theta})$ from (2.26) is computed as part of the Newton-Raphson algorithm for finding the MLE, $\hat{\Theta}$, of Θ_o there is no extra computation required for computing an estimate of V in Eq. (2.21). The other required matrix, $J(\Theta_*, \Theta_o)$ is also easy to estimate since

$$J(\Theta_*, \Theta_o) = -\,E\Big(\frac{\partial^2 L(\Theta_*)}{\partial\Theta\partial\Theta'}\Big) = E \sum_n \sum_j R_n Z_{jn} \frac{\partial P_{jn}(\Theta_*)}{\partial\Theta'} = \tag{2.21}$$

$$= \sum_n R_n \sum_j (Z_{jn} - \bar{Z}_n(\Theta_*)) P_{jn}(\Theta_*)(Z_{jn} - \bar{Z}_n(\Theta_*))'$$

where the last step uses the fact that the operation $Z_{jn} \to Z_{jn} - \bar{Z}_n$ is idempotent for all Θ,

Eq. (2.21) is computed as part of the Newton-Raphson algorithm for finding the MLE of Θ_*, $\hat{\Theta}_*$. These results show that, for ML models, Chow's Information Criteria estimator (2.6) can be computed as cheaply as the likelihood ratio test statistic since only a few additional small matrix manipulations are required.

Computation of the EIC estimators can also be simplified for ML models. If Θ and Θ_* are estimated by maximum likelihood, then $Q(\hat{\beta}) = L(\hat{\Theta}_*)$ so that no extra computations are required to get $Q(\hat{\beta})$. To see this identity, consider a simple binominal ML model with no repetitions (i.e., $R_n = 1$ for all n). For convenience, assume that Θ_* is a scalar derived from Θ via exclusion restrictions and that the first alternative is always chosen. Then

$$L(\hat{\Theta}_*) - Q(\hat{\beta}) = \sum_{n=1}^{N} \log P_{1n}(\hat{\Theta}_*) - \sum_{n=1}^{N} P_{1n}(\hat{\Theta})\log P_{1n}(\hat{\Theta}_*)$$

$$+ \left(1-P_{1n}(\hat{\Theta})\right)\log\left(1-P_{1n}(\hat{\Theta}_*)\right) \qquad (2.21)$$

$$= \sum_{n=1}^{N} \hat{\Theta}_*(z_{zn}-Z_{1n})\left(1-P_{1n}(\hat{\Theta})\right) = 0$$

since $\log\left(1-P_{1n}(\hat{\Theta}_*)\right) = \log P_{1n}(\hat{\Theta}_*) + \Theta_*(z_{2n}-z_{1n})$

and $0 = \dfrac{\partial L(\hat{\Theta})}{\partial\Theta} = \displaystyle\sum_{n=1}^{N} \dfrac{\partial \log P_{1n}(\hat{\Theta})}{\partial\Theta} =$

$$= \sum_{n=1}^{N} - (z_{2n}-z_{1n})\left(1-P_{1n}(\hat{\Theta})\right).$$

Similar arguments can be used to prove the proposition for the general ML model, but the notation is considerably more complex. It should also be noted that this equality does not hold for other qualitative choice models, including Multinominal Probit.

The only computation required beyond computing the maximum likelihood estimators is the first block of the hessian of Q:

$$\frac{\partial Q}{\partial\theta\partial\theta'} = \sum_n R_n \sum_i \log P_{in}(\theta_*)\left[(Z_{in}-\bar{Z}_n(\theta_o))P_{in}(\theta_o)\right.$$

$$(Z_{in}-\bar{Z}_n(\theta_o))' - 2\sum_j (Z_{jn}-\bar{Z}_n(\theta_o))P^2_{jn}(\theta_o)$$

$$\left.(Z_{jn}-\bar{Z}_n(\theta_o))'\right] \qquad (2.23)$$

This section can best be summarized by giving explicit formulas for the ML model Information Criterion estimators in the case of exclusion restrictions on a subset of parameters. Using the notation of Eq. (2.16), Chow's estimator (i.e., Eq. 2.6) is:

$$L(Y,\hat{\theta}_*) - \frac{1}{2}\,\mathrm{tr}\,\left(J_{11}(\hat{\theta}_o,\hat{\theta}_o)\; J_{11}^{-1}(\hat{\theta}_*,\hat{\theta}_o)\right) \qquad (2.24)$$

The DEIC estimator is:

$$L(Y,\hat{\theta}_*) + \frac{1}{2}\,\mathrm{tr}\,\left(\frac{\partial Q_{11}(\hat{\theta}_*,\hat{\theta}_o)}{\partial\theta\partial\theta'}\, J^{-1}(\hat{\theta}_o,\hat{\theta}_o) + \right. \qquad (2.25)$$

$$\left. + J_{11}(\hat{\theta}_o,\hat{\theta}_o)\; J_{11}^{-1}(\hat{\theta}_*,\hat{\theta}_o)\right)$$

The IEIC estimator is:

$$L(Y,\hat{\Theta}_*) + \frac{1}{2}\,\mathrm{tr}\,\Big(\frac{\partial Q_{11}(\hat{\Theta}_*,\hat{\Theta}_o)}{\partial\Theta\partial\Theta'}\,J^{-1}(\hat{\Theta}_o,\hat{\Theta}_o) - \qquad (2.26)$$

$$- J_{11}(\hat{\Theta}_o,\hat{\Theta}_o)\,J_{11}^{-1}(\hat{\Theta}_*,\hat{\Theta}_o)\Big)$$

If the DEIC is used to compare nested models and $L(Y,\hat{\Theta}_*)$ is close to $L(Y,\hat{\Theta}_{**})$, then

$$\mathrm{tr}\,\Big(J_{11}(\hat{\Theta}_o,\hat{\Theta}_o)\,J_{11}^{-1}(\hat{\Theta}_*,\hat{\Theta}_o) -$$

$$- J_{11}(\hat{\Theta}_o,\hat{\Theta}_o)\,J_{11}^{-1}(\hat{\Theta}_{**},\hat{\Theta}_o) \approx k_* - k_{**}$$

where $k*$ is the rank of Θ_*. It is then obvious that the DEIC defined in Eq. (2.25) will always yield a higher expected information for the larger model Θ_*.

3 Monte Carlo Results

This section describes the results from some Monte Carlo experiments on the Binominal ML model. These experiments are designed to examine the relative performance of the Information Criterion estimators derived in the previous section. An important by-product is an examination of the sampling distribution of the maximum likelihood estimator for the binominal logit model. Since McFadden's (1973) Monte Carlo study, there have been countless applications of ML models where it has been implicitly assumed that the asymptotic Normal distribution (e.g. Eq. 2.15) is correct. The results described here cast considerable doubt on the validity of the asymptotic approximation for realistic sample sizes.

The Binominal ML model is defined by Equations (2.9) and (2.10) with $J_n = 2$. The choice probability for the first alternative is:

$$P_{in} = \frac{1}{1 + \exp(\Theta_o Z_n)} \tag{3.1}$$

The true model for all the experiments here has two components, and the approximate model restricts the second component of Θ_o to equal zero.

The independent variables, Z_n, for all the experiments are independent draws from a Uniform distribution on (-1/2, 1/2).[1] These basic variables are then transformed to generate different levels of correlation between the two components for different experiments. During initial trials it became apparent that the MLE is very sensitive to outliers in the independent variables. Therefore, averages over a number of draws from the underlying Uniform distribution were used instead of replicating a fixed data set as in Davidson and MacKinnon (1984). The Monte Carlo results here have a higher variance but give a more realistic view of the behavior of the MLE for this particular data generation process.

The experiment definitions are given in Table 1. Each experiment was run for sample sizes: 50, 200, 500, and 1,000. 100 independent choice variables, Y, were drawn for each experiment and each sample

---

[1] All random numbers were generated using the standard APL uniform random variate operator.

Table 1 Definition of experiments

Expt. No.	Correlation between components of Z_n	Variance of True[a] Probabilities	Θ_o	Θ_*	Information[b] of True Model	Information[c] of Approximate Model
1	0.8	.158	5, 1	5.76	-.485	-.4872
2	0.95	.156	5, 1	5.94	-.4787	-.4793
3	0.8	.112	7.8, 1.56	8.98	-.3596	-.3633
4	0.8	.121	5, 4	7.62	-.384	-.409
5	0.3	.167	5, 1	5.2	-.5068	-.5126
6	0.8	.248	0.5, 0.1	0.578	-.6898	-.6898

Notes: a This is $\left[\sum_{n=1}^{N} P_{in}(\Theta_o)(1-P_{in}(\Theta_o))\right]/N$ and a is a measure of how close the true probabilities are to .5. The maximum value is .25 and the minimum is 0.

b This is $\frac{1}{1,000}\sum_{n=1}^{1,000}\sum_{i} P_{in}(\Theta_o)\log P_{in}(\Theta_o) = E_{\tilde{Y}}\, L(\tilde{Y},\Theta_o)/N$

c This is $\frac{1}{1,000}\sum_{n=1}^{1,000}\sum_{i} P_{in}(\Theta_o)\log P_{in}(\Theta_*) = E_{\tilde{Y}}\, L(\tilde{Y},\Theta_*)/N$

size. The six experiments were chosen to represent realistic models with varying degree of misspecification. A more ambitious experimental design would clearly be very useful, but the computational costs are prohibitive.

The most striking result from the Monte Carlo experiments is the poor performance of the standard asymptotic approximation to the distribution of the MLE. Percentage biases for parameter estimates and standard error estimates are given in Tables 2A and B. These biases decline as the sample size increases, but they are still quite large for 1,000 observations.

White (1982 and 1983) and Chow (1981 and 1982) have both claimed that it is important to correct the usual maximum likelihood standard error estimates for model misspecification. Using the notation from Equation (2.16) the incorrect asymptotic variance of $\hat{\Theta}_*$ is

$$J_{11}^{-1}(\Theta_*,\Theta_o) \tag{3.2}$$

while the correct formula is:

$$J_{11}^{-1}(\Theta_*,\Theta_o)\ J_{11}(\Theta_o,\Theta_o)\ J_{11}^{-1}(\Theta_*,\Theta_o) \tag{3.3}$$

Numeric values of these estimators are given in Table 3. Although the "correct" formula performs slightly better in larger samples, the differences

Table 2a Percentage bias of MLE of θ_o

Expt. No.	Sample Size 50	200	500	1,000
1	-3.92, 85.54	-2.75, 24.18	-2.59, 17.04	-2.36, 15.86
2	-16.79, 147.05	-7.86, 42.12	-7.86, 42.12	-6.19, 31.57
3	8.94, 34.72	-1.76, 9.21	-2.35, 13.25	-1.71, 7.47
4	-5.82, 35.52	-3.03, 8.22	-2.73, 7.47	-2.65, 5.98
5	2.79, 69.71	-2.17, 19.63	-1.16, 12.81	-.96, 11.74
6	-39.08, 335.37	-6.88, 82.92	-16.40, 95.81	-17.91, 134.65

Table 2b Percentage bias of standard error estimate MLE of θ_o

Expt. No.	Sample Size 50	200	500	1,000
1	-.09, 8.38	-1.81, 11.73	-6.83, 12.00	-5.76, 12.06
2	-.58, 4.66	6.42, 8.57	4.92, 10.02	6.87, 12.55
3	-13.56, -14.14	-5.39, 9.75	-6.84, -.62	-2.94, -2.65
4	-11.45, -6.87	-9.15, 2.42	-7.53, -.15	-7.01, 8.57
5	-10.60, .65	6.73, 12.23	-9.91, 18.69	-14.63, 17.04
6	2.07, -.85	20.57, 23.03	10.87, 25.56	5.80, 10.71

Table 3 **Percentage bias in standard error estimates for MLE of approximate model θ_***
("correct", "incorrect")

Expt. No.	Sample Size 50	200	500	1,000
1	-8.18, -7.35	-7.35, -7.14	-11.65, -11.59	-16.26, -16.25
2	-17.28, -16.65	-9.77, -9.66	-12.13, -12.13	-16.83, -16.86
3	-16.42, -14.46	.96, 1.53	-4.96, -4.53	3.55, 3.90
4	-5.86, -1.39	-13.88, -11.65	-4.75, -2.51	-9.50, -7.50
5	-11.91, -10.60	8.29, 8.82	-8.30, -8.04	-15.24, -15.03
6	-5.89, -4.98	-2.58, -2.42	.97, 1.04	1.96, 2.01

Table 4a **MSE of information criterion estimators for experiment 1 (· 100)**

Estimator	Sample Size 50	200	500	1,000
Chow	7.91	4.11	2.77	1.92
DEIC	7.26	4.02	2.75	1.91
IEIC	7.78	4.09	2.76	1.91
Akaike	8.11	4.13	2.78	1.92

Table 4b **MSE of information criterion estimators for experiment 2 (• 100)**

Estimator	Sample Size			
	50	200	500	1,000
Chow	8.31	4.22	2.82	1.94
DEIC	7.65	4.14	2.80	1.93
IEIC	8.18	4.20	2.81	1.94
Akaike	8.49	4.25	2.83	1.94

Table 4c **MSE of information criterion estimators for experiment 3 (• 100)**

Estimator	Sample Size			
	50	200	500	1,000
Chow	7.77	3.76	2.61	1.63
DEIC	7.48	3.72	2.60	1.63
IEIC	8.21	3.82	2.62	1.64
Akaike	7.95	3.79	2.61	1.63

Table 4d **MSE of information criterion estimators for experiment 4 (• 100)**

Estimator	Sample Size			
	50	200	500	1,000
Chow	7.03	4.45	2.61	1.82
DEIC	6.65	4.38	2.60	1.81
IEIC	7.28	4.45	2.62	1.82
Akaike	7.25	4.47	2.62	1.82

Table 4e **MSE of information criterion estimators for experiment 5 (• 100)**

Estimator	Sample Size			
	50	200	500	1,000
Chow	7.56	3.42	2.56	1.83
DEIC	6.88	3.34	2.54	1.82
IEIC	7.34	3.41	2.55	1.82
Akaike	7.77	3.45	2.57	1.83

Table 4f **MSE of information criterion estimators for experiment 6 (• 100)**

Estimator	Sample Size			
	50	200	500	1,000
Chow	1.98	.75	.37	.26
DEIC	1.75	.72	.36	.26
IEIC	1.92	.74	.37	.26
Akaike	2.64	.86	.41	.27

Table 5a MSE of estimates of first parameter of Θ_o for experiment 1

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	2.41	1.31	.92	.70
Likelihood Ratio 10 %	2.43	1.33	.88	.65
Chow	2.38	1.22	.81	.56
DEIC	2.39	1.16	.77	.54
IEIC	2.41	1.22	.81	.56
Akaike	2.41	1.34	.87	.61

Table 5b MSE of estimates of first parameter of Θ_o for experiment 2

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	3.65	1.71	1.37	1.12
Likelihood Ratio 10 %	3.83	1.89	1.36	1.11
Chow	4.26	2.00	1.37	1.01
DEIC	4.61	2.02	1.30	.93
IEIC	4.10	2.00	1.38	1.01
Akaike	3.89	1.95	1.37	1.07

Table 5c MSE of estimates of first parameter of Θ_o for experiment 3

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	3.85	1.69	1.23	.80
Likelihood Ratio 10 %	3.90	1.69	1.15	.78
Chow	3.91	1.59	1.08	.72
DEIC	4.00	1.57	1.01	.69
IEIC	3.94	1.59	1.08	.72
Akaike	3.95	1.68	1.12	.74

Table 5d MSE of estimates of first parameter of Θ_o for experiment 4

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	3.66	1.54	.89	.61
Likelihood Ratio 10 %	3.47	1.52	.85	.61
Chow	3.11	1.40	.85	.61
DEIC	3.03	1.38	.85	.61
IEIC	3.11	1.40	.85	.61
Akaike	3.35	1.46	.85	.61

Table 5e **MSE of estimates of first parameter of Θ_o for experiment 5**

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	1.81	.73	.54	.40
Likelihood Ratio 10 %	1.82	.72	.54	.40
Chow	1.82	.71	.53	.40
DEIC	1.82	.71	.53	.40
IEIC	1.82	.71	.53	.40
Akaike	1.82	.71	.54	.40

Table 5f **MSE of estimates of first parameter of Θ_o for experiment 6**

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	1.36	.56	.35	.27
Likelihood Ratio 10 %	1.38	.56	.37	.31
Chow	1.62	.64	.45	.36
DEIC	1.73	.70	.48	.36
IEIC	1.61	.64	.45	.36
Akaike	1.48	.55	.39	.32

Table 6a

MSE of prediction of $(\sum_{n=1}^{N} z_n\theta_o)/N$ for experiment 1 • 100

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	9.41	2.73	1.55	.86
Likelihood Ratio 10 %	9.82	2.75	1.48	.79
Chow	9.99	2.71	1.24	.77
DEIC	9.96	2.62	1.21	.75
IEIC	10.00	2.71	1.24	.77
Akaike	9.88	2.74	1.44	.77

Table 6b

MSE of prediction of $(\sum_{n=1}^{N} z_n\theta_o)/N$ for experiment 2 • 100

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	9.94	2.43	1.23	.84
Likelihood Ratio 10 %	10.53	2.56	1.33	.82
Chow	10.73	2.59	1.32	.78
DEIC	11.01	2.68	1.24	.76
IEIC	10.73	2.59	1.32	.78
Akaike	10.59	2.54	1.34	.81

Table 6c

MSE of prediction of $\left(\sum_{n=1}^{N} z_n\theta_o\right)/N$ for experiment 3 • 100

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	14.65	3.52	1.82	.98
Likelihood Ratio 10 %	14.78	3.43	1.77	.95
Chow	15.55	3.35	1.63	.91
DEIC	16.19	3.14	1.60	.90
IEIC	15.58	3.35	1.63	.91
Akaike	14.91	3.40	1.69	.92

Table 6d

MSE of prediction of $\left(\sum_{n=1}^{N} z_n\theta_o\right)/N$ for experiment 4 • 100

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	18.40	3.95	1.63	1.01
Likelihood Ratio 10 %	18.20	3.95	1.61	1.01
Chow	17.62	3.71	1.61	1.01
DEIC	17.55	3.67	1.61	1.01
IEIC	17.62	3.71	1.61	1.01
Akaike	18.09	3.79	1.61	1.01

Table 6e

MSE of prediction of $\left(\sum_{n=1}^{N} z_n\theta_o\right)/N$ for experiment 5 • 100

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	9.84	3.06	1.30	.75
Likelihood Ratio 10 % / Chow	9.8[illegible]	2.9[illegible]	1.1[illegible]	.73
DEIC	9.89	2.41	1.13	.73
IEIC	9.75	2.52	1.12	.73
Akaike	9.98	2.91	1.13	.73

Table 6f

MSE of prediction of $\left(\sum_{n=1}^{N} z_n\theta_o\right)/N$ for experiment 6 • 100

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	5.13	1.18	.50	.40
Likelihood Ratio 10 %	5.45	1.24	.52	.41
Chow	6.22	1.51	.62	.43
DEIC	6.04	1.63	.67	.45
IEIC	6.12	1.49	.62	.43
Akaike	5.70	1.37	.56	.43

Table 7a Percentage rejection of approximate model for experiment 1

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	12	18	37	66
Likelihood Ratio 10 %	19	28	50	79
Chow	41	56	86	93
DEIC	100	100	100	100
IEIC	37	56	86	93
Akaike	23	35	61	87

Table 7b Percentage rejection of approximate model for experiment 2

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	11	9	13	26
Likelihood Ratio 10 %	16	19	27	39
Chow	40	39	44	67
DEIC	100	100	100	100
IEIC	33	39	43	67
Akaike	20	23	32	49

Table 7c Percentage rejection of approximate model for experiment 3

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	13	26	62	86
Likelihood Ratio 10 %	21	44	73	89
Chow	50	66	85	96
DEIC	100	100	100	100
IEIC	47	66	85	96
Akaike	28	49	78	94

Table 7d Percentage rejection of approximate model for experiment 4

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	48	93	99	100
Likelihood Ratio 10 %	59	94	100	100
Chow	82	99	100	
DEIC	100	100	100	100
IEIC	82	99	100	100
Akaike	68	96	100	100

Table 7e Percentage rejection of approximate model for experiment 5

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	23	44	84	98
Likelihood Ratio 10 %	34	56	94	100
Chow	58	85	99	100
DEIC	100	100	100	100
IEIC	54	84	99	100
Akaike	40	65	96	100

Table 7f Percentage rejection of approximate model for experiment 6

Selection Rule	Sample Size			
	50	200	500	1,000
Likelihood Ratio 5 %	6	1	2	7
Likelihood Ratio 10 %	9	4	5	15
Chow	35	28	28	38
DEIC	100	100	100	100
IEIC	28	26	28	38
Akaike	16	6	9	22

between the estimators are swamped by the magnitude of their common biases.

The Information Criterion estimators are derived by using the same asymptotic approximations as the MLE. Therefore, it is not surprising that the poor performance of these asymptotic approximations clouds comparisons between the estimators. Tables 4a-f show the MSE of the various Information Criterion estimators for the approximate models across the experiments. As expected, the DEIC has uniformly lower MSEs, and the difference between the estimators declines as the sample size increases. Chow's estimator is comparable to the IEIC, and Akaike's estimator is inferior to the others.

The purpose of estimating the Information Criterion is to generate model selection procedures. Better estimators will usually yield better model selection criterion. The DEIC, which is generally the best estimator for sample sizes greater than 200, always rejects the smaller model. This selection rule is in sharp contrast to the usual likelihood ratio test rules, and it generally performs better. It could be argued that this study is unfair since all the approximate models are misspecified. In practical applications all models are misspecified, and the experiments run here do cover a realistic range of misspecifications.

Tables 5-7 compare the behavior of the selection rules based on the various Information Criterion estimators. For the sake of comparison, these tables also include the standard likelihood ratio selection rule: reject if the likelihood ratio test statistic for the parameter restrictions on Θ_* is significant at the 5 or 10 percent level.

Tables 5a-f give the MSEs for estimating the first parameter of Θ_o. For each Monte Carlo repetition either $\hat{\Theta}_o$ or $\hat{\Theta}_*$ was used according to which model is selected as correct. Tables 6a-f show the MSEs for predicting the average value of $Z_n\Theta_o$. This quantity is inversely related to the predicted demand for alternative 1, but is much faster to compute. There is sometimes a tradeoff between prediction and estimation accuracy. Comparing Tables 5 and 6, there are cases where the best selection rule for prediction is the worst rule for estimation, and vice versa. Tables 7a-f give the percentage rejection of the approximate model for the various selection rules. As expected, the DEIC always rejects the approximate model. Chow's and the IEIC are the second most likely to reject, followed by Akaike's rule and the likelihood ratio tests.

The only clear conclusion from Tables 5 and 6 is that more study is needed before choosing the "best" prediction criterion. It does appear that the Information Criterion selection rules perform better for the larger sample sizes used in these experiments. Chow's criterion and the IEIC perform reasonably well in most of the experiments. Chow's criterion also has the advantage of being easier to compute. The main exception to these conclusions is experiment 6, where the likelihood ratio test at the 5 percent level dominates all others. This experimental design has the worst behavior of the MLEs (see Tables 2) and also contains the least information. Only when the sample size reaches 1,000 is it possible to reject the hypothesis that $\Theta_o \equiv 0$ using the likelihood ratio test at the 10 percent level.

Finally, many of the experiments described in this section were replicated for the Multinominal Probit model. This model does not allow any of the computational simplifications derived for the ML model in Section 2. Nevertheless, the results were essentially identical to those given here for the ML model.

4 Conclusions

This study has a number of implications for applied econometricians. As usual, the positive implications are very tentative, and the negative implications are much clearer. The usual asymptotic approximation to the sampling behavior of the MLE for ML models can be very poor even in "large" samples. Previous Monte Carlo studies (McFadden (1973) and Domencich and McFadden (1975)) used dichotomous independent variables which yield much better small sample behavior. The designs considered here and in Davidson and MacKinnon (1984) have continuous independent variables, but they are nevertheless small, simple models. Increasing the number of discrete alternative and independent variables will only make the small sample behavior worse for a fixed sample size. Therefore the Monte Carlo results given in Section 3 raise considerable doubts regarding the quality of the estimation and testing procedures in most applied work with Multinominal Logit and Probit models.

This paper also shows that the Information Criterion is a promising theoretical basis for modelling actual econometric practice. The estimation and model selection procedures generated by the Information Criterion appear to be good, computationally simple, alternatives to current standard practices. The more traditional specification testing procedures for qualitative choice models are also easy to compute, and they also suffer from the same small sample problems noted above. Unlike the Information Criterion, these specification tests cannot be used for comparing non-nested models, and they only provide pairwise comparisons for nested models. Much more work needs to be done to determine the best selection procedures for qualitative choice models. Once a particular procedure is chosen it will also be necessary to derive the sampling distribution of the implied sequential estimation procedures.

APPENDIX

Covariance Estimator for MLE $\hat{\beta}$

An estimator for the covariance matrix of $\hat{\beta}$ will be derived following Silvey (1959). Since $\hat{\Theta}$ is a maximum likelihood estimator we have

$$N^{-1}\frac{\partial L(\Theta_o)}{\partial\Theta} + \left[N^{-1}\frac{\partial^2 L(\Theta_o)}{\partial\Theta\partial\Theta'} + o(1)\right](\hat{\Theta}-\Theta_o) = 0 \qquad (A.1)$$

Similarly $\partial L(\hat{\Theta}_*)/\partial\Theta$ can be expanded about Θ_* to yield

$$N^{-1}\frac{\partial L(\Theta_*)}{\partial\Theta} + \left[N^{-1}\frac{\partial^2 L(\Theta_*)}{\partial\Theta\partial\Theta'} + o(1)\right](\hat{\Theta}_*-\Theta_*) + H\hat{\lambda} = 0 \quad (A.2)$$

($\hat{\lambda}$ is the MLE of the lagrange multiplier for the constraint in Eq. (2.3).)

Subtraction of the first order conditions for the program in Eq. (2.2) from (A.2) yields

$$\begin{bmatrix} -N^{-1}\dfrac{\partial^2 L(\Theta_o)}{\partial\Theta\partial\Theta'} + o(1) & 0 & 0 \\ 0 & -N^{-1}\dfrac{\partial^2 L(\Theta_*)}{\partial\Theta\partial\Theta'} + o(1) & -H \\ 0 & -H' & 0 \end{bmatrix} \begin{bmatrix} \hat{\Theta}-\Theta_o \\ \hat{\Theta}_*-\Theta_* \\ \hat{\lambda}-\lambda \end{bmatrix} =$$

$$= \begin{bmatrix} N^{-1}\dfrac{\partial L(\Theta_o)}{\partial\Theta} \\ N^{-1}\dfrac{\partial L(\Theta_*)}{\partial\Theta} - N^{-1}\dfrac{\partial EL(\Theta_*)}{\partial\Theta} \\ 0 \end{bmatrix} \tag{A.3}$$

The asymptotic distribution of $N^{1/2}$ times the right hand side of (A.3) is Normal by the Central Limit Theorem with mean zero and covariance matrix

$$\lim_{N\to\infty} N^{-1} \begin{bmatrix} J(\Theta_o,\Theta_o) & K & 0 \\ K' & V & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{A.4}$$

where $K_{ij} = \mathrm{Cov}\left[\dfrac{\partial L(\Theta_o)}{\partial\Theta_i}, \dfrac{\partial L(\Theta_*)}{\partial\Theta_j}\right]$

and $V = \mathrm{Cov}\left[\dfrac{\partial L(\Theta_*)}{\partial\Theta}\right]$.

Solving (A.3) yields an asymptotic Normal distribution for $N^{1/2}(\hat{\beta}-\beta_o)$ and $N^{1/2}(\hat{\lambda}-\lambda)$ with mean 0 and covariance matrix

$$\lim_{N\to\infty} \begin{bmatrix} NJ^{-1}(\Theta_o,\Theta_o) & J^{-1}(\Theta_o,\Theta_o)KS' & J^{-1}(\Theta_o,\Theta_o)KQ \\ SK'J^{-1}(\Theta_o,\Theta_o) & N^{-1}SVS' & N^{-1}SVQ \\ Q'K'J^{-1}(\Theta_o,\Theta_o) & N^{-1}Q'VS' & N^{-1}Q'VQ \end{bmatrix} \qquad \text{(A.5a)}$$

where

$$\begin{bmatrix} S & Q \\ Q' & R \end{bmatrix} = \begin{bmatrix} N^{-1}J(\Theta_*,\Theta_o) & -H \\ -H' & 0 \end{bmatrix}^{-1} = \qquad \text{(A.5b)}$$

$$= \begin{bmatrix} NJ^{-1}(\Theta_*,\Theta_o)-NJ^{-1}(\Theta_*,\Theta_o)H(H'J^{-1}(\Theta_*,\Theta_o)H)^{-1}H'J^{-1}(\Theta_*,\Theta_o) & -J^{-1}(\Theta_*,\Theta_o)H(H'J(\Theta_*,\Theta_o)H)^{-1} \\ -(H'J^{-1}(\Theta_*,\Theta_o)H)^{-1}H'J^{-1}(\Theta_*,\Theta_o) & -N^{-1}(H'J^{-1}(\Theta_*,\Theta_o)H)^{-1} \end{bmatrix}$$

Estimates of $K, V, J(\Theta_o,\Theta_o)$ and $J(\Theta_*,\Theta_o)$ for qualitative choice models can be obtained by evaluating the following formulas at $\hat{\Theta}$ and $\hat{\Theta}_*$:

$$K = \sum_n R_n \Big(\sum_i \frac{\partial \log P_{in}(\Theta_o)}{\partial \Theta} P_{in}(\Theta_o) \frac{\partial \log P_{in}(\Theta_*)}{\partial \Theta'} \Big) -$$

$$- \Big(\sum_j P_{jn}(\Theta_o) \frac{\partial \log P_{jn}(\Theta_o)}{\partial \Theta} \Big) \Big(\sum_k P_{kn}(\Theta_o) \frac{\partial \log P_{kn}(\Theta_*)}{\partial \Theta'} \Big) =$$

$$= \sum_n R_n \sum_i \frac{\partial \log P_{in}(\Theta_o)}{\partial \Theta} P_{in}(\Theta_o) \frac{\partial \log P_{in}(\Theta_*)}{\partial \Theta'} \qquad \text{(A.6a)}$$

$$\Big(\text{since } E\Big(\frac{\partial L(\Theta_o)}{\partial \Theta}\Big) \text{ implies } \sum_j P_{jn}(\Theta_o) \frac{\partial \log P_{jn}(\Theta_o)}{\partial \Theta} = 0\Big)$$

$$V(\Theta_*) = \sum_n R_n \Big(\sum_i \frac{\partial \log P_{in}(\Theta_*)}{\partial \Theta} P_{in}(\Theta_o) \frac{\partial \log P_{in}(\Theta_*)}{\partial \Theta'} -$$

$$- \sum_j P_{jn}(\Theta_o) \frac{\partial \log P_{jn}(\Theta_*)}{\partial \Theta} \Big) \Big(\sum_k P_{kn}(\Theta_o) \frac{\partial \log P_{kn}(\Theta_*)}{\partial \Theta'} \Big) \Big)$$

$$\text{(A.6b)}$$

$$J(\Theta_*, \Theta_o) = - \sum_n R_n \sum_i P_{in}(\Theta_o) \frac{\partial^2 \log P_{in}(\Theta_*)}{\partial \Theta \partial \Theta'} \qquad \text{(A.6c)}$$

where $R_n = \sum_i Y_{in}$.

These formulas can be easily derived from

$$E\ Y_{in} = R_n P_{in}(\Theta_o) \qquad \text{and}$$

$$E\ Y_{in}\ Y_{jn} = \delta_{ij} R_n P_{in} + R_n(R_n - 1)P_{in}P_{jn}$$

$$\delta_{ij} \begin{cases} = 1 \text{ if } i = j \\ = 0 \text{ if } i \neq j \end{cases}$$

The hessian of Q can be evaluated from:

$$\frac{\partial^2 Q}{\partial\Theta\partial\Theta'} = \sum_n R_n \sum_i \frac{\partial^2 P_{in}(\Theta_o)}{\partial\Theta\partial\Theta'} \log P_{in}(\Theta_*) \qquad \text{(A.7a)}$$

$$\frac{\partial^2 Q}{\partial\Theta\partial\Theta'_*} = K \qquad \text{(from Eq. 2.21a)} \qquad \text{(A.7b)}$$

$$\frac{\partial^2 Q}{\partial\Theta_*\partial\Theta'_*} = -\ J(\Theta_*, \Theta_o) \qquad \text{(A.7c)}$$

References

Akaike, H., 1973, Information Theory and an Extension of the Maximum Likelihood Principle, in B.N. Petrov and F. Csåki, eds., Proceedings of the 2nd International Symposium on Information Theory, (Akademi Kiadó, Budapest).

Chow, G.C., 1981, A Comparison of the Information and Posterior Probability Criteria for Model Selection, Journal of Econometrics 16, 21-33.

Chow, G.C., 1982, Note on Maximum Likelihood Estimation of Misspecified Models, Econometric Research Program Memorandum No. 300, (Princeton University).

Domencich, T. and D. McFadden, 1975, Urban Travel Demand, A Behavioral Analysis, (North-Holland, Amsterdam).

Davidson, R. and J.G. MacKinnon, 1984, Convenient Specification Tests for Logit and Probit Models, Journal of Econometrics 25, 241-262.

Kullback, S., 1959, Information Theory and Statistics, (John Wiley and Sons, New York).

Kullback, S. and R.A. Leibler, 1951, On Information and Sufficiency, Annals of Mathematical Statistics 22, 79-86.

McFadden, D., 1973, Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zarembka, ed., Frontiers in Econometrics, (Academic Press, New York).

Silvey, S.D., 1959, The Lagrangian Multiplier Test, Annals of Mathematical Statistics 30, 389-407.

White, H., 1982, Maximum Likelihood Estimation of Misspecified Models, Econometrica 50, 1-25.

White, H., 1983, Corrigendum ..., Econometrica 51, 513-515.

Zellner, A. and P.E. Rossi, 1984, Bayesian Analysis of Dichotomous Quantal Response Models, Journal of Econometrics 25, 365-394.

WORKING PAPERS (Missing numbers indicate publication elsewhere)

1976

1. Corporate and Personal Taxation and the Growing Firm
   by Ulf Jakobsson

8. Estimation and Analysis with a WDI Production Function
   by Göran Eriksson, Ulf Jakobsson and Leif Jansson

1977

15. Pay Differentials between Government and Private Sector Employees in Sweden
    by Siv Gustafsson

1980

25. On Unexplained Price Differences
    by Bo Axell

34. Imperfect Information Equilibrium, Existence, Configuration and Stability
    by Bo Axell

1981

36. Energi, stabilitet och tillväxt i svensk ekonomi (Energy, Stability and Growth in the Swedish Economy)
    by Bengt-Christer Ysander

38. Utiliy in Local Government Budgeting
    by Bengt-Christer Ysander

40. Wage Earners Funds and Rational Expectations
    by Bo Axell

42. The Structure of the ISAC Model
    by Leif Jansson, Tomas Nordström and Bengt-Christer Ysander

43. An Econometric Model of Local Government and Budgeting
    by Bengt-Christer Ysander

44. Local Authorities, Economic Stability and the Efficiency of Fiscal Policy
    by Tomas Nordström and Bengt-Christer Ysander

45. Growth, Exit and Entry of Firms
    by Göran Eriksson

56. Central Control of the Local Government Sector in Sweden
by Richard Murray

59. Longitudinal Lessons from the Panel Study of Income Dynamics
by Greg J. Duncan and James N. Morgan

1982

61. Var står den nationalekonomiska centralteorin idag?
av Bo Axell

63. General Search Market Equilibrium
by James W. Albrecht and Bo Axell
General Equilibrium without an Auctioneer
by James W. Albrecht, Bo Axell and Harald Lang

64. The Structure and Working of the ISAC Model
by Leif Jansson, Thomas Nordström and Bengt-Christer Ysander

65. Comparative Advantage and Development Policy Twenty Years Later
by Anne O. Krueger

67. Computable Multi-Country Models of Production and Trade
by James M. Henderson

69. Relative Competitiveness of Foreign Subsidiary Operations of a Multinational Company 1962-77
by Anders Grufman

71. Technology, Pricing and Investment in Telecommunications
by Tomas Pousette

72. The Micro Initialization of MOSES
by James W Albrecht and Thomas Lindberg

75. The MOSES Manual
by Fredrik Bergholm

76. Differential Patterns of Unemployment in Sweden
by Linda Leighton and Siv Gustafsson

77. Household Market and a Nonmarket Activities (HUS) - A Pilot Study
by Anders Klevmarken

1983

79. Energy Prices, Industrial Structure and Choice of Technology; An International Comparison with Special Emphasis on the Cement Industry
by Bo Carlsson

81. ELIAS - A Model of Multisectoral Economic Growth in a Small Open Economy
by Lars Bergman

84. Monopoly and Allocative Efficiency with Stochastic Demand
by Tomas Pousette

86. The Micro (Firm) Foundations of Industrial Policy
by Gunnar Eliasson

87. Excessive Government Spending in the U.S.: Facts and Theories
by Edward M. Gramlich

88. Control of Local Authority Expenditure - The Use of Cash Limits
by Noel Hepworth

89. Fiscal Containment and Local Government Finance in the U.K.
by Peter Jackson

90. Fiscal Limitations: An Assessment of the U.S. Experience
by Wallace E. Oates

91. Pricing and Privatization of Public Services
by George E. Peterson

94. Job Mobility and Wage Growth: A Study of Selection Rules and Rewards
by Bertil Holmlund

96. The Machine Tool Industry - Problems and Prospects in an International Perspective
by Bo Carlsson

97. The Development and Use of Machine Tools in Historical Perspective
by Bo Carlsson

99. An Equilibrium Model of Search Unemployment
by James W. Albrecht and Bo Axell

100. Quit Behavior under Imperfect Information: Searching, Moving, Learning
by Bertil Holmlund and Harald Lang

102. Collecting Data for Micro Analysis.
Experiences from the Hus-Pilot Study
- by N. Anders Klevmarken

103. The Swedish Micro-to-Macro Model
- Idea, Design and Application
by Gunnar Eliasson

105. Estimation of Wage Gains and Welfare Gains from
Self-Selection Models
by Anders Björklund and Robert Moffitt

106. Public Policy Evaluation in Sweden
by Bengt-Christer Ysander

108. Entry, Industry Growth and the Microdynamics
of Industry Supply
by John C. Hause and Gunnar Du Rietz

112. Schumpeterian Competition in Alternative Technological
Regimes
by Sidney Winter

113. The Industrial Finance Systems; Europe, U.S. and Japan
by Tad Rybczynski

117. Micro Heterogeneity of Firms and the Stability of Industrial
Growth
by Gunnar Eliasson

118. The Initialization Process - The Moses Manual, Part 2
by Fredrik Bergholm

1984

119. Forestalling the Demise of Empirical Economics: The Role
of Microdata in Labor Economics Research
by Frank Stafford

120. Unemployment and Mental Health - Some Evidence from
Panel Data
by Anders Björklund

121. Comparative Keynesian Dynamics
by Richard H. Day

122. The Firm and Financial Markets in the Swedish
Micro-to-Macro Model (MOSES)
- Theory, Model and Verification
by Gunnar Eliasson

123. What Can Be Learned from the Chrysler Experience?
by Frank Stafford

125. Estimates of Bounds for Export Price Elasticities in an Underidentified Simultaneous Two Equation System
by Eva Christina Horwitz

126. An Easy Fixup of the Flaw in Lucas' "Neutrality of Money"
by Harald Lang

127. Comment on Feldstein and Poterba's "Unemployment Insurance and Reservation Wages"
by Harald Lang

128. Information Criterion and Estimation of Misspecified Qualitative Choice Models
by David Brownstone