CHAPTER 7

DIAGNOSTIC TESTING OF HOUSING CHOICE MODELS

(P Longley, N Wrigley and R Dunn)

Modeling and Simulation, 1984, 15(1), 215-219

Diagnostic Testing of Housing Choice Models

by
Paul Longley (Department of Geography, University of Reading, Whiteknights, Reading RG6 2AB, UK),
Neil Wrigley (Department of Geography, University of Bristol, University Road, Bristol BS8 1SS, UK) and
Richard Dunn (Department of Geography, University of Bristol, University Road, Bristol BS8 1SS, UK)

1. INTRODUCTION

Recent years have seen the extension of discrete choice modelling to a number of new fields such as residential location and housing choice, retail location, car ownership decisions, occupational choice and attainment and the selection of routes for urban freeways. In these new areas, specification of the representative component of utility has become an increasingly important issue, and it has been realized that discrete choice models are beset by a wider variety of specification problems than standard econometric models. It is the contention of this paper that diagnostic procedures have an important role to play in the process of specification and refinement of discrete choice models.

A wide range of diagnostic techniques now exist within the standard regression analysis literature, and which are appropriate to an interactive computing environment (7). A more recent development has involved the extension of diagnostic tests to the binary logit model (5), (8). The main driving force behind the development of diagnostics is that standard fitting procedures, such as maximum likelihood, may be unduly influenced by outlying data points. When such methods are applied to data obtained from social surveys, rather than from controlled design experiments, the probability of such extreme data points increases, and it is important that the presence of these points is detected and their influence on the model fit is evaluated.

In this paper, these methods are introduced and discussed in the context of an empirical logit model of housing choice. Our example uses data taken from the 1976 English House Condition Survey (EHCS: (1), (2)). Specifically, we consider tenure choice within the rented sector of the housing market for 151 households resident in the Manchester conurbation.

2. THE EMPIRICAL EXAMPLE: BACKGROUND AND INITIAL RESULTS

There is an emerging consensus that some European perspectives upon housing choice have become too derivative of established United States models and continue to over-emphasize budget constraints within a rent-bid framework at the expense of state housing policy considerations. The net effect of the range of direct and indirect policy incentives in Britain, for example, is to manipulate housing choices in a number of tenure-specific ways. In particular, subsidized local government 'social housing' is allocated in accordance with prespecified need criteria, whilst tax incentives to owner-occupied dwellings distort housing consumption patterns within a general housing budget framework. The significance of needs versus budget frameworks will vary between local government areas, since local governments have varying stocks of 'social' housing as well as varying commitments towards finance of this sector using local and national taxes.

In using statistical techniques such as logit models to compare needs versus budget frameworks (eg. (4)), it is important to be able to distinguish between results which reflect genuine differences in behaviour patterns (eg. because of different local state attitudes to housing policies) from results influenced by anomalies in the data to which the models are fitted. Hence there is a need to use diagnostic procedures as an essential part of the model building process.

For our empirical example we start by estimating a binary logit model of the form

$$\log_e \{P_i/(1-P_i)\} = \beta_0 + \beta_1 \text{HOHAGE}_i + \beta_2 \text{HHSIZE}_i + \beta_3 \text{HHY}_i \qquad (1)$$

where P_i is the probability that the revealed preference of household i will be to rent within the private sector rather than the local authority sector,

HOHAGE is the age of head of household,

HHSIZE is household size, and

HHY is household income.

Similar specifications have been used previously within logit models of tenure choice ((3), (9)), and Longley (4) gives a rationale for alternative needs and budget based specifications similar to (1).

The standard maximum likelihood (ML) estimates of the logit model are:

$$\log_e \{\hat{P}_i/(1-\hat{P}_i)\} = \underset{\substack{(0.912)\\ \{2.58\}}}{2.353} - \underset{\substack{(0.124)\\ \{-3.75\}}}{0.464\ HOHAGE_i} - \underset{\substack{(0.175)\\ \{-2.93\}}}{0.512\ HHSIZE_i} + \underset{\substack{(0.015)\\ \{0.47\}}}{0.007\ HHY_i} \qquad \rho^2=0.126 \qquad (2)$$

(standard errors in brackets)
{t-statistics in square brackets}

The significance of the HOHAGE and HHSIZE parameter estimates is very encouraging and lends further support to our interpretation of the workings of the British rented housing market: that is, large households and later stage-in-lifecycle households are less likely to rent within the private sector. However, the HHY parameter estimate is not significant, despite some evidence to the contrary from other British empirical studies which argue the continued relevance of budget considerations in conjunction with public policy allocation criteria.

We now apply a series of diagnostic procedures to the data from which the model in (2) was calibrated. In particular, one role for these procedures is to ascertain whether the coefficient on income is unduly influenced by one or two unusual data points and whether it is worthy of further study given our a priori beliefs.

3. RESIDUALS AND LEVERAGES

The two basic building blocks for regression diagnostics are residuals and leverages. These measures aim to identify two types of unusual or 'bad' data (5) : see Figures 1 and 2. Residuals measure the distance between the observed and predicted values of the response variable, and serve to identify outliers in the dependent variables.

Residuals for logit regression models can be defined on a number of scales, but two have been found to be particularly useful. These are the components of chi-squared, χ_i, and the components of deviance d_i (See (5), (8) for further discussion and definitions). In the example of tenure choice in the Manchester conurbation, the χ_i and d_i measures identify observations 48, 49 and 77 as outliers. For example, observation 48 is a private renter who, given the parameter estimates in (2), has a predicted probability of renting in that sector of only 0.033.

Further analysis of the pattern of χ_i and d_i residuals is both possible and desirable - for example, to detect omitted independent variables, suitable transformations of the independent variables in the model, or the need to segment the sample into subgroups. This is not reported here, because of constraints of space.

The second building block for diagnostics are leverages, denoted h_{ii}, where large values identify data points which are extreme in the design space and which are influential in determining the form of the model. (For the binary logit model the leverage diagnostics are discussed and defined in (5) and (8)).

Visual interpretation of a plot of the leverage values associated with the Manchester data set reveals that observations 42 and 108 have very large leverage values. Both cases are households with very high income and very small household size. Moreover, and rather unexpectedly, one of these high income households is a local authority renter.

Use of leverage values in this way identifies observations which are unusual and may highlight data points which are so extreme that it may be decided to exclude them from the analysis - for example, if they are obviously miscoded. For the moment we see no need to exclude either of these data points.

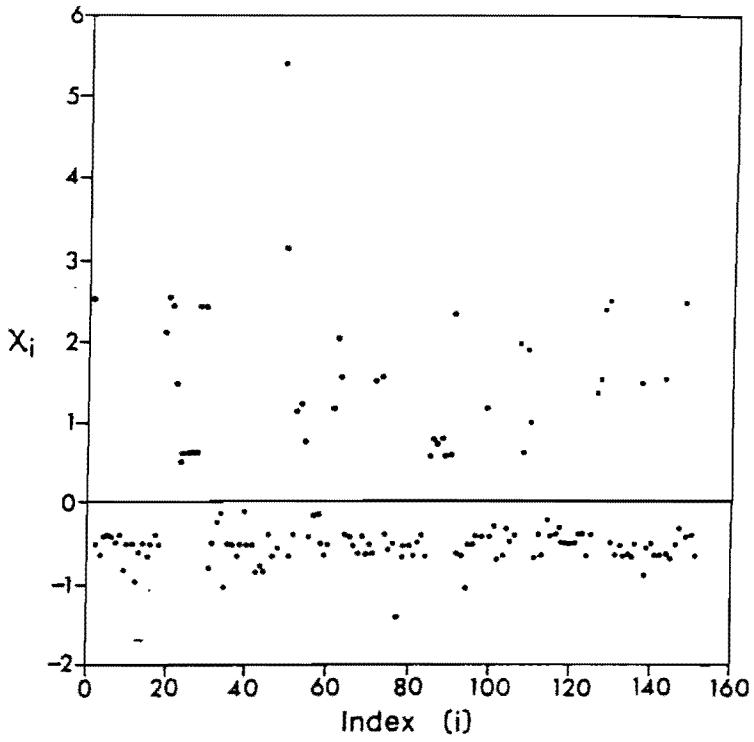

Figure 1: Components of Chi-Squared (χ_i) for the Manchester data

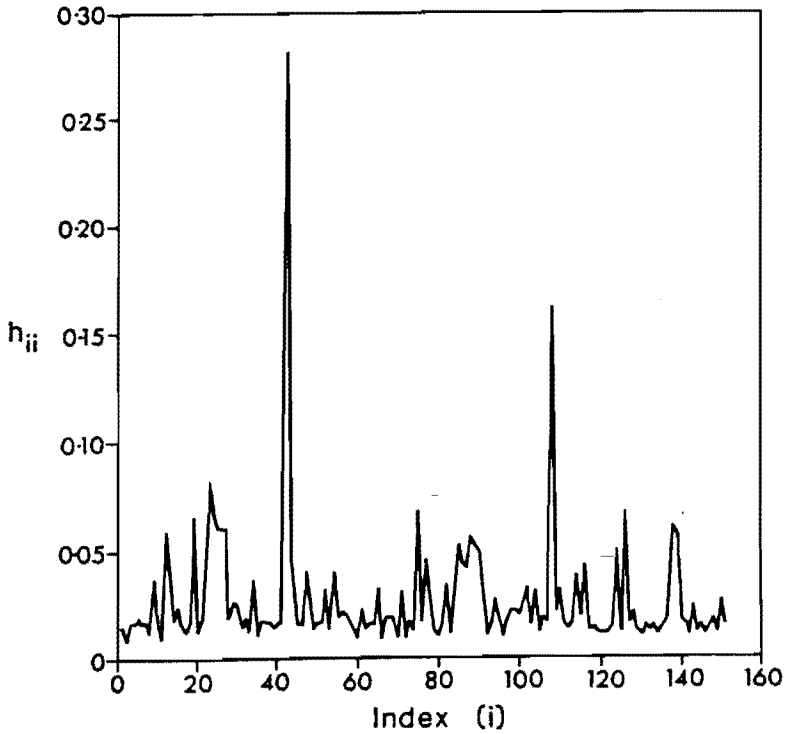

Figure 2: Leverage values (h_{ii}) for the Manchester data

4. SENSITIVITY ANALYSIS

Further diagnostic procedures explicitly consider the effect of deleting each observation in turn from the model.

The most important of these assess the influence of individual observations on the parameter estimates in the model: they are denoted $DBETA_{ik}$, where the subscript i indicates that observation i has been deleted, k indicates that we are concerned with the effect on the k'th parameter estimate. (For computational details, see (5) and (8))

Plots of $DBETA_{ik}$ for each explanatory variable k, against the index number i, allow identification of the data points which cause coefficient sensitivity and the coefficients which these data points affect. Since the $DBETA_{ik}$ are standardised, the plots also show approximately how great an effect the deletion of any one data point would have upon the coefficient in question.

In Figure 1 the $DBETA_{ik}$ plots are shown for HOHAGE (β_1: Figure 3 (a)), HHSIZE (β_2: Figure 3(b)) and HHY (β_3: Figure 3(c)). The plot $\hat{\beta}_1$ shows considerable stability, since no single observation changes the parameter estimates by more than about 0.3 of a standard error. However, in the case of $\hat{\beta}_2$ excluding observation 48 decreases $\hat{\beta}_2$ by about 0.9 of a standard error. Re-estimating the binary logit model without observation 48 yields:

$$\log_e \{\hat{P}_i/(1-\hat{P}_i)\} = \underset{(0.971)\ \{2.91\}}{2.827} - \underset{(0.130)\ \{-3.88\}}{0.503\ HOHAGE_i} - \underset{(0.200)\ \{-3.47\}}{0.694\ HHSIZE_i} + \underset{(0.016)\ \{0.676\}}{0.011\ HHY_i}$$

$$\rho^2 = 0.157 \qquad (3)$$

The rho-squared goodness of fit statistic improves from 0.126 to 0.157 and the overall prediction success index increases marginally from 67.04% to 68.23%. The effect of dropping observation 48 is therefore to increase the magnitude of the $\hat{\beta}_2$ parameter and to promote marginal increases in the efficiency and predictive success of the model.

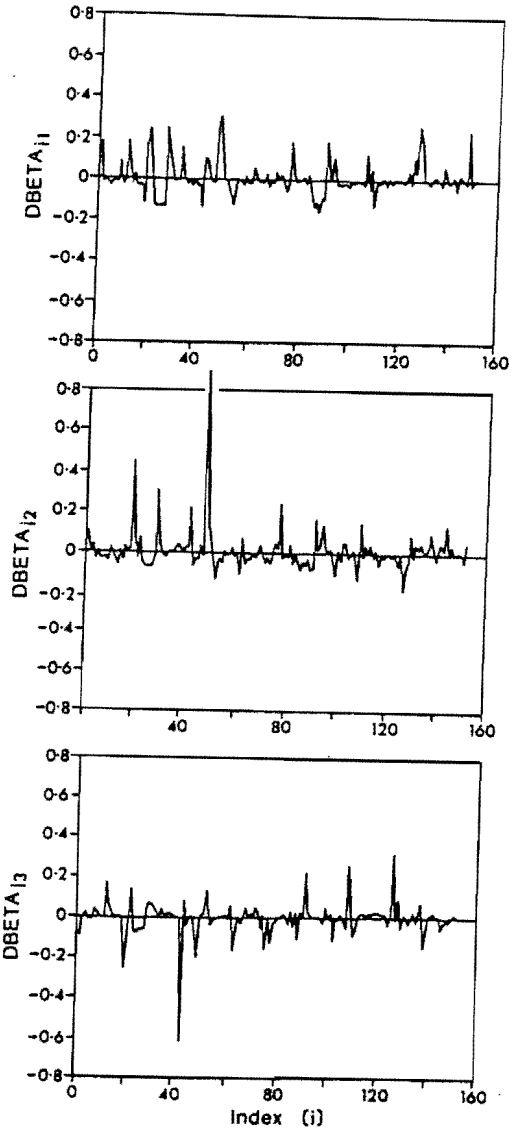

Figure 3 : $DBETA_{ik}$ plots for HOHAGE, HHSIZE and HHY

The $DBETA_{ik}$ plot for $\hat{\beta}_3$ again shows general stability, with the exception that observation 42 increases $\hat{\beta}_3$ by a fairly large amount (0.6 of a standard error). Observation 42 was identified as a high leverage point in the discussion above. Dropping this observation (but replacing observation 48) we find that:

$$\log_e \{\hat{P}_i/(1-\hat{P}_i(1.P_i)\} = \underset{(0.919)\ \{2.37\}}{2.183} - \underset{(0.124)\ \{-3.60\}}{0.447\ HOHAGE_i} - \underset{(0.185)\ \{-3.00\}}{0.554\ HHSIZE_i} + \underset{(0.018)\ \{0.92\}}{0.017\ HHY_i} \qquad (4)$$

$$\rho^2 = 0.130$$

The effect of dropping just observation 42 is therefore to increase the magnitude of the $\hat{\beta}_3$ parameter, although the increase relative to the standard error is insufficient to warrant reevaluation of the statistical significance of the household income variable.

The combined effect of dropping observations 42 and 48 yields:

$$\log_e\{\hat{P}_i/(1-\hat{P}_i)\} = \underset{(0.975)\ \{2.70\}}{2.634} - \underset{(0.130)\ \{-3.73\}}{0.435\ HOHAGE_i} - \underset{(0.213)\ \{-3.54\}}{0.753\ HHS_i} + \underset{(0.020)\ \{1.17\}}{0.023\ HHY_i} \qquad (5)$$

$\sigma^2 = 0.164$

5. CONCLUDING COMMENTS

We make two concluding points on the basis of the $DBETA_{ik}$ diagnostics. First, the sensitivity of the individual coefficients when individual observations are deleted highlights the instability inherent in maximum likelihood estimation procedures when used on data which include 'extreme' observations. This suggests that 'robust' or 'resistant' fitting procedures for logit models may be important ((8), (6)).

Second, assessment of the $DBETA_{ik}$ plots can reinforce or change our confidence in the significance of particular variables in the model. In the empirical example considered here, our confidence in the significance of the HOHAGE and HHSIZE variables is not changed by the deletion of observations, although it does reveal (in the case of $\hat{\beta}_2$) the sensitivity of the ML estimators to a single observation. Similarly, the magnitudes of the $\hat{\beta}_3$ parameter and its t statistic move in line with our a priori expectations, but it does not become statistically significant.

Over-all, the use of the logit model diagnostics on the Manchester data suggest that the results of the original binary logit model (Equation (2)) are reasonably robust. In view of this, we now have the confidence to question the validity of the original utility specification. For example, income may not be an important determinant of tenure choice given the wide availability of local authority housing in the Manchester conurbation. The implication is that the utility function should be respecified and the model should possibly be couched within a different decision context (4).

ACKNOWLEDGEMENTS

The authors are indebted to the Department of the Environment (UK) for supply of data from the 1976 English House Condition Survey. The Department bears no responsibility for the views and analysis set out in this paper, which are wholly the responsibility of the authors. Financial assistance from the Economic and Social Research Council (UK), grant D00230033 is also acknowledged by the last author.

REFERENCES

(1) Department of the Environment, English House Condition Survey 1976: Part 1 Report of the Physical Condition Survey, HMSO, London, 1978

(2) Department of the Environment, English House Condition Survey 1976: Part 2 Report of the Social Survey, HMSO, London, 1979.

(3) Li, M.M., A logit model of homeownership, Econometrica, 45, 1977, pp. 1081-1097.

(4) Longley, P.A., Comparing discrete choice models: some housing market examples, London Papers in Regional Science, Vol. 14 (Discrete Choice Modelling in Regional Science) Ed. D. Pitfield, Pion, London, 1984

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

(6) Pregibon, D., Resistant fits for some commonly used logistic models with medical applications, Biometrics, 38, 1982, pp. 485-498

(7) Wrigley, N., Quantitative methods: on data and diagnostics, Progress in Human Geography, 7, 1983, 565-575

(8) Wrigley, N. and Dunn, R., Diagnostics and resistant fits in logit choice models, London Papers in Regional Science, Vol. 14 (Discrete Choice Modelling in Regional Science), Ed. D. Pitfield, Pion, London, 1984

(9) Wrigley, N. and Longley, P.A. Discrete choice modelling in urban analysis, Ch.2. in Geography and the Urban Environment, Vol. 6, Eds. D.T. Herbert and R.J. Johnston, John Wiley, Chichester, 1984, pp. 45-94.