

GRAPHICAL PROCEDURES FOR IDENTIFYING FUNCTIONAL FORM IN BINARY DISCRETE CHOICE MODELS*

A Case Study of Revealed Tenure Choice

R. DUNN

University of Bristol, Bristol BS8 1SS, UK

P. LONGLEY and N. WRIGLEY

University of Wales Institute of Science and Technology, Cardiff CF1 3EU, UK

Received December 1985, final version received June 1986

This paper introduces procedures for plotting and analysing partial residuals for binary logit choice models. Seen within the general context of a movement away from purely inferential statistical model-building towards more 'data-analytical' or exploratory approaches, these graphical methods are presented as offering significant advantages over established methods of identifying the functional forms of discrete choice models. An empirical example is developed in the context of revealed tenure choices in London, England, in which a number of revisions are made to the functional form in the light of the analysis of partial residual plots. These results are discussed in relation to the dynamic and policy-dominated British housing market.

1. Introduction

In this short paper we describe an empirical analysis of urban tenure choice which demonstrates the importance of two recent research themes in the geographical and statistical literatures. The first theme is the utility of discrete choice models in urban analysis. The second is the importance of graphical diagnostic procedures to assess goodness-of-fit in statistical model building.

The potential of discrete choice models for urban geographers and planners has recently been reviewed by Wrigley and Longley (1984), and the reader is referred to that review for a full discussion. Two aspects of this approach, one substantive and the other methodological, form the point of departure for this paper. First, discrete choice models enable analysis of

*Thanks are due to W.S. Cleveland for providing software for the smoothing procedures used here. Data were supplied by the Department of the Environment: the Department bears no responsibility for the views and analyses set out in this paper, which are wholly the responsibility of the authors.

observed choice and judgement at the level of the *individual* decision maker. The importance of this relates to the belief that

'a theory which works well at the level of the individual is a better theory in itself, and is also likely to work better at the market level, than one which does not' [Green (1976)].

Secondly, discrete choice models have been developed within an interdisciplinary research context which embraces statistics, econometrics, transportation science and psychology, and strenuous efforts are now being made to assess the performance of these models in a variety of empirical contexts. Graphical diagnostics provide a valuable addition to the range of techniques currently available for the assessment of model performance, and have the additional advantage of linking discrete choice modelling to an important and rapidly developing area of statistical methodology.

Graphical diagnostics are now well established in conventional (OLS) regression analysis [see Belsey et al. (1980), Chambers et al. (1983), Cook and Weisberg (1982) and Wrigley (1983)] yet extension of these techniques to logit models – the workhorse of discrete choice modelling – is a more recent innovation which has been less fully explored. Diagnostic procedures for assessing the influences of individual observations in binary logit models have been described by Pregibon (1981, 1982) – for a geographical example see Wrigley and Dunn (1984) – and some further graphical methods for assessing binary logit models have been suggested by Landwehr et al. (1984).

In this paper, we concentrate on partial residual plots for logit models, as described by Landwehr et al. (1984). A primary aim of partial residuals is to assist in determining the correct functional form of exogenous variables included in a regression model. The use of partial residual plots in standard regression models is widely advocated and such plots are frequently included in empirical analyses [e.g., Jones (1984), Larsen and McCleary (1972), Weisberg (1980) and Wrigley (1983)]: the extension of such procedures to binary logit models marks an important advance in the specification and refinement of discrete choice models.

This emphasis on graphical procedures reflects a wider movement away from inferential statistical model-building based upon hypothesis testing towards more 'data analytical' or exploratory approaches. These latter, more recent, methods stress the need for an iterative approach to model building, using easily-executed graphical displays at all stages of the model building process: to explore the nature of the original data, to identify model structure and to assess the final model. It should be stressed, however, that rather than being alternatives, traditional 'confirmatory' approaches and exploratory (often graphical) methods should be seen as complementary, with both having an important role in statistical model building [Tukey (1980)].

The importance of graphical displays per se derives from the fact that the

'eye-brain system is the most sophisticated information processor ever developed, and through graphical displays we can put this system to good use to obtain deep insight into the structure of data.' [Chambers et al. (1983, p. 1).]

In the specific context of plots of partial residuals, graphical summaries have the potential to identify non-linearities of all kinds, from quadratic and cubic functions and other power transformations to clear discontinuities or step-functions in the data. A further important statistical motivation behind such approaches is that certain individual observations may be extremely important in model fitting [Pregibon (1981, 1982)] and model refinement [Atkinson (1982)] and it is important to identify if such observations exist in any empirical analysis, a task for which graphical methods are particularly well suited.

In contrast, the most widely used procedures for testing functional form in logit models at the time of writing, the Box-Cox and Box-Tukey transformations [see Wrigley and Longley (1984) and Wrigley (1985) for more details], do not explicitly consider individual observations or use graphical procedures. Typically, two additional parameters are specified in order to facilitate a range of statistical transformations of the representative component of utility. Incremental adjustments in these parameters permit a 'search' for an appropriate functional form, the results of which are assessed using log-likelihood ratios. Hensher and Taylor (1983) emphasise that such transformations are

'no perfect substitute for direct behavioural specification, but can assist in the specification of an improved initial set of assumptions upon which models are derived from theory.'

In the absence of theoretical guidelines specific to a particular empirical context, the Box-Tukey method remains the standard procedure for identifying functional form. It nevertheless remains an undesirably mechanistic means of deriving functional form which also incurs considerable computational expense. We suggest that the Box-Cox and Box-Tukey transformations are unnecessarily clumsy and cumbersome in comparison with appropriate partial residual plots.

The remainder of this paper takes the following form. In the next section, the derivation and analysis of partial residuals for binary logit models is discussed. In section 3, an empirical example of urban tenure choice is presented, in which we concentrate on the use of partial residual plots to ascertain the correct functional form of the exogenous variables in the model. We note that important behavioural implications are revealed by a changed model specification. The final section emphasises the importance of such

methods, both in general and for the analysis of urban tenure choice, and outlines areas for further research.

2. Partial residual plots for binary logit models

We begin by defining the basic binary logit model for observation/individual i . The observed response of dependent variable, Y_i , takes the value of 1 or 0, with probability p_i that $Y_i=1$. The logit transformation $\text{logit}(p_i)=\log_e(p_i/(1-p_i))$ is then modelled as a linear function of a vector of exogenous variables, x_i , giving

$$\text{logit}(p_i) = x_i\beta, \quad i = 1, \dots, N, \quad (1)$$

where β is a vector of parameters and N is the number of observations. This may be re-expressed in matrix form for the whole sample as

$$\text{logit}(p) = X\beta, \quad (2)$$

where p is a vector of choice probabilities (p_i 's), X the $N \times k$ matrix of independent variables which includes a constant and k is the number of β parameters. The vector of observed responses is denoted y . Maximum likelihood (ML) estimates of the $(k \times 1)$ vector β , denoted $\hat{\beta}$, are readily obtained for a given data set using, for example, the GLIM computer package [Baker and Nelder (1978)].

The logit partial residual with respect to a particular exogenous variable z , r_{par} , is then given by

$$r_{\text{par}} = (y - \hat{p})/(\hat{p}(1 - \hat{p})) + z\hat{\beta}_z, \quad (3)$$

where $\hat{\beta}_z$ is the parameter estimate for the particular exogenous variable z . A plot of this partial residual against z then forms the basis for assessing the correct functional form of z in the model.¹ If the relationship between $\text{logit}(p)$ and z is linear, then the partial residual plot should be linear in form: non-linear relationships should be identified by non-linear plots. This is in essence the same procedure as is well established for ordinary regression models. The complicating factor for the binary logit model is that the partial residuals fall into two 'clouds', which correspond to the values $Y_i=1$ and $Y_i=0$: this makes the identification of the underlying functional form in the partial residual plots considerably more difficult than is the case for ordinary regression.

Landwehr et al. (1984) show that smoothing the partial residual plots provides a powerful way of identifying the correct relationship between $\text{logit}(p)$

¹In these plots z is usually expressed in terms of deviations from its mean - we follow this procedure below.

and z . This smoothing is equivalent in statistical terms to the grouping or pooling of data, a procedure which is made necessary by the discreteness of Y_i , and which is common to a number of procedures for assessing logit models. The smoothing of partial residual plots is thus equivalent to grouping observations using values of the independent variable z .

The way in which partial residual plots are smoothed is crucial, since the smoothing identifies the functional form for z . Landwehr et al. (1984) advocate using the lowess (locally weighted scatterplot smoothing) procedure due to Cleveland (1979): see also Chambers et al. (1983) and Cleveland and McGill (1984). Lowess produces non-robust and robust smooths for any scatter plot: given a set of points (x_i, y_i) where y is regarded as in some sense dependent on x , lowess computes and plots a set of points (x_i, \hat{y}_i) called smoothed values, where \hat{y}_i aims to summarise the middle of the distribution of y at $x = x_i$. (In the specific case here we plot r_{par} on the y axis and z on the x axis and aim to summarise the form of dependence of r_{par} on z).

Because of the importance of the smoothing procedure in assessing partial residual plots for logit models we now briefly outline lowess, following Cleveland and McGill (1984). More detailed accounts are to be found in Cleveland (1979) and Chambers et al. (1983).

Step 1. The user chooses a number f between 0 and 1. Let $q = fn$ rounded to an integer, where n is the number of observations.

Step 2. For each x_i , fit a line to the q points on the scatterplot whose abscissas are closest to x_i . The fitting is done by weighted least squares in which points close to x_i receive large weight, and those far away receive less weight.

Step 3. The fitted value at x_i , \hat{y}_i , is the y value of the fitted line at $x = x_i$.

These three steps define a non-robust smooth, since the least-squares fitting in step two may be distorted by a small number of outlying y_i values. To make lowess a robust procedure, that is one that is not unduly influenced by a small number of outliers, two further steps may be added:

Step 4. Residuals $r_i = y_i - \hat{y}_i$ are computed. Weights are assigned at each point (x_i, y_i) according to the size of the residual; if r_i is close to 0 the weight is large, and if r_i is far from 0 it is small.

Step 5. Steps two and three are repeated, but now in the fitting of the line to get \hat{y}_i , the weights described in step two are multiplied by the robustness weights from step four so that outliers are down weighted.

Together steps four and five are termed a robustness iteration or step; smooths with one or more robustness iterations may be derived, but generally one or two iterations are considered sufficient. The non-robust smooth is sometimes referred to as one with no robustness iteration.

The parameter f controls the degree of smoothing, with smaller values giving more resolution at the cost of increased noise; most users of lowess use values of f between 0.33 and 0.67. In any particular situation different values of f , and different numbers of robustness iterations, will give different results, and it is usual, and necessary, to look at a series of smooths for each scatterplot.

The need to consider a series of smooths is especially important when using partial residuals from logit models. First, because the composition of these scatter plots is complex, with two clouds of points, and, second, because what little research currently exists into the behaviour of these plots and their smooths under the null hypothesis suggests the need for caution. For example, Chesher (1985) has shown that under certain circumstances robust smooths may take highly non-linear forms when the true relationship between logit (p) and z is linear.

These methods for smoothing partial residual plots are illustrated using empirical material in the following section. From a computational standpoint it is important to note that partial residuals can be calculated for each independent variable in turn using only one model fit,² and that they may readily be calculated in a computer package such as GLIM. A further smoothing algorithm is, however, necessary.

3. An empirical example: Tenure choice within the rented sector

The empirical example presented below concerns revealed tenure choices *within the rented sector* of the British housing market. The dependent variable takes the form $Y_i=1$ for renting in the private sector, and $Y_i=0$ for renting from the local authority. We use a sample of 309 households, all from the London conurbation, the data for which are taken from the 1976 English House Condition Survey [Department of Environment (1978, 1979)].

The background to this example is the debate between 'needs/use-value' versus 'budget/consumption' frameworks for the analysis of urban tenure choice. Some European perspectives on housing choice have been criticised as being too derivative of established models from the United States which stress the paramount importance of budget constraints within a bid-rent framework. The alternative is to stress the range of direct and indirect housing policy incentives provided by the State, which conspire to manipulate

²Strictly, the derivation of partial residuals requires that all variables in the model are correctly specified except the one of interest. However, Landwehr et al. (1984, p. 82) state that in practice little power appears to be lost when several variables are simultaneously misspecified.

housing choice in a number of tenure-specific ways. Important in the present context is the fact that British local authority housing tends to be subsidised and to fulfil a role as 'social housing', with allocation being in accordance with prespecified socially-defined need criteria. Our case study may thus be seen in the context of one of the recurrent debates in behavioural geography, namely the relative merits of postulating observed behaviour as revealed in a constraint-dominated versus essentially constraint-free choice context. In the following statistical analysis we aim to determine how the choice between private and local authority renting varies with important household characteristics, and to relate these results to a broader knowledge of the British housing market.

Our starting point is the following binary logit model, which considers the odds of renting in the private rather than the public (i.e., local authority 'social housing') sector of the housing market.

$$\text{logit}(\hat{p}) = 2.42 + 0.0157I - 0.0428A - 0.589S, \quad (4)$$

(0.63)	(0.0069)	(0.0083)	(0.126)
[3.84]	[2.27]	[5.16]	[4.67]

$$\text{deviance} = 351.7, \quad df = 305, \quad \rho^2 = 0.126$$

(deviance of 'constant only' model = 402.4),

where

I = household income (£'00 per annum, 1976 prices),

A = age of head of household, and

S = household size.

Standard errors are in round parentheses, t tests in square parentheses. The deviance is minus twice the maximised log-likelihood.

The significance of these parameter estimates is encouraging, and lends further support to previous studies of the British housing market. Within the rented sector the probability of renting from a private landlord decreases significantly with increasing age of head of household and with increased household size, in accordance with 'needs-based' interpretations of the provision of rented local authority housing. This is because local authorities are pledged to respond to the particular needs of the elderly and of large families. The probability of renting from a private landlord increases significantly as income increases, suggesting the continued relevance of budget considerations in determining the split between state housing and the remnants of a market-clearing private rental sector.

Models of this genre abound in the social science literature (notably in studies of travel demand), whilst the variable specification is fairly typical of the increasing number of discrete choice formulations of housing market behaviour. With the exception of those studies which use Box-Cox and Box-

Tukey transformations (such as those referred to above), most of these studies adopt a linear functional form for the exogenous variables. However, in view of our preceding comments concerning the complex and multidimensional decision environment characteristic of the British housing market, we suggest that it may be inappropriate to impose the linear functional form in such an arbitrary manner. For example, income effects might be complicated by the division of private renters into a low income group unable to secure access to local authority accommodation and a high income group for whom this sector retains advantages in a comparatively free choice situation.

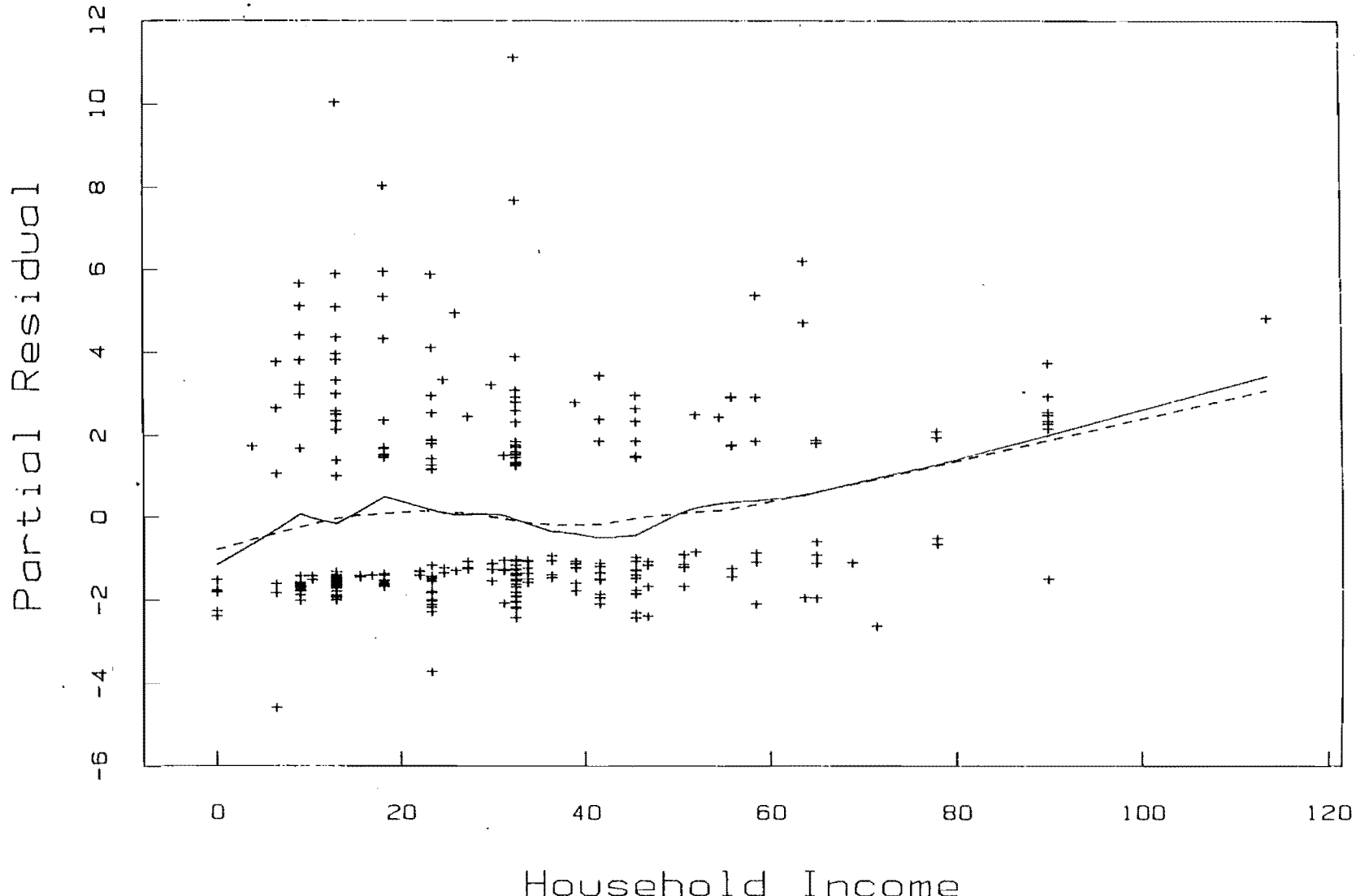
To assess the appropriateness of this model, with regard to its functional form, partial residuals were calculated for all three exogenous variables. In fig. 1 the partial residual plot for household income (I) is shown together with two non-robust smooths. The more erratic smooth has $f=0.35$, while the smoother curve has $f=0.65$; here the larger value of f appears to pick out the general nature of the dependence of the partial residual more successfully. [The clustering on certain values of household income – that is in the x direction – reflects the nature of individuals' responses to the survey question on yearly income: for example, the many values at $I=13$ (£1,300 per annum) reflects a response of a weekly income of £25.]

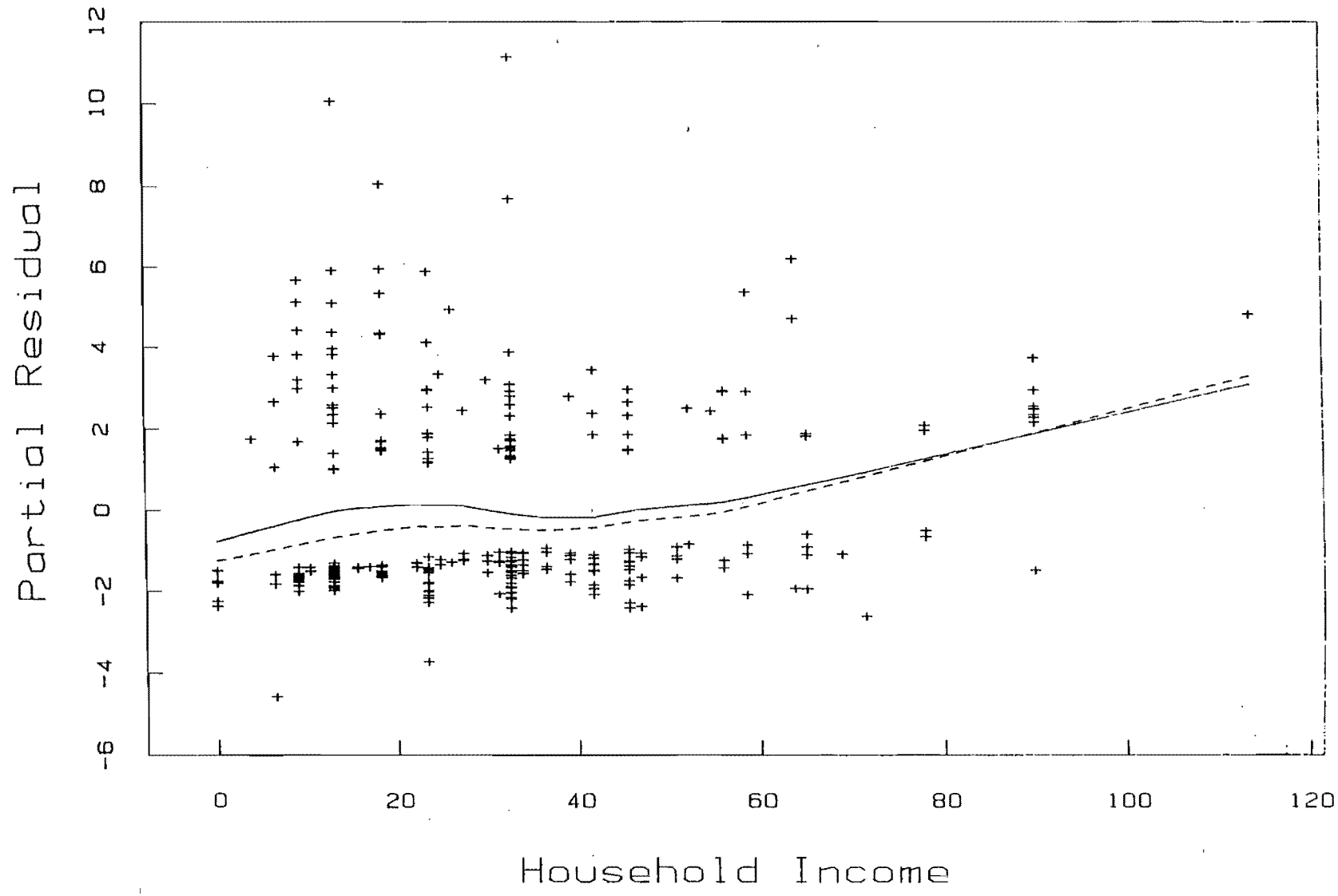
Fig. 2 shows the same partial residual plot with the non-robust smooth with $f=0.65$, together with a robust smooth with two robustness iterations and $f=0.65$. (The robust smooth with one robustness iteration lies between these two curves.) The robust smooth is generally lower, particularly on the left of the graph, where the robustness iterations downweight the few very large positive residuals which greatly influence the non-robust smooth. However, the overall shapes of the two curves are reassuringly similar.

The important inference to be drawn from figs. 1 and 2, and from other smooths calculated but not shown here, is that the relationship between $\text{logit}(p)$ and the household income variable is not linear. Furthermore, all this evidence suggests one particular functional form, namely a cubic. A possible rationale for this unexpected functional plot is discussed below.

In fig. 3 the partial residual plot for the age of head of household variable (A) is shown, with a non-robust smooth and a robust smooth with one robustness iteration; both have $f=0.65$. Again the non-robust curve is higher, as it is pulled towards a small number of large positive residuals. The influence of two large positive residuals at the top right of the scatterplot is particularly noticeable, which pulls the curve up for high values of A . In this case the two smooths differ in nature but both do suggest a clear non-linearity, as did other smooths with differing values of f , and the robust smooth with two robustness iterations.

Deciding the correct functional form for this exogenous variable is less straightforward, since a quadratic function, a logarithmic transformation, and a step function with a break at about $A=40$ are all possible candidates.







Deciding which of these forms is in some way 'best' is difficult, since the problem is one of testing between non-nested models, and the criteria for assessing goodness-of-fit are not necessarily simple (largely because of the influence of outliers). After a series of model specifications had been investigated the quadratic form was adopted for the following reasons:

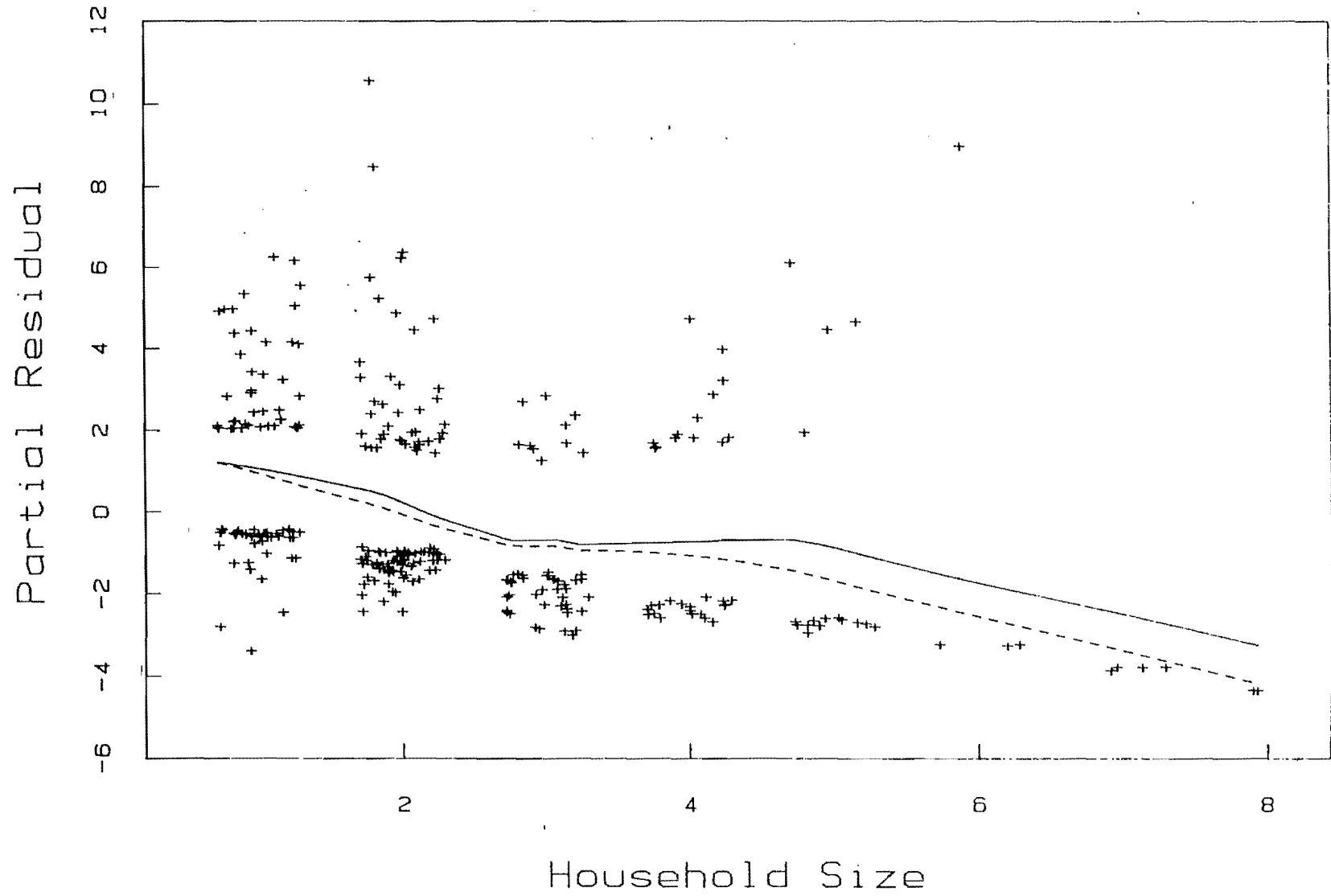
- (a) when the logarithm of A was included in the model its partial residual plot was not linear, as it should be if that transformation were appropriate;
- (b) various step functions, estimated by dummy variables on the slope coefficient, were not statistically significant when included in the model;
- (c) the quadratic function was statistically significant when included; and
- (d) the coefficients on the quadratic form for age of head of household gave intuitively plausible results, given a prior knowledge of the British housing market, as discussed below.

The way in which the appropriate functional form was decided upon for this exogenous variable provides an interesting example of the way in which exploratory graphical procedures suggest a series of possible functional forms (some unexpected), which may then be investigated further – by both confirmatory methods and a second round of graphical diagnostics.

The third variable in the model, household size, differs in nature from the previous two, as it is measured on an integer scale over a small range (here the range is 1 to 8 with only ten values >5). To treat this as a continuous variable, as we have done in eq. (4), assumes the metric scale of measurement is appropriate so that, for example, the effect on $\text{logit}(p)$ of changing household size by one is the same whether it be from 1 to 2, 2 to 3 or 3 to 4 and so on.

The use of partial residuals here is complicated by the discrete nature of the exogenous variable, but such plots may help to determine whether the linear formulation of eq. (4), which has the great advantages of simplicity and parsimony, is appropriate. Alternatively the partial residuals may indicate that this assumption of linearity is unjustified and suggest an alternative treatment of the variable.

Fig. 4 shows the partial residual plot for household size (S) from eq. (4), where, to avoid the problem of multiple over plotting of characters, the values of S have been 'jittered' by adding a random value between -0.3 and 0.3 to each observation. [For a further discussion of jittering for scatterplots see Chambers et al. (1983)]. In fig. 4 the upper smooth is a non-robust smooth with $f=0.65$, the lower curve is a robust smooth with one robustness iteration and the same f value. Alternative values of f and other robust smooths showed very similar patterns, as did smooths in which no jittering was performed.



As expected the partial residual plots for this variable are more difficult to interpret than those discussed so far, but it does not seem that linearity is a reasonable assumption given the shape of these smooths. However, for the present we choose not to alter the functional form of this exogenous variable, for two reasons: first, it is not clear from fig. 4 in what form the household size variable should enter the model, and, second, the clear evidence of the need to redefine the functional form of the two other variables in the model suggests that this should be the next step, since those redefinitions may produce a different pattern of partial residuals for household size.

These arguments suggest a new model specification which includes a cubic form for I (household income) and a quadratic for A (age of head of household). The ML parameter estimates of this model are:

$$\begin{aligned} \text{logit}(\hat{p}) = & 4.56 + 0.0866I - 0.00226I^2 + 0.0000232I^3 \\ & (1.23) (0.0613) (0.00162) (0.0000123) \\ & [3.71][1.45] [1.40] [1.89] \\ & -0.161A + 0.00115A^2 - 0.5527S, \quad (5) \\ & (0.048) (0.00048) (0.1359) \\ & [3.35] [2.40] [4.07] \end{aligned}$$

$$\text{deviance} = 335.9, \quad df = 302, \quad \rho^2 = 0.165.$$

The quadratic term on A is clearly significant, and the joint inclusion of the quadratic and cubic terms in I also produces a significant improvement in the model fit ($\chi^2 = 8.7$ with two degrees of freedom). Our investigation of the functional form of I and A thus leads to a respecification of the tenure choice model which, in statistical terms, is a significant improvement over the original model of eq. (4). However, the question of the functional specification for household size is still unresolved.

The partial residuals for the household size variable were next calculated from the model of eq. (5), and when various smooths were investigated very similar patterns to those of fig. 4 were observed; that is, a clearly non-linear functional relationship between $\text{logit}(\hat{p})$ and S was identified. It seems that the simplifying assumption of entering S as a quasi-continuous variable is not justified.

To investigate in more detail the influence of household size in the model this variable was recoded as five binary dummy variables: household sizes 1, 2, 3, 4 and 5 or more, which we denote as S_1 , S_2 , S_3 , S_4 , and S_5 . A new formulation of the model then excluded S and entered S_1 , S_3 , S_4 and S_5 as binary dummies – so that the base category is a two-person household, the largest of the five categories.

The ML parameter estimates of this model are:

$$\begin{aligned}
 \text{logit}(\hat{p}) = & 4.07 + 0.1036I - 0.00306I^2 + 0.0000269I^3 \\
 & (1.36) (0.0635) (0.00167) (0.0000126) \\
 & [2.98] [1.63] [1.83] [2.13] \\
 & -0.184A + 0.00131A^2 + 0.791S_1 \\
 & (0.0512) (0.000500) (0.358) \\
 & [3.59] [2.62] [2.21] \\
 & -1.838S_3 - 0.623S_4 - 1.547S_5, \\
 & (0.500) (0.488) (0.619) \\
 & [3.68] [1.28] [2.50]
 \end{aligned} \tag{6}$$

$$\text{deviance} = 323.3, \quad df = 299, \quad \rho^2 = 0.245.$$

These results are encouraging in many respects. The parameter estimates on the variables A and I are stable compared to those of eq. (5). The significance levels increase for each of the three I parameters, and the parameters on I^2 and I^3 are now significant at conventionally accepted levels. The deviance drops by 12.6 for the loss of three degrees of freedom, and the ρ^2 measure increases to 0.245, which falls within the frequently quoted range 0.2–0.4 said to represent a satisfactory model.

The parameter values on the four dummy variables for household size follow a pattern which is almost representative of a linear trend, with the important exception of S_3 . In this sample three-person households are the most likely of all household sizes to rent in the local authority, rather than the private, sector. One possible explanation of this result is that an interaction exists between the age of head of household and household size: for example, rather than dealing separately with these two variables it may be preferable to define, and use in the model, household types, such as young couples with one or two children, extended families of mixed ages, or households of one or two elderly persons. (Unfortunately this is not possible in the current case due to data limitations.) However, the important point to stress is that use of household size as a quasi-continuous variable with linear functional form is clearly not justified. Disaggregating this variable has improved model performance and suggested possible further areas for model development.

4. Discussion

It is important and interesting to consider the behavioural implications of the new specifications of eqs. (5) and (6). To do this fig. 5 shows a plot,

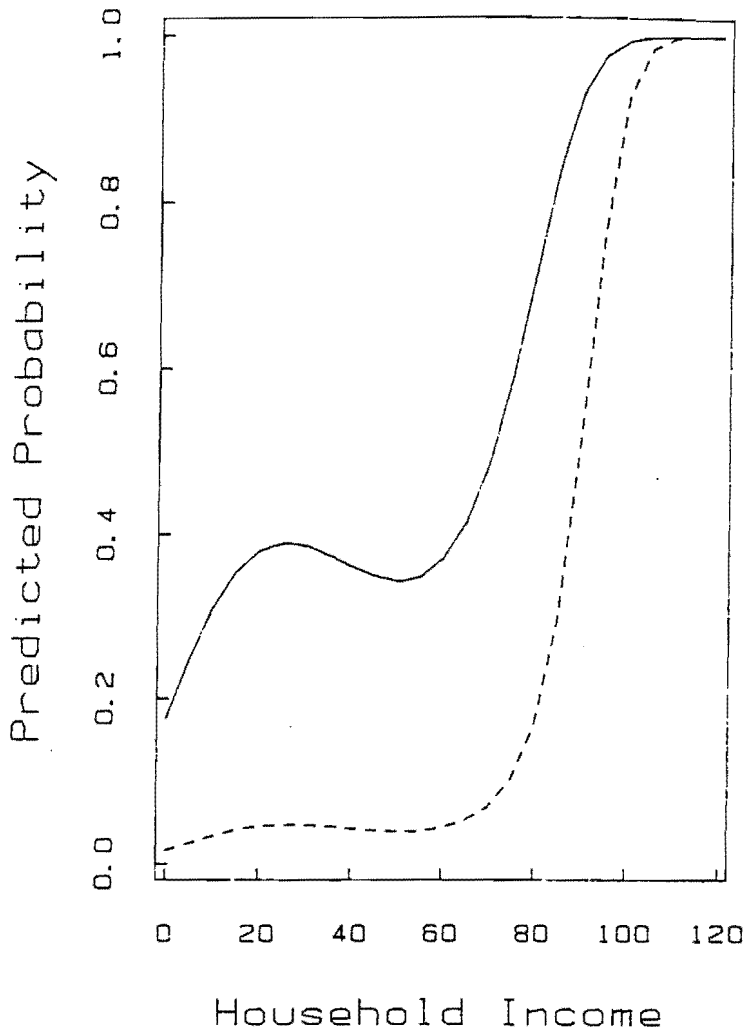


Fig. 5. The predicted probability of renting in the private sector as a function of household income (I) (£'00 per annum, 1976 prices). The dashed line is for age of head of household of 60, the solid line for age 25. Both lines for a household size of three.

derived from the parameter estimates of eq. (6), of the predicted probability of renting in the private sector against household income (I). The lower line is for a case with the head of household aged 60, the upper line aged 25; both are for households with three persons.

Two effects are to be noted from fig. 5. First, the effect of increasing the age of head of household from 25 to 60 has a large effect on the predicted probabilities of revealed tenure choices, with 25 year olds exhibiting a consistently higher propensity to rent in the private sector. This accords with an a priori expectation that lifecycle groups at or about retirement age are generally more likely to exhibit socially-recognised needs than the more variable requirements of households with young heads.

Second, the predicted probability of renting in the private sector increases as income rises to £2,500, falls as income rises from £2,500 to £5,000, and

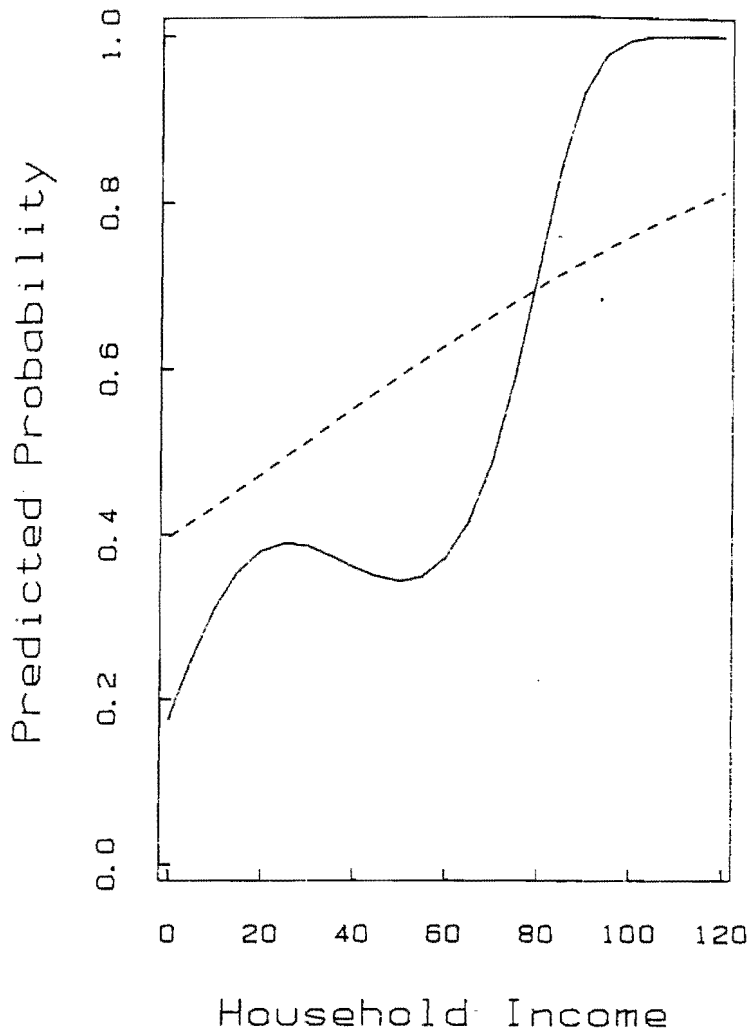


Fig. 6. The predicted probability of renting in the private sector as a function of household income (I) (£'00 per annum, 1976 prices), assuming the standard linear functional form (dashed line) and the cubic functional form (solid line). Both lines for age of head of household 25 and household size of three.

increases again as income rises beyond £5,000, reflecting the cubic function of I in the model. Assuming that local authority allocations closely mirror patterns of demonstrable household needs, this result is interesting insofar as it demonstrates that high social need groups do not coincide wholly with those on lowest incomes. In short, local authority subsidies do not have a simple effect upon the housing consumption of low income households.

This phenomenon is obscured, however, if we use the linear functional form for I of eq. (4). For example, fig. 6 compares the predicted probabilities of renting in the private sector as a function of household income for a 25-year-old head of household, holding household size constant at three, using eqs. (4) and (6). Our assessment of the influence of income upon the choice of tenure within the renting sector is markedly different between the two models. The linear functional form of eq. (4) seems, in this case, to represent an unnecessary and undesirable simplification of that relationship.

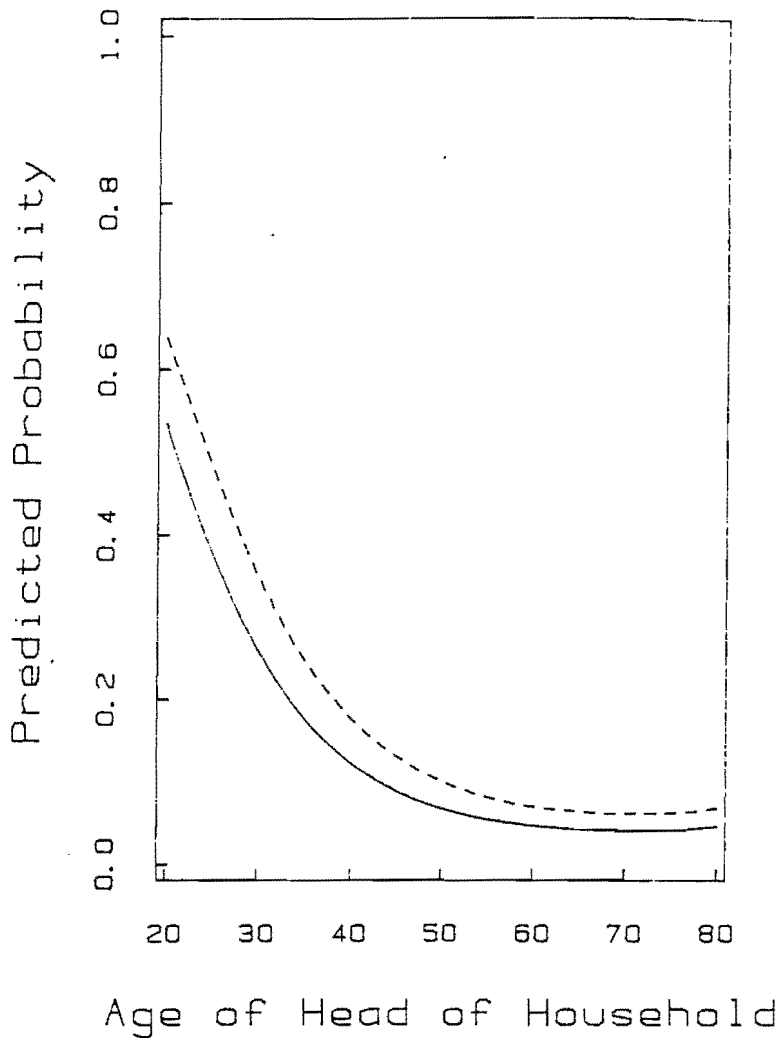


Fig. 7. The predicted probability of renting in the private sector as a function of age of head of household (A). The solid line is for a household income of £2,000 p.a., the dashed line for £7,000 p.a. Both lines for a household size of three.

Fig. 7 shows a similar plot to fig. 5, this time measuring the effect of age of head of household on the probability of renting in the private sector, using the parameter estimates of eq. (6). The two curves are for households with incomes of £2,000 p.a. (lower line) and £7,000 p.a. (upper line), both of three persons. The main feature here is that the probability of renting in the private sector does not fall consistently as a linear functional form would predict: rather, when the age of head of household exceeds about 50 the probability of renting in the private sector remains constant or increases slightly. Fig. 8 compares the predicted probabilities obtained using eqs. (4) and (6) for a three-person household with income £2,000 per annum. Again, our interpretation of how a particular exogenous variable (in this case age) affects the choice process is altered by the move from the first linear specification to the preferred (quadratic) specification.

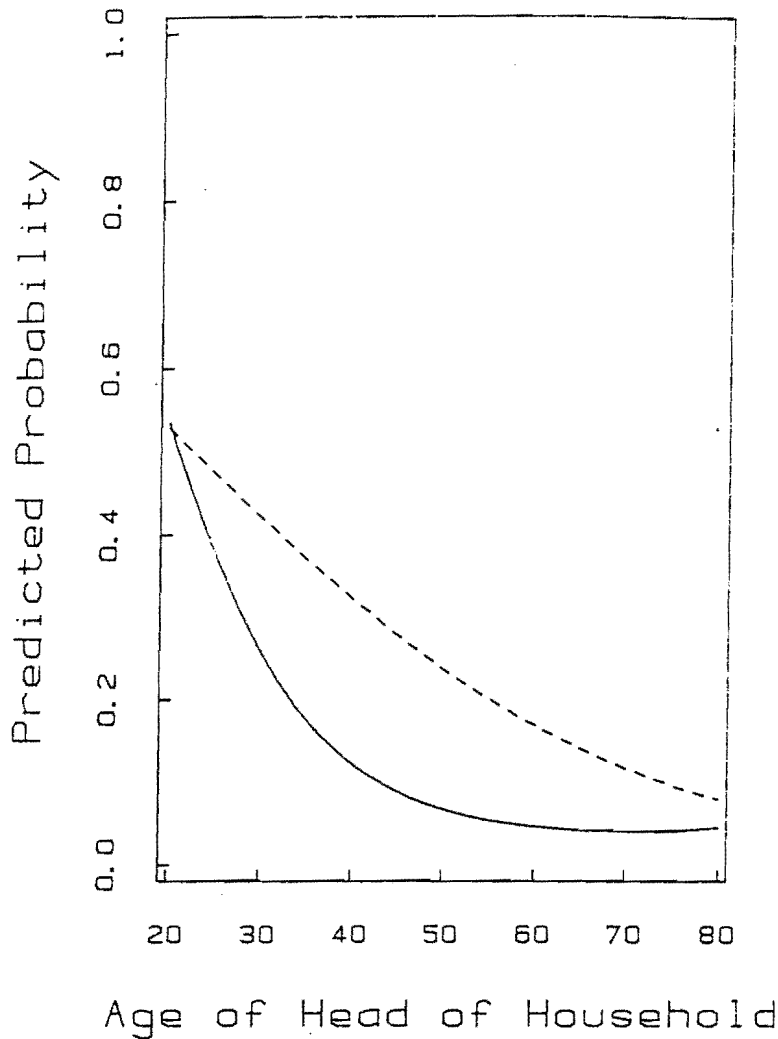


Fig. 8. The predicted probability of renting in the private sector as a function of age of head of household (A), assuming the standard linear functional form (dashed line) and the quadratic functional form (solid line). Both lines for a household income of £2,000 p.a., and household size of three.

Substantive assessment of this result requires that we make passing comment about the decreasing relative magnitude in recent years of the aggregated rental sector in the face of an expansion of owner-occupation. The British owner-occupied sector has undergone remarkable growth in recent decades [Ball (1983)] and currently houses approximately 61% of all households. In view of the fact that access to this sector has been controlled by the lending policies of mortgage finance institutions such as building societies, there is evidence to suggest that those becoming owner-occupiers are primarily households with high current incomes and a future income stream which is long enough for loan repayment purposes. To an extent, therefore, we might anticipate that the comparatively recent growth of owner-occupation has been fuelled by the allocation of loans to younger households, and that some older households have been left behind in the increasingly residualised private renting sector.

5. Conclusions

This paper has applied just one of a developing range of graphical diagnostic techniques for logistic regression to an urban tenure choice problem. In conclusion, we make three comments which arise from the previous discussion. First, graphical data-analytical approaches to the investigation of functional form offer a considerable improvement in flexibility compared to the mechanistic and restrictive Box-Cox/Box-Tukey traditions. Second, the interpretation of how particular exogenous variables affect the choice process may be significantly altered in the move from the initial (conventionally linear) functional form to the finally selected functional specification. Third, the development of diagnostic graphical tests for discrete choice models remains at a formative stage, and the future holds the prospect of a number of important developments: in particular, we may now anticipate the extension of these procedures to multinomial choice situations, and the integration of these procedures into an interactive computing environment as a standard technique in discrete choice modelling.

References

- Atkinson, A.C., 1982, Regression diagnostics, transformations, and constructed variables, *Journal of the Royal Statistical Society, Series B*, 44, 1-36.
- Baker, K.J. and J.A. Nelder, 1978, *The GLIM system release 3: Generalised linear interactive modelling* (Numerical Algorithms Group, Oxford).
- Ball, M.J., 1983, *Housing policy and economic power: The political economy of owner occupation* (Methuen, London).
- Belsey, D.A., E. Kuh and R.E. Welsch, 1980, *Regression diagnostics* (Wiley, New York).
- Chambers, J.M., W.S. Cleveland, B. Kleiner and P.A. Tukey, 1983, *Graphical methods for data analysis* (Wadsworth, Boston, MA).
- Chesher, A., 1985, Smoothing logit residual plots, Discussion paper 85/164 (Department of Economics, University of Bristol).
- Cleveland, W.S., 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* 74, 829-836.
- Cleveland, W.S. and R. McGill, 1984, The many faces of a scatterplot, *Journal of the American Statistical Association* 79, 807-822.
- Cook, R.D. and S. Weisberg, 1982, *Residuals and influence in regression* (Chapman-Hall, London).
- Department of the Environment, 1978, *English house condition survey 1976, part 1: Report of the physical condition survey* (HMSO, London).
- Department of the Environment, 1979, *English house condition survey 1976, part 2: Report of the social survey* (HMSO, London).
- Green, H.A.J., 1976, *Consumer theory* (Macmillan, London).
- Hensher, D.A. and A.K. Taylor, 1983, Intraurban residential relocation choices for students: An empirical enquiry, *Environment and Planning A* 15, 815-830.
- Jones, K., 1984, Graphical methods for exploring relationships, in: G. Bahrenberg, M.M. Fischer and P. Nijkamp, eds., *Recent developments in spatial data analysis: Methodology, measurement, models* (Gower, Aldershot) 375-392.
- Landwehr, J.M., D. Pregibon and A.C. Shoemaker, 1984, Graphical methods for assessing logistic regression models (with discussion and rejoinder), *Journal of the American Statistical Association* 79, 61-83.
- Larsen, W.A. and S.J. McCleary, 1972, The use of partial residual plots in regression analysis, *Technometrics* 14, 781-790.

- Pregibon, D., 1981, Logistic regression diagnostics, *Annals of Statistics* 9, 705–724.
- Pregibon, D., 1982, Resistant fits for some commonly used logistic models with medical applications, *Biometrics* 38, 485–498.
- Tukey, J.W., 1980, We need both exploratory and confirmatory, *American Statistician* 34, 23–25.
- Weisberg, S., 1980, *Applied linear regression* (Wiley, New York).
- Wrigley, N., 1983, Quantitative methods: on data and diagnostics, *Progress in Human Geography* 7, 567–577.
- Wrigley, N., 1985, *Categorical data analysis for geographers and environmental scientists* (Longman, London).
- Wrigley, N. and R. Dunn, 1984, Diagnostics and resistant fits in logit choice models, in: D. Pitfield, ed., *London papers in regional science*, Vol. 14, *Discrete choice modelling in regional science* (Pion, London) 44–66.
- Wrigley, N. and P.A. Longley, 1984, *Discrete choice modelling in urban analysis* in: D.T. Herbert and R.J. Johnston, eds., *Geography and the urban environment*, Vol. 6 (Wiley, Chichester) 45–94.