Herbert Nagel and Reinhold Hatzinger

Diagnostics in some Discrete Choice Models

Working Paper

Diagnostics in some Discrete Choice Models

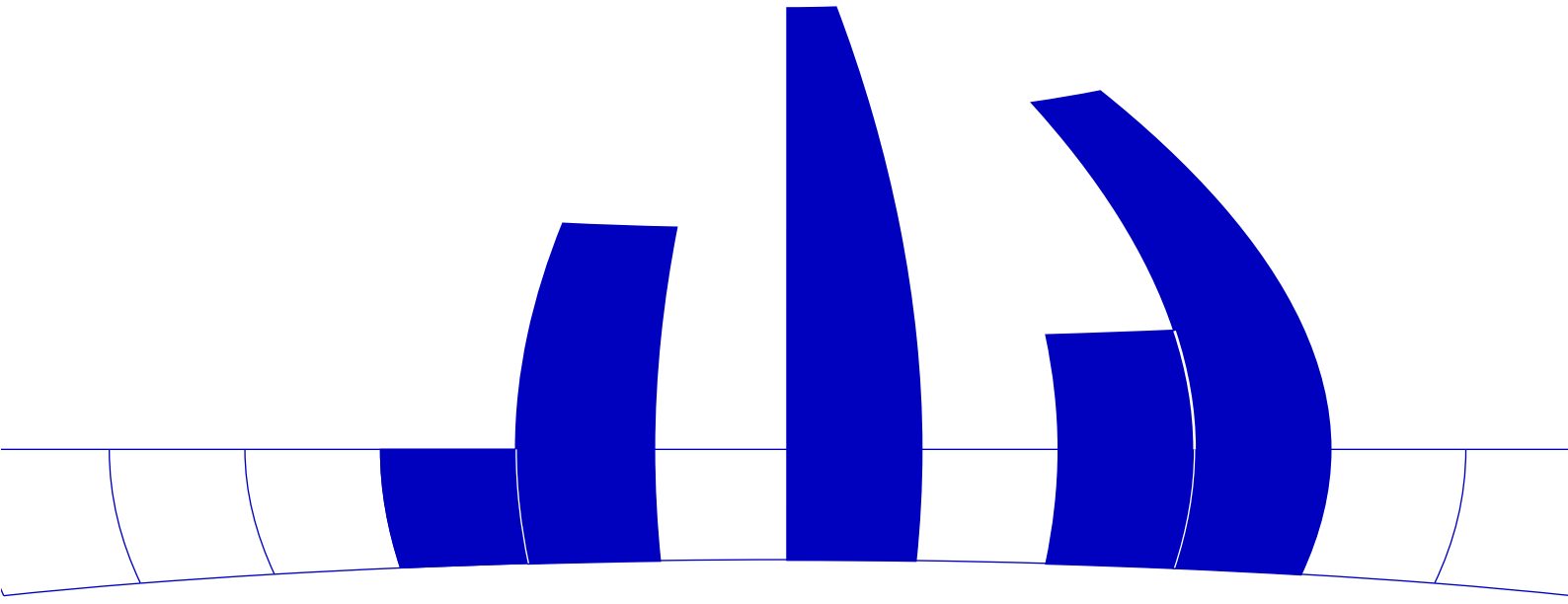
Herbert Nagel, Reinhold Hatzinger

Institut für Statistik
Wirtschaftsuniversität Wien

Forschungsberichte

Bericht 7
1990

http://statmath.wu-wien.ac.at/



Diagnostics in some discrete choice models

By Herbert Nagel and Reinhold Hatzinger

Abstract: Discrete choice models form a class of models widely used in econometrics for modelling the individual choice from a finite set of alternatives. The most widely used model is the multinomial logit model, implicitly assuming independence of irrelevant alternatives. A generalization is the nested multinomial logit model, relaxing this strong assumption. Viewing both models as nonlinear regression models a set of diagnostics is derived. This includes a hat matrix, measures of leverage, influence and residuals and an approximation to the parameters for case deletion. In an example for the multinomial logit model a good performance of these diagnostics is observed and the parameter approximation by the proposed formula is better than a one step Newton-Raphson procedure. In an example for the nested logit model a constructed outlier with high influence is revealed by the measures of leverage and residual, but the parameter approximation is insufficient.

Keywords: Discrete choice model, Multinomial logit model, Nested multinomial logit model, Diagnostics, Residual, Leverage, Influence, Hat matrix

1. Discrete choice models

An approach to model the choice from a set of mutually exclusive and collectively exhaustive alternatives are discrete choice models. Using the principle of utility maximisation a decision maker is modelled as selecting the alternative with the highest utility among those available to him at the time the choice is made, often called the choice set of that decision. To make the model operational a parameterized utility function connects observable explanatory variables with unknown parameters to be estimated from a sample of observed choices. There are different reasons why this utility function does not exactly describe the utility function of the decision maker. The concept of random utility considers the true utilities as random variables, so the probability that an alternative is chosen is defined as the probability that it has the greatest utility among the available alternatives. A discrete choice model (DCM) is completely specified by

- the separation of the total utility into its deterministic and random parts (mostly an additive separation is chosen)
- the specification of the deterministic part (linear in the parameters)
- the specification of the random part.

Let C_n denote the choice set for decision n (it is not necessary that all choice sets consist of the same set of alternatives, i.e. not all decision makers must have the same set of alternatives available), let U_n^i, D_n^i and ε_n^i be the utility, the deterministic and the random part of the utility of alternative i for decison maker $n, i = 0, \ldots, J_n$, then the probability, that the choice of n is for alternative i

$$p_n^i = prob(U_n^i \geq U_n^j; \forall j \in C_n) = prob(\varepsilon_n^j \leq D_n^i - D_n^j + \varepsilon_n^i; \forall j \in C_n) \qquad (1)$$

If f_n denotes the joint density function of the random terms, the probability for alternative 0 is given by

$$p_n^0 = \int_{-\infty}^{\infty} \int_{-\infty}^{D_n^0 - D_n^1 + \varepsilon_n^0} .. \int_{-\infty}^{D_n^0 - D_n^{J_n} + \varepsilon_n^0} f_n(\varepsilon_n^0, \varepsilon_n^1, .., \varepsilon_n^{J_n}) d\varepsilon_n^{J_n} .. d\varepsilon_n^1 d\varepsilon_n^0 \qquad (2)$$

This formula is adapted easily for alternatives other than $i = 0$, but computation of these probabilities often requires cumbersome integration. As shown in Domencich and McFadden (1975), the assumption that the disturbances are independently and identically distributed with the Type I extreme–value distribution, $F(y) = \exp(-\exp(-y))$, expression (2) reduces to

$$p_n^i = \frac{\exp(D_n^i)}{\sum_{j \in C_n} \exp(D_n^j)} = \frac{\exp(\sum_k x_{nk}^i \beta_k)}{\sum_{j \in C_n} \exp(\sum_k x_{nk}^j \beta_k)} \qquad (3)$$

This model is called the multinomial logit discrete choice model (MNL model) and is the most widely used discrete choice model. The aim of this model is to consider the effects of choice characteristics on the determinants of choice probabilities. Compared to the multinomial model in categorical regression the choice probabilities depend on individual characteristics only. If $J_n = J$ for all n, i.e. the same number of alternatives is available in all decisions, there is a $1-1$ relationship between both models (see Maddala, 1983, p.42).

A great advantage of the MNL model is its computational feasibility but *independence of irrelevant alternatives* (IIA – property) is assumed implicitly. Since

$$\frac{p_n^i}{p_n^j} = \frac{\exp(D_n^i)}{\exp(D_n^j)} = \frac{\exp(\sum_k x_{nk}^i \beta_k)}{\exp(\sum_k x_{nk}^j \beta_k)} \qquad (4)$$

the ratio of the choice probabilities for two alternatives does only depend on the characteristics of these two alternatives. (Note that in categorical regression this ratio is $\exp(\sum_k x_{nk}\beta_{ik})/\exp(\sum_k x_{nk}\beta_{jk})$.)

A generalization of the MNL model is the nested multinomial logit model (NMNL) where the IIA assumption is relaxed. The choice set for individual n is a subset of the Cartesian product of F and S (F is the set of alternatives for the first choice

and S is the set of alternatives for the second choice) $C_n = F \times S \setminus C_n^*$, where C_n^* is the set of elements of the Cartesian product infeasible for individual n.

The random utility of the alternative $(f, s) \in C_n$ for individual n is given by

$$U_n^{f,s} = D_n^{f,s} + \varepsilon_n^{f,s} + \varepsilon_n^f + \varepsilon_n^s \tag{5}$$

where ε_n^f is the random part attributable to $f \in F$, ε_n^s is the random part attributable to $s \in S$ and $\varepsilon_n^{f,s}$ the remaining part of the random variation. Assuming $\varepsilon_n^s = 0$ and $\varepsilon_n^{f,s}$ is iid Type I extreme–value distributed with

$$p_n^{s|f} = \frac{\exp(D_n^{f,s})}{\sum_{s' \in S_n^f} \exp(D_n^{f,s'})} \tag{6}$$

where $S_n^f = \{s' \in S \mid (f, s') \in C_n\}$. Moreover, if ε_n^f and $\varepsilon_n^{f,s}$ are independent for all $f \in F_n, F_n = \{f \mid \exists s \text{ so that } (f, s) \in C_n\}$, $s \in S_n, S_n = \{s \mid \exists f \text{ so that } (f, s) \in C_n\}$ and ε_n^f is distributed so that $\max_{s \in S^f} U^{f,s}$ is Type I extreme–value distributed, $F(y) = \exp(-\exp(-\mu^f y))$, with scale parameter μ^f then

$$p_n^f = \frac{\exp(\mu^f I_n^f)}{\sum_{f' \in F_n} \exp(\mu^f I_n^{f'})} \tag{7}$$

where I_n^f is denotes the *inclusive value* of f,

$$I_n^f = \ln \sum_{s \in S_n^f} \exp(D_n^{f,s}) \tag{8}$$

Thus

$$p_n^{f,s} = p_n^{s|f} \cdot p_n^f = \frac{\exp(D_n^{f,s})}{\sum_{s' \in S_n^f} \exp(D_n^{f,s'})} \frac{\exp(\mu^f I_n^f)}{\sum_{f' \in F_n} \exp(\mu^f I_n^{f'})} \tag{9}$$

These results are easily extended to cases with more nesting levels.

For parameter estimation usually maximum likelihood methods are applied. Under weak conditions the likelihood function of the MNL model is globally concave (Amemiya, 1985). A similar result for the NMNL model does not generally hold.

2. Diagnostics

A hat matrix in multinomial DCMs. In the context of DCM few is known about model diagnostics (ie. outlier detection, influential points, etc.) except in the case of the multinomial logit model in categorical regression equivalent to the MNL model with the same number of alternatives in all decisions (see Lesaffre and Albert, 1989). The aim of this paper is to present a method how deletion of decisions affects parameter estimates. In linear models the hat matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (Hoaglin and Welsch, 1978) is used to detect critical observations in the design space. Pregibon (1981) gave an approximation to the change in the maximum likelihood estimate $\hat{\beta}$ on deletion of single observations in logistic regression models. A geometric construction of a "hat matrix" $\tilde{\mathbf{H}}$ in nonlinear regression (Moolgavkar, Lustbader and Venzon, 1984) yields a generalization including the normal hat matrix and Pregibon's hat matrix for the exponential family as special cases. This matrix is given as

$$\tilde{\mathbf{H}} = \hat{\mathbf{V}}^{-1/2}\mathbf{J}_{\hat{\beta}}(\mathbf{J}'_{\hat{\beta}}\hat{\mathbf{V}}^{-1}\mathbf{J}_{\hat{\beta}})^{-1}\mathbf{J}'_{\hat{\beta}}\hat{\mathbf{V}}^{-1/2} \tag{10}$$

where $\hat{\mathbf{V}}$ is the estimated expected information matrix and $\mathbf{J}_{\hat{\beta}}$, the Jacobian of the expected values $f(\beta)$ at $\hat{\beta}$.

In multinomial DCMs $\hat{\mathbf{V}}^{-1}$ is block diagonal with elements $\hat{p}_{ni}(\delta_{ij} - \hat{p}_{nj})$, the covariance structure for decision n and alternatives i and j, $\delta_{ij} = 1$ if $i \neq j$, 0 otherwise. The structure of $\mathbf{J}_{\hat{\beta}}$ depends on the functional relationship between θ, the vector of natural parameters with elements $\theta_n^i = \ln(p_n^i/p_n^0)$ of the multinomial distribution in the exponential familiy and β, the vector of parameters for the explanatory variables, $\theta = h(\beta)$. In the MNL model

$$\theta_n^i = \ln(p_n^i/p_n^0) = D_n^i - D_n^0 = \sum_k (x_{nk}^i - x_{nk}^0)\beta_k \tag{11}$$

Thus $h(\beta) = \mathbf{X}\beta$, where $\mathbf{X}$ is the design matrix adjusted for a reference alternative to avoid overparameterization and $\mathbf{J}_{\hat{\beta}} = \hat{\mathbf{V}}\mathbf{X}$.
In the NMNL model

$$\theta_n^{f,s} = D_n^{f,s} - D_n^{f_0,s_0} - \sum_{s' \in S_n^f} D_n^{f,s'} + \sum_{s' \in S_n^{f_0}} D_n^{f_0,s'} + \mu^f I^f - \mu^f I^{f_0} \tag{12}$$

and $h(\beta)$ is not linear in β. $\mathbf{X}$ in $h(\beta) = \mathbf{X}\beta$ has therefore to be replaced by $\mathbf{X}_{\hat{\beta}} = \partial\theta/\partial\beta$. Differentiating (12) with respect to β_k and μ^f, respectively, yields

$$\begin{aligned} \frac{\partial \theta_n^{f,s}}{\partial \beta_k} &= x_{nk}^{f,s} - x_{nk}^{f_0,s_0} - \sum_{s' \in S_n^f} x_{nk}^{f,s'} + \sum_{s' \in S_n^{f_0}} x_{nk}^{f_0,s'} + \\ &+ \mu^f \frac{\sum_{s' \in S_n^f} x_{nk}^{f,s'} \exp(D_n^{f,s'})}{\sum_{s' \in S_n^f} \exp(D_n^{f,s'})} - \mu^f \frac{\sum_{s' \in S_n^{f_0}} x_{nk}^{f_0,s'} \exp(D_n^{f_0,s'})}{\sum_{s' \in S_n^{f_0}} \exp(D_n^{f_0,s'})} \quad , \end{aligned} \tag{13}$$

$$\frac{\partial \theta_n^{f,s}}{\partial \mu^f} = \ln \frac{\sum_{s' \in S_n^f} \exp(D_n^{f,s'})}{\sum_{s' \in S_n^{f_0}} \exp(D_n^{f_0,s'})} \tag{14}$$

Substitution into (10) leads to

$$\mathbf{H}^* = \hat{\mathbf{V}}^{1/2} \mathbf{X}_{\hat{\beta}} (\mathbf{X}'_{\hat{\beta}} \hat{\mathbf{V}} \mathbf{X}_{\hat{\beta}})^{-1} \mathbf{X}'_{\hat{\beta}} \hat{\mathbf{V}}^{1/2} \quad , \tag{15}$$

where $\mathbf{X}_{\hat{\beta}} = \mathbf{X}$ in the simpler case of the MNL model. Using (15) a measure of leverage can be obtained by the determinant of $\mathbf{M}_n = \mathbf{I} - \mathbf{H}_n^*$, where $\mathbf{H}_n^*$ is restricted to the n^{th} decision. Small values of $\det(\mathbf{M}_n)$ might imply large changes in parameter estimates $\hat{\beta}$ when the n^{th} decision is deleted. Note that $\sum_n (1 - \det(\mathbf{M}_n))$ is not necessarily equal to the number of parameters in contrary to the case of the linear model.

Some other diagnostics. To determine the actual influence of decision n an analogue to the Cook distance (Cook,1977) is given by

$$d_n = \left((\hat{\beta} - \hat{\beta}_{(n)})' (\mathbf{X}'_{\hat{\beta}} \hat{\mathbf{V}} \mathbf{X}_{\hat{\beta}}) (\hat{\beta} - \hat{\beta}_{(n)}) \right)^{1/2} \quad , \tag{16}$$

where $\hat{\beta}_{(n)}$ is the vector of parameters $\hat{\beta}$ estimated with decision n deleted. To avoid the computational expense of fully iterating until convergence for all decisions approximations for $\hat{\beta}_{(n)}$ are required. A proposal given by Moolgavkar, Lustbader and Venzon (1984) using deletion algebra in the weighted linear least squares problem $z = \mathbf{J}_{\hat{\beta}} \beta + r = \hat{\mathbf{V}} \mathbf{X}_{\hat{\beta}} \beta + r$ with $E(r) = 0$ and $var(r) = \hat{\mathbf{V}}$ is

$$\hat{\beta}_{(n)} \approx \hat{\beta} - (\mathbf{J}'_{\hat{\beta}} \hat{\mathbf{V}}^{-1} \mathbf{J}_{\hat{\beta}})^{-1} \mathbf{J}'_{\hat{\beta}n} \hat{\mathbf{V}}_n^{-1/2} (\mathbf{I} - \tilde{\mathbf{H}}_n)^{-1} \hat{\mathbf{V}}_n^{-1/2} \hat{r}_n \tag{17}$$

with $\mathbf{J}'_{\hat{\beta}} \hat{\mathbf{V}}^{-1} \mathbf{J}_{\hat{\beta}}$, the expected Fisher information at $\hat{\beta}$ and $\hat{r}_n = y_n - \hat{p}_n^i$, the ordinary residuals of decision n. A different approach is to estimate $\beta_{(n)}$ by a one-step Newton Raphson approximation with $\hat{\beta}$ as starting values.

Familiar residual plots can be obtained by using the quadratic form

$$| \hat{\varepsilon}_n |^2 = \hat{r}_n \hat{\mathbf{V}}_n^{-1} \hat{r}_n = (y_n - \hat{p}_n^i)' \hat{\mathbf{V}}_n^{-1} (y_n - \hat{p}_n^i) \tag{18}$$

with the obvious difference to usual residuals that the $| \hat{\varepsilon}_n |^2$ are always positive. Plotting the residuals (18) against the leverages $(1 - \det(\mathbf{M}_n))$ might provide additional information about the impact of decision n.

3. Two Examples

Example 1: MNL model for transport modal choice. A study on the modal split in the region of Vienna (Otruba, Gampe 1986), including only trips to work with a maximal choice set of the four alternatives: a_1=motorcycle,a_2=private car, a_3=taxi, a_4=public transport led to a final model for 308 decisions with nine variables, where only the estimates for the most important covariates are given below

Variable	estimate	s.e.	t value
Time a_1 to a_3	-0.2692	0.0351	-7.6617
Time a_4	-0.1195	0.0208	-5.7465
Costs a_1 to a_3	-0.0351	0.0122	-2.8788
Costs a_4	-0.1577	0.0456	-3.4591

The log likelihood for the null model was $l(0) = -227.56$ with $df = 343$ and for the final model $l(\hat{\beta}) = -95.89$ with $df = 334$. Figure 1a shows a plot of the residuals versus leverages $1 - \det(\mathbf{M}_n)$.

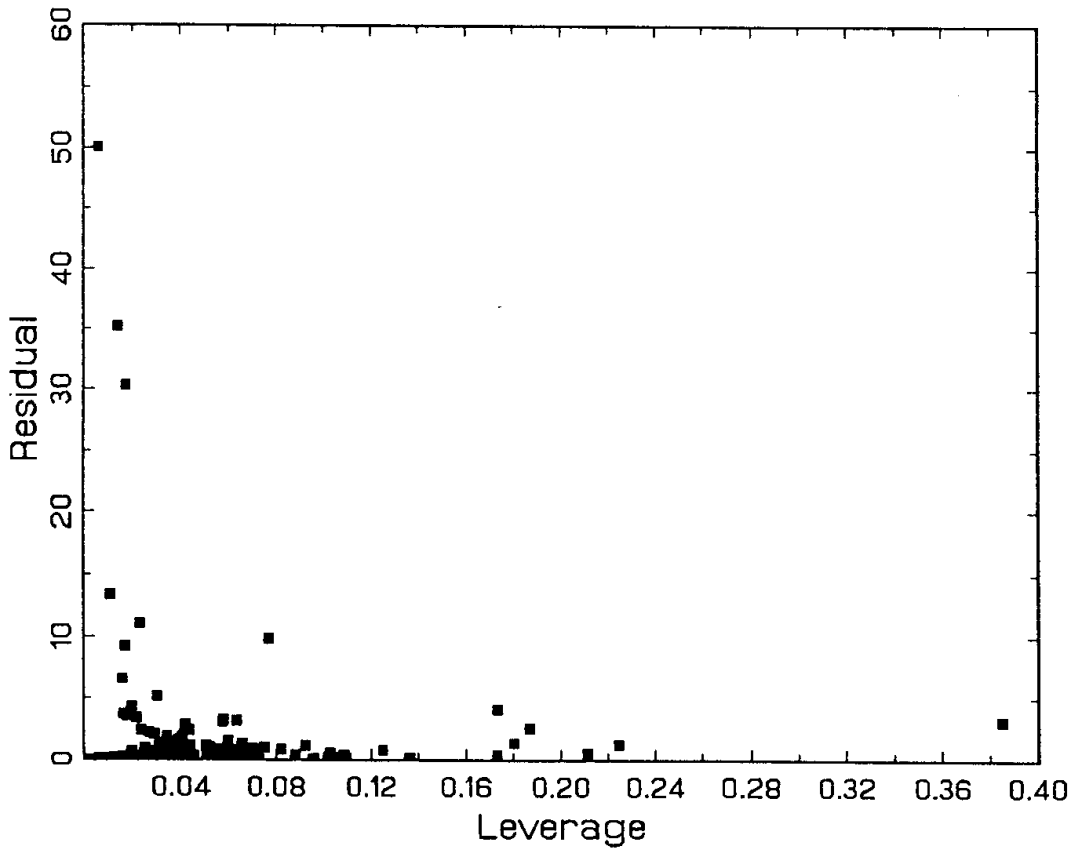


Fig 1a: Residuals versus leverages

Decisions with an unusual combination of covariates will have a value far above the average leverage, which is $8.99/308 = 0.029$. There are some decisions with outstanding leverage values but relatively small residuals and vice versa. One decision has both high residual (≈ 10) and high leverage (≈ 0.08) and should be examined carefully. Figure 1b gives an index plot of the influences, when decision n is deleted. For decisions with high influence (17) approximates the value obtained

by full iteration generally better than the Newton-Raphson one-step estimate (cf. decision 116).

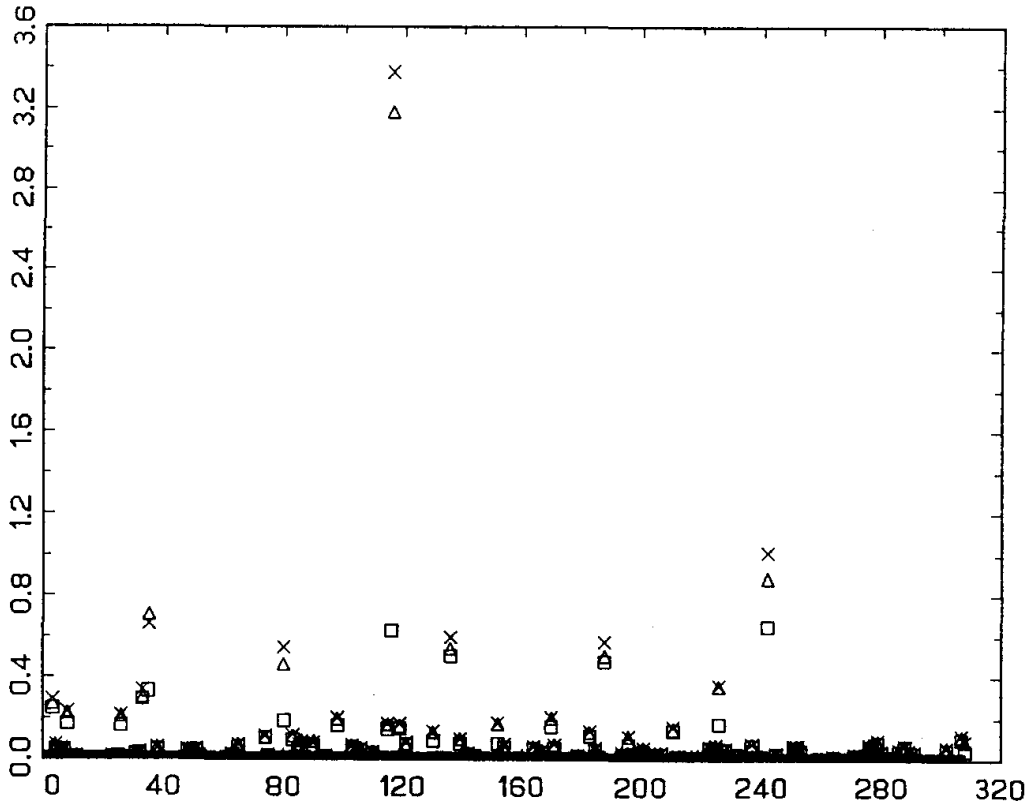


Fig 1b: Influences. The symbols ×, △ and □ denote the values for the full iteration, the approximation (17) and the Newton-Raphson one-step procedure.

Example 2: NMNL model. To demonstrate the diagnostics proposed for the NMNL model a data set of 100 decisions with 2 levels of decisions was constructed. In the first level 2 choices were assumed to be available (f_1, f_2) in the second level 3 alternatives were available given choice 1 of the first level $S^{f_1} = \{s_1^{f_1}, s_2^{f_1}, s_3^{f_1}\}$ and 2 alternatives given choice 2 of the first level $S^{f_2} = \{s_1^{f_2}, s_2^{f_2}\}$. An example for this kind of data structure could be the following: Suppose the simultaneous choice to have dinner at a certain kind of restaurant S and the choice to stay in a city f_1 or go to the countryside f_2 where the third kind of restaurant is not available.

Three covariates were included to specify the deterministic part of the utility, a binary dummy for f_1, a binary dummy included in the deterministic part for $s_3^{f_1}$ only and a continuous covariate for all alternatives. The random part was generated using the Type I extreme-value distribution. An artificial outlier was constructed by changing one decision to an alternative with low $\hat{p}_n^{f_1 s}$. By inspecting Figure 2a this outlier has a high leverage value of ≈ 0.14 and an incredible residual of ≈ 24000. Other decisions with high residuals are below the average leverage of 0.03.

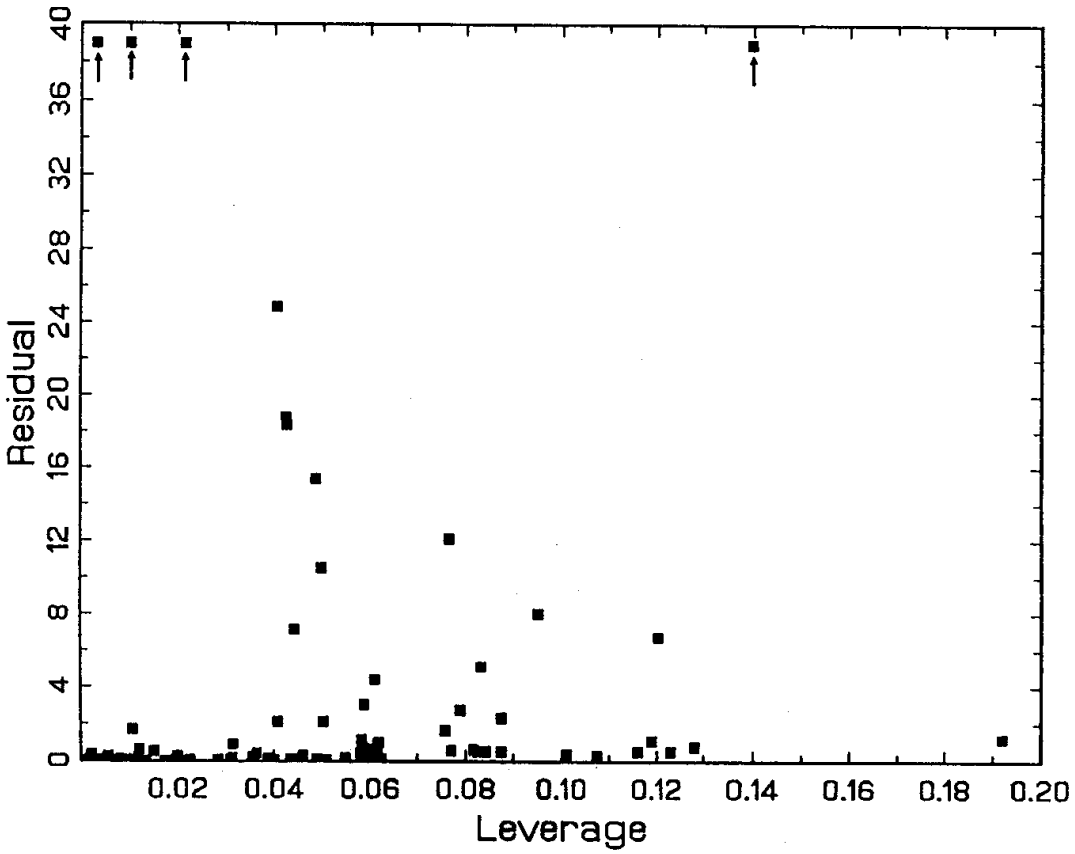


Fig 2a: Residuals versus leverages

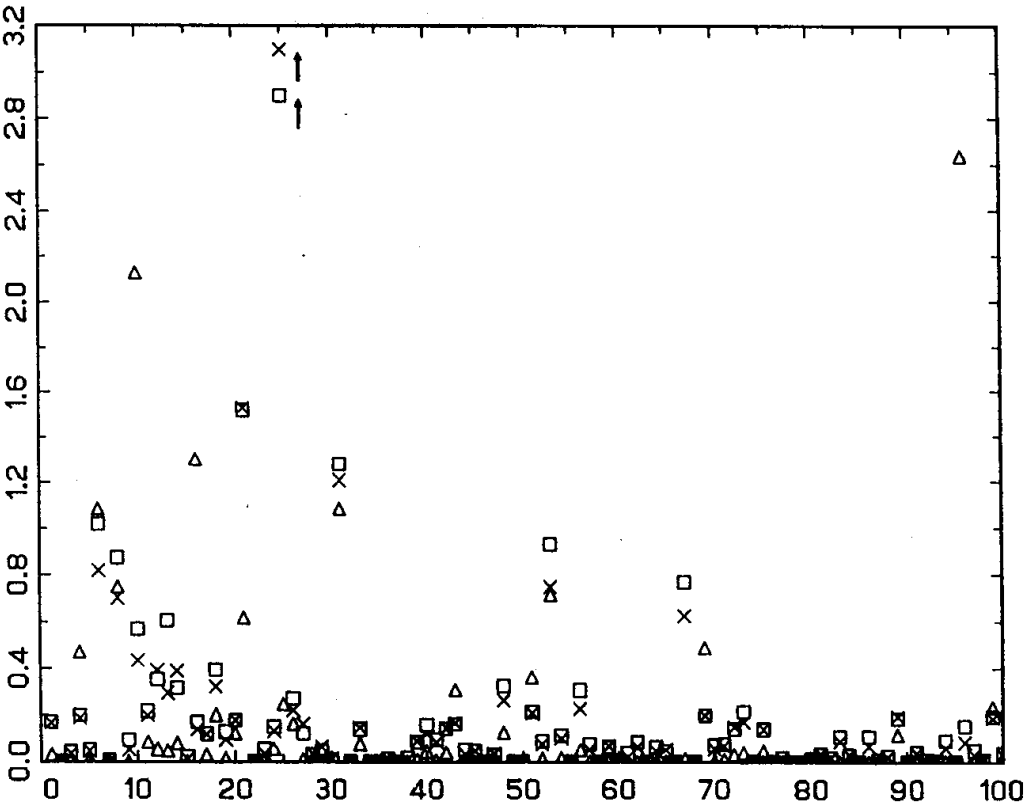


Fig 2b: Influences. The symbols ×, △ and □ denote the values for the full iteration, the approximation (17) and the Newton-Raphson one-step procedure.

Figure 2b gives the index plot for the influences in the constructed data set. Generally the approximation (17) is worse than in the MNL model, for the most influential decision (the constructed outlier) with index 25 the value obtained by full iteration is 141, by the one-step Newton-Raphson 33 and by the approximation 0.24. For decision 96 the influence was highly overestimated by the approximation. Furthermore the approximation is generally worse than the Newton-Raphson one-step procedure.

To conclude, the approach of Moolgavkar et al. provides diagnostics for widely used discrete choice models with the computational advantage to avoid iterations for all decisions. In the MNL model the approximation (17) performs well whereas in the NMNL model the performance is dubious. This might be due to the high nonlinearity of the likelihood in the parameters.

Both examples have been calculated using GAUSS 2.0 R29. To avoid exceeding space limitations concerning the dimensions of the matrices in (15) in a first run $\hat{\mathbf{V}}^{1/2}\mathbf{X}_{\hat{\beta}}$, then $(\mathbf{X}'_{\hat{\beta}}\hat{\mathbf{V}}\mathbf{X}_{\hat{\beta}})^{-1}$ and in a second run the diagnostics (15) – (18) were computed.

References

Amemiya T. (1985). Advanced Econometrics. Basil Blackwell, Oxford

Cook R.D. (1977). Detection of influential observations in linear regression. Technometrics 19, 15-18

Domencich T.A., McFadden D. (1975). Urban travel demand. North Holland, Amsterdam

Hoaglin D.C., Welsch R.G (1978). The hat matrix in regression and ANOVA. Amer. Statist. 32, 17-22

Lesaffre E., Albert A. (1989). Multiple-group Logistic Regression Diagnostics. Applied Statistics 38, 425-440

Maddala G.S. (1983). Limited-dependent and qualitative variables in econometrics. Cambridge University Press

Moolgavkar S.H., Lustbader E.D., Venzon D.J. (1984). A geometric approach to nonlinear regression diagnostics with apllication to matched case-control studies. Annals of Statistics 12, 816-826

Otruba H., Gampe J. (1986). Untersuchung des Modal-Split im Ballungsraum Wien. Forschungsarbeiten aus dem Verkehrswesen, Bundesministerium für öffentliche Wirtschaft und Verkehr, Wien

Pregibon D.(1981). Logistic regression diagnostics. Annals of Statistics 9, 705-724