Statistical and Learning Approaches to Nonlinear Modeling of Labour Force Participation

Georgios Paliouras
paliourg@iit.demokritos.gr
Inst. of Informatics
and Telecommunications
NCSR "Demokritos"
Aghia Paraskevi 15310
Athens
Greece

Hans C. Jessen
hans.jessen@uk.millwardbrown.com
Millward Brown
Olympus Avenue
Tachbrook Park
Warwick CV34 6RJ
UK

Abstract

The decision of whether or not to join the labour market is complex and often involves nonlinearities. However, most econometric decision models are linear and therefore may not be able to capture all aspects of the decision problem. In recent years several interesting Machine Learning methods have emerged for estimating nonlinear models in a relatively straightforward manner. It is shown here that some of these methods achieve significantly better classification performance than the standard linear model. Furthermore, a graphical approach is taken for interpreting the nonlinear models for the examined problem.

1 Introduction

The decision to participate in the labour market and how many hours to supply is influenced by a multitude of factors interacting in a nonlinear fashion. This fact has long been recognised in the literature [13], but econometric models of labour supply are often simplified to ease estimation and interpretation. For the labour force participation problem the models used are mostly logistic regression models and usually restricted to linear discrimination. This modelling approach stands in sharp contrast to other sciences, e.g. pattern recognition or medical diagnostic systems, where nonlinear models are widely used for real–world problems. The dogmatic acceptance of linear models for labour supply can also be seen by the fact that the adequacy of the generated models is seldom tested statistically and the possibility of nonlinearity rarely examined. This belief is partly explained by the opinion that:

> "the use of nonlinear expressions as arguments of the logistic function ...seldom adds anything of substance to the analysis" [5, page 9].

The work presented here aims to contribute to an emerging effort to change this situation and introduce the use of nonlinear modelling methods for labour force participation. In 1987, Thomas Mroz carried out an extensive evaluation of some popular econometric models in labour economics [16]. In estimating labour supply in terms of hours supplied, he concluded that estimates were very sensitive to the choice of functional form for the econometric model used. We perform a similar comparison of labour force participation models, and ask how well the standard procedures perform in comparison to new Machine Learning methods. Following standard practice, Mroz evaluated his models in terms of *in-sample fit*, i.e., performance on the training set, but in the past years work with new Machine Learning algorithms has suggested that in-sample fit is an unsuitable criterion for model comparison [28]. The reason being that many of these algorithms can in theory achieve a perfect fit in-sample, without having learned the true data generation process. The models investigated in this paper will therefore be evaluated in terms of predictive power on unseen cases.

Machine Learning methods, mainly Neural Networks (NNs), have been used in economics and other social sciences for some time [9], [25]. However, most work in economics using NNs has concentrated on time–series modelling. This is caused in part by a growing interest in time–series econometrics towards nonlinear models [19, page 5] but also by successful applications of NNs in Finance e.g. [23], [20]. There is, however, no reason why NNs should not be applied to *cross–sectional* data, i.e., data collected from a large sample on a specific point in time. Indeed, from both a theoretical and practical point of view, the estimation of NNs requires more data than the standard techniques, e.g. logistic regression models, and cross–sectional data sets are often much larger than time series ones.

Another Machine Learning tool, Decision Trees, have become popular in many research areas but so far not in economics. Apart from being nonlinear classifiers, Decision Trees use easy–to–interpret classification rules which sometimes can be directly compared to economic theory. This characteristic of Decision Trees has led us to include them in our study.

We estimate a simple logistic regression participation model based on the Mroz data set and compare the classification performance with NNs and Decision Trees. A graphical analysis is used to explain the classification results and to illustrate the *elasticities* of the corresponding probability surfaces, i.e., people's response to changes in the parameters that affect labour force participation, especially the wage level.

2 The Data

The data used is the same as that used by Mroz [16] and comes from a 1975 study of income dynamics conducted at the University of Michigan. It consists of 753 observations on married women of whom 428 worked for pay and 325 did not. In order to test the performance of the models, the data set was split into two parts.

The first part consisted of 600 observations and was used for the estimation of the models, while the second consisted of the remaining 153 observations and was used for testing the out–of–sample performance. This division was repeated randomly generating five pairs of sets.

The interviewed women included in the data set were questioned about several aspects of their social and economic background. In order to narrow down the number of relevant variables, it was decided to choose the variables for the comparison on the basis of optimising the logistic regression model. This was done by eliminating variables from the model if a Wald test showed that they were insignificant on a 10% level on the training set.[1] This procedure was carried out separately for each of the five random data sets but lead to a similar variable set in all five cases. This set consisted of the following variables:

- Number of children less than 6 years old.
- Number of children between 6 and 18 years old.
- The woman's age.
- The woman's level of education measured in years.
- The husband's annual working hours.
- The husband's hourly wage.
- The marginal tax rate facing the woman.
- The woman's previous labour market experience.
- The woman's hourly wage.
- The family income excluding any earnings by the woman.
- The husband's level of education.

The first ten variables were selected in all five experiments, while the last, the husband's education, was only used in two. Some preprocessing was also carried out involving the wage variable. In economic theory the wage rate is usually believed to be the most important variable in determining labour supply. Since no observations are available on this variable for non-working women, these were estimated using linear regression as suggested in [3]:

$$LWW_i = a + b_1 WA_i + b_2 WE_i + b_3 CIT_i + b_5 WA_i^2 + b_6 AX + U_i. \qquad (1)$$

The regression relates the log of the ith woman's wage rate (LWW) with her age (WA), level of education (WE), whether she lives in a big city or not (CIT), previous

[1] The wage variable was kept out of the selection process as it is considered the main economic variable.

labour market experience (AX) and the square of her age (WA^2).[2] The model was estimated using the working women in the data set and then used to estimate expected wage rates for the non-working women.

3 The models

3.1 Logistic discrimination

The main discrete-class models used in economics are static probability models which describe the outcome of a choice between S alternatives. Most work is dominated by logistic regression modeling. Logistic regression models are often used as arbitrary probability models but they can have a sound basis as the logistic form follows if the class–conditional densities are given by normal distributions with common covariance matrices [1, pp. 281–283].

The Logistic regression model relates the dependent variable y to the independent variables x in assuming that

$$\Pr(y = 1 \mid x) = \Lambda\left(\beta_0 + \sum_{i=1}^{I} \beta_i x_i\right) = p, \tag{2}$$

where $\Lambda(u) = 1/(1 + \exp(-u))$ denotes the logistic function. The unknown regression coefficients β_i are estimated from the data and are directly interpretable as log-odds ratios. Estimation is usually based on the maximum likelihood principle, i.e., on maximising the log–likelihood function

$$L(\beta) = \sum_{n=1}^{N} \left[y^{(n)} \log p\left(x^{(n)}, \beta\right) + \left(1 - y^{(n)}\right) \log\left(1 - p\left(x^{(n)}, \beta\right)\right)\right], \tag{3}$$

where $\left(x^{(n)}, y^{(n)}\right)$ denote the observed data from the nth individual, $n = 1, ..., N$. The maximum likelihood estimate $\hat{\beta}$ in this paper is obtained by applying the Fischer–scoring method, which is the default optimisation algorithm in the `Logit` procedure $SAS_{®}$version 6.11.

3.2 Feedforward Neural Networks

The most widely-used type of Neural Network is the feedforward NN with one hidden layer which is fully connected using logistic activation functions. Such Networks can be written as

$$\Pr(y = 1 \mid x) = \Lambda\left(W_0 + \sum_{j=1}^{J} W_j \times \Lambda\left(\sum_{i=0}^{I} w_{ij} x_i\right)\right) = p, \tag{4}$$

[2]Both the age and the squared age are usually included in wage estimations to allow for positive but diminishing age effects.

where $\Lambda(u) = 1/(1+\exp(-u))$ is the logistic function as in the logistic regression model. This group of NNs can therefore be seen as logistic regression models but with the simple linear index, $\sum_{i=1}^{I} \beta_i x_i$, replaced by a linear combination of other logistic functions. These "internal" functions are the hidden nodes. The unknown NN coefficients W_0, W_j, w_{ij} are typically found either by minimising the squared error criterion

$$SE(w) = \sum_{n=1}^{N} \left(y^{(n)} - p\left(x^{(n)}, w\right)\right)^2 \tag{5}$$

or by maximising the log–likelihood

$$L(w) = \sum_{n=1}^{N} \left[y^{(n)} \log p\left(x^{(n)}, w\right) + \left(1 - y^{(n)}\right) \log\left(1 - p\left(x^{(n)}, w\right)\right)\right]. \tag{6}$$

Squared error minimisation was chosen here using the Levenberg-Marquardt minimisation algorithm.[3] Because of the complexity of the Neural Network function the error surface often contains many local minima [14] and the choice of starting values for the optimization can have a large influence on the final solution. For that reason, several different starting values were tried but although some did lead to suboptimal solutions the optimization remained largely unaffected.

The main attraction of NN models is their function approximation capabilities. It has been shown that feedforward Networks can uniformly approximate any reasonable function with arbitrary precision, if the number of internal functions is unrestricted [6], [8], [12]. The main approximation results have also been extended to classifier–type mappings [6].

As NNs are highly nonlinear and used as approximation models rather than true representations of the data generation process, the distribution of the NN estimators becomes complicated. The theory of least squares for misspecified nonlinear regression models can be used for inference of least-square NNs and similar results have been derived for NNs estimated using the maximum likelihood criterion [29]. However, the issue of variable selection for the NNs did not arise in this work as the variable set was selected to maximise the fit of the logistic regression model. In order to facilitate the comparison of the models no further selection was done for the NN.

In the estimation of the NNs the so–called "early stopping rule", which has been linked to Ridge Regression [24], was used. As NNs have strong approximation capabilities they often overfit the data and thereby give rise to models which perform poorly on unseen observations. The "early stopping rule" avoids this by suggesting that the data set be split into 3 parts; training data, validation data and test data. The NN parameters are estimated using the training data until standard convergence criteria have been reached and at every iteration, $k = 1, ..., K$, the current parameter vector is saved. The K parameter vectors are used to calculate a fit on the validation data and the vector giving the best fit selected for further use. An unbiased estimate of the 'true' fit of the model can then be found by measuring the error on the

[3] Warren Sarle at the SAS Institute provided a set of SAS macros which considerably simplified the estimation.

test data. The use of the "early stopping rule" is generally thought of as a heuristic which might improve the out–of–sample fit but lacking any firm statistical justification. However, there are some links to Regularisation, which is used in optimisation theory to control the size of the parameter estimates. Shrinkage estimation and ridge regression can improve generalization in linear models by reducing the size of the estimated parameters as compared to the estimates that give the best fit in–sample. The "early stopping rule" performs a similar task in a NN. As the starting values in a NN estimation are usually small in size, the "early stopping rule" ensures that the estimated parameters cannot become "too large" and that they generally are smaller than at the best in–sample fit. We split the 753 observations into a training set consisting of 500 observations, a validation set of 100 observations and a test set of 153 observations. Finally, 3 hidden nodes were used as this gave the best performance on the validation data set.

3.3 Decision Trees

Symbolic Machine Learning techniques were initially invented to help in the construction of expert systems, but the process performed is essentially one of classification and can therefore be used for various classification tasks. The main advantage of those systems is the interpretability of the estimated models, which are to a large extent easily comprehensible and can be used for building transparent models. The main representatives of this field of work are systems generating *Decision Trees* (e.g. ID3 [21], CART [4]).

The symbolic Machine Learning algorithm, which has been used in this work is C4.5 [22], which is a recent version of the most widely used Decision Tree algorithm, ID3 [21].

Estimation of the Decision Trees is based on the *gain ratio* which is a function of the entropy measure known from information theory. For a two class problem the entropy for the whole data set is

$$entropy = -f_1 \times \log_2 f_1 - f_2 \times \log_2 f_2, \tag{7}$$

where f_i is the proportion of class $i = 1, 2$ in the data set. The estimation is stepwise where at each step the objective is to split the data set into subsets with lower entropy. The entropy is calculated on the subsets and denoted $entropy_j$ for the j-th subset. Finally, the *gain ratio* for a particular split is

$$gain\ ratio = \frac{entropy - \sum_j \frac{n_j}{n} entropy_j}{-\sum_j \frac{n_j}{n} \times \log_2 \left(\frac{n_j}{n}\right)}, \tag{8}$$

where n is the total number of observations, n_j the number of observations in subset j. The gain ratio is calculated for a number of possible splits of all independent variables and the one with the highest gain ratio is chosen. The split creates two new data subsets,[4] each of which is subjected to further splitting and the procedure

[4]The assumption that binary variables are used is made in this description. Numeric variables are always binarised by selecting an appropriate splitting threshold. In the case of multi-valued discrete variables, one subset for each value of the variable is created.

continues until a stopping criterion is met.[5] The process creates a tree–like structure as in figure 1.

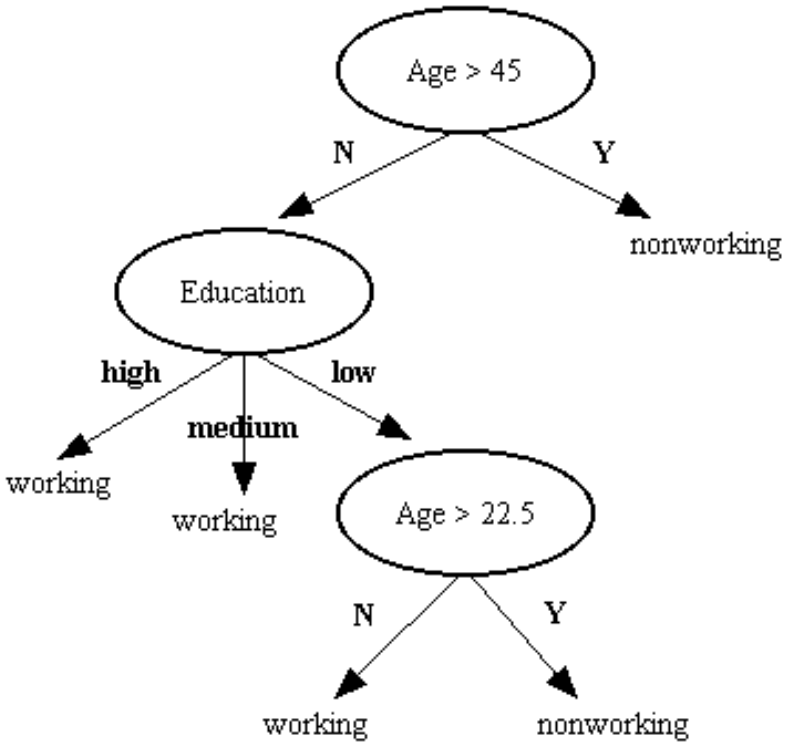

Figure 1: A simple decision tree.

Each non-leaf node of the tree can be seen as a question. Depending on the answer to this question for each particular case a different branch of the tree is traversed, asking further questions and eventually reaching a decision node, i.e., a leaf node in the tree.

In the presence of noise, and especially when many continuous independent variables are used, C4.5 may overfit the given data set [18]. In this case, the resulting tree can be pruned, i.e., reduced in size by collapsing some of the splits made. C4.5 provides a facility for pruning Decision Trees based on a heuristic. Decision Trees can be represented by a set of mutually exclusive rules which sometimes are preferable due to their interpretability. For example, the Tree in figure 1 can be translated to the following rule–set:

IF	$age > 45$	**THEN**	class = non-working
ELSE IF	$education = high$	**THEN**	class = working
ELSE IF	$education = medium$	**THEN**	class = working
ELSE IF	$education = low$		
AND	$age > 22.5$	**THEN**	class = non-working
ELSE			class = working

[5] One common stopping criterion is to set a lower limit on the number of observations in a subset.

4 Results and Analysis

4.1 Criteria

There is a tradition in time–series econometrics for conducting out–of–sample prediction tests but not in cross–sectional modelling. This is surprising since cross–sectional data is often acquired in large quantities and some researchers have recently called for more predictive tests [17]. As many new Machine Learning methods are capable of obtaining a perfect fit in–sample by overfitting the data, predictive tests become even more important in model evaluation. The main criterion used in these comparisons is the percentage of correct classifications, a measure known as the Count R^2 in econometrics and an unbiased estimator given random sampling and if measured on a hold–out sample.

The examined models provide the probability of whether or not a woman will work. Thus given a vector $\mathbf{x}$ of independent variables and the parameters of the model (θ) the fitted value is $p = \Pr(y = 1 \mid \mathbf{x}, \theta)$. In order to make a prediction based on this output, a decision rule is needed which assigns a class label according to the generated probability. Following standard procedures we use the maximum probability rule, which predicts the most probable class. For a dichotomous problem this corresponds to a threshold value $t = 0.5$, for which:

$$\text{Predicted class} = \begin{cases} 1 & \text{if } p > t \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Despite the easy comparison that it provides, there has been substantial criticism expressed about the Count R^2 measure [5], [10] and the maximum probability rule. The critics emphasise that if a sample is unbalanced (one group larger than the other) then the classification will be biased towards the larger group. However, as the Mroz data is relatively balanced this problem does not arise.

Finally, one methodological consideration is that a good fit does not necessarily prove the usefulness of an econometric model [2]. Often economists emphasise the interpretability of the model and its links with known and accepted economic theory. They are thus prepared to trade off predictive power for model interpretability. Realising the importance of this issue, the paper takes a few cautious steps towards the interpretation of the generated models by a graphical examination of the discrimination boundaries and elasticity surfaces for a simplified problem.

4.2 Classification performance

The first set of experiments involved the comparison of the models in terms of their pure predictive accuracy on the five test sets. Table 1 presents the results of these experiments for the different methods.The prediction accuracy of the Decision Tree models is very similar, but the pruned Tree has an advantage in terms of comprehensibility over the original Tree since on average the pruned Tree is 20% smaller than the unpruned one. The logistic regression models' classification performance is not very different from the Decision Trees but the NNs seem to perform much better than

the other methods.

Data Sets	Models			
	Logit	Neural Networks	C4.5 pruned	C4.5 unpruned
A	80.39	90.20	79.08	77.78
B	79.74	90.20	80.39	81.05
C	79.08	92.16	77.12	77.12
D	80.39	92.81	81.05	75.16
E	75.16	90.85	77.12	79.08
Average	78.96	91.24	78.95	78.04
Std. err.	2.19	1.19	1.82	2.20

Table 1: Performance as percentage correct over the 5 data sets.

As all models are estimated on the same data sets, it is possible to make a pairwise comparison between them using the McNemar test [26] which is a nonparametric test of the hypothesis of equality of classification success between two models (see appendix A). A summary table for the Z-test statistics when compared to the NN models is given in table 2.

Data Sets	Models		
	Logit	C4.5 pruned	C4.5 unpruned
A	-3.0	-3.57	-3.41
B	-3.66	-2.20	-2.56
C	-2.92	-4.74	-4.13
D	-3.65	-3.14	-4.70
E	-4.0	-2.19	-2.78
Average	3.45	3.17	3.52

Table 2: McNemar test statistics for comparisons with the NNs.

At a 5% level the H_0 of equal classification success with the NNs can be rejected for all five data sets in favour of the alternative hypothesis that the NNs classify significantly better than the other models which suggests that the NNs model aspects of the data generation process not captured by the other methods.

4.3 Extended logistic model

Most economists would be hesitant to include nonlinear transformations of the independent variables without some economic justification. However, in terms of potential fit, nonlinear terms in the logistic regression model are very important as without any

such terms the logistic regression model performs a simple linear discrimination. It is therefore expected that a logistic regression model using a polynomial instead of a linear index will be able to classify much better if there are nonlinearities in the classification problem and, given a large enough number of nonlinear transformations, the function approximation capability of the logistic regression model should begin to approach that of the Neural Network [11]. It is important to note, though, that the number of transformations grows exponentially with the number of independent variables and that this places a constraint on the practical flexibility of the logistic regression model whereas adding hidden units, which is the equivalent for NNs, increases only linearly the number of parameters to be estimated.

To test whether the relatively poor classification ability of the simple logistic regression model could be attributed to missing nonlinear transformations of the independent variables, the logistic regression model was re–estimated using the original explanatory variables, and their squares, cubes and all possible cross-products of the original variables. In order to reduce the resulting large number of explanatory variables (roughly 100), a backwards selection procedure was used where the variable with the least significant Wald statistic was eliminated until all variables left were significant on a 5% level.[6] Backwards selection is effective in finding a parsimonious model but since the first step consists of estimating a model with some 100 variables using only 600 observations, individual parameters estimates have high variance adding a spurious element to the variable selection. It should also be noted that had the number of 'base' variables been higher than the 11 used here the problem would have been much worse.

Data sets	Original	Extended
A	80.39	90.20
B	79.74	85.62
C	79.08	88.24
D	80.39	92.16
E	75.16	92.20
Average	78.95	89.68
Std. err.	2.19	2.80

Table 3: Performance of the Logit models using the original and the extended variable set.

Classification performance increased for all 5 models by an average of ten percentage points. To test the difference in classification performance between the extended logistic regression models and the NNs, the sign test [26] was used as the total number of misclassifications was too small for a Z approximation (see appendix A). On an approximate 5% level none of the hypotheses of equal classification success between

[6] In all data sets, some variables had almost perfect correlation with each other. This was partly due to the many categorical variables and meant that the Logit model could not be estimated. To be able to estimate the model the most extremely correlated variables were excluded *a priori*.

the extended logistic regression models and the NNs could be rejected although for data set B and C the acceptance was marginally. The conclusion is that even if the NNs did slightly better on average, than the extended logistic regression models, this difference is not statistically significant.

From a methodological point of view, the use of a large set of transformed variables can be criticised because it makes the interpretation of the logistic regression model difficult and a similar criticism applies to the NN. The following section takes a closer look at this problem.

5 Discrimination illustration

5.1 Decision Rules

A major benefit from using symbolic Machine Learning methods in the construction of econometric models is that their output can be interpreted as a set of rules which are easily understood. To illustrate this process, the following discussion examines the significance of one of the variables of the model, the woman's wage, and its relation to the propensity of the individual to work.

Labour theory would suggest that there is a positive relationship between the wage rates and labour force participation, i.e., the higher the wage, the more willing an individual would be to join the work force. This theory is partly validated by the constructed rules. For example, the following rule, developed for the first training set, says that women with higher than average wage rates[7] are likely to work.

Rule 1:
IF wage $>$ \$5.54
THEN class = *Working*

The same model, however, contains the following rule which counter-intuitively states that women with very low wage rates are expected to work.

Rule 2:
IF wage $<$ \$1.25
THEN class = *Working*

In order to explain this apparent contradiction, the composition of the observed sample set has to be examined.

Figure 2 presents the distribution of the wage rate for workers/non-workers, and shows that the reality is very different from what intuition might suggest. There seems to be a large number of women working at low wage rates and while the average wage for working women is higher than the average wage for non-working, the two distributions overlap to a large extent. It is known that part-time hourly wages tend to be lower than full-time ones [7] which might explain the existence of many low paid working women in the sample. However, the correlation between the

[7] Original 1975 prices are used. Average wage for workers=\$4.18.

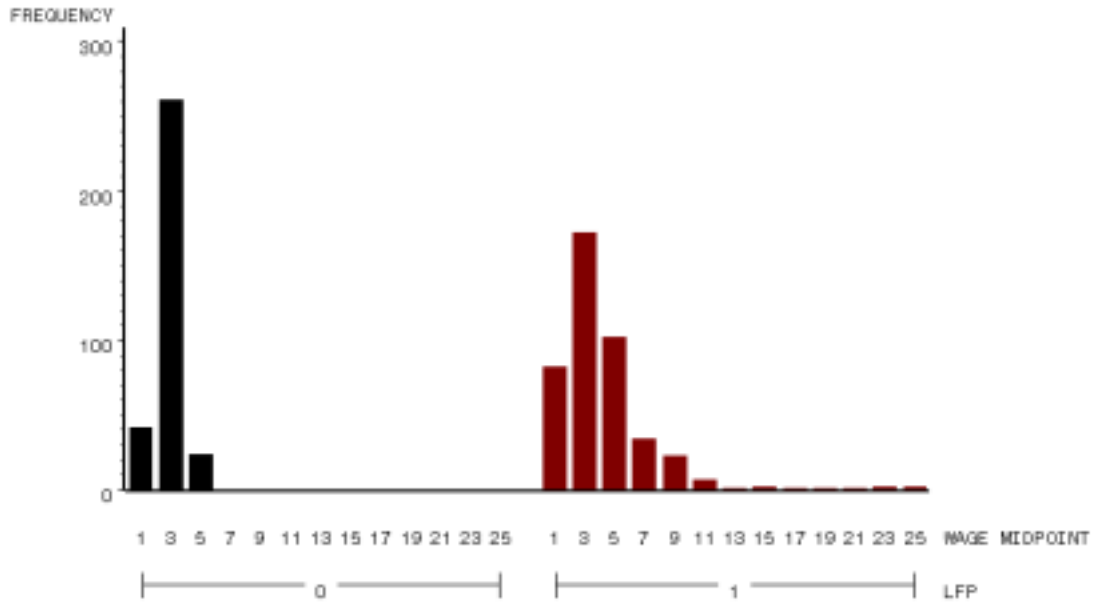

Figure 2: Distribution of wages for workers (LFP=1) and non-workers (LFP=0).

number of working hours and the hourly wage for working individuals in the data was very low and many individuals did work long hours on low pay.

In addition to the wage rate there are other factors which are taken into account in the decision making process and it seems that the combination of these factors provides a much better distinction between workers and non-workers. As an example, the following two rules, which correspond to specific groups of women, seem to be quite successful in identifying the corresponding type of individuals.

Rule 3:
IF wage $\leq$ \$2.15
AND Education $>$ 10 years
THEN class = *Working*

Rule 4:
IF \$2.15 $<$ wage $\leq$ \$3.45
AND Labour–experience $\leq$ 10 years
THEN class = *Non–Working*

The first of the two rules suggests that "not–low" educated women are likely to work even when their wage is not very high. This is expected to be true, especially for young women, who due to their qualifications expect better earnings in the future. Another influential factor, according to the second rule, is the labour experience of an individual, which is included in most of the generated rules. The argument is that women, who have been in the labour market for a long time, rarely decide to stop working unless they exit the labour market altogether. Thus, according to the latter rule, low experience decreases the likelihood of women working.

The above discussion shows that the decision process, involved in the examined problem, is quite complex. This is the reason why comprehensible nonlinear models, like Decision Trees and Rules, can provide a deeper understanding of the examined problem and might be used in the development/evaluation of economic theories.

To graphically illustrate how the Decision Tree separates workers from non-workers, we look at a simplified problem. Instead of using all variables, we select only the wage rate and the labour market experience and re-estimate the decision tree on this reduced labour force participation problem. An example using data set A is given in figure 3.[8] The graph shows the classification of observations in the holdout data where a non-worker is represented by a triangle and a worker by a cross. The variable space is split orthogonally to the axes by the rules and each rectangle corresponds to either a working or non-working classification.

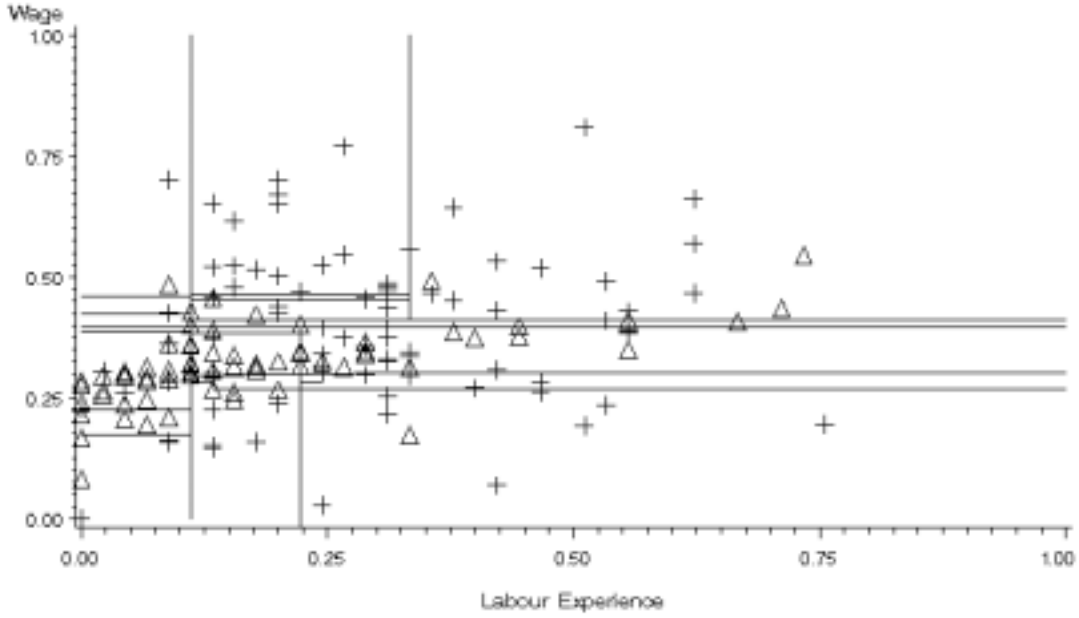

Figure 3: C4.5 discrimination lines. Each rectangle defines an area of common classification.

5.2 Logistic regression

As with C4.5, the discrimination of a simple logistic regression model can also be illustrated on the reduced labour force participation problem. The parameter estimates gave the following logistic model:

$$p_i = \Lambda\left(4.758AX_i + 1.130LWW_i - 1.132\right). \tag{10}$$

Using the maximum probability rule, a woman can be classified as working if:

[8] The equations and graphs in this section present the results for data set A. The other data sets gave similar results. Both variables were normalised to lie between zero and one.

$$\Pr(Working) = \Lambda\left(4.758AX_i + 1.130LWW_i - 1.132\right) > 0.5. \tag{11}$$

By taking logs and simplifying, the inequality becomes:

$$4.758AX_i + 1.130LWW_i - 1.132 > 0. \tag{12}$$

This inequality defines a discrimination line

$$LWW = -4.211AX + 1.002, \tag{13}$$

which splits the variable space into two parts. Alternatively, the logistic regression model could be interpreted as forming a weighted average of the wage rate and labour experience for each person and classifying women with a positive weighted average as workers and those with a negative weighted average as non-workers. Graphically, the discrimination line can be plotted as in figure 4.

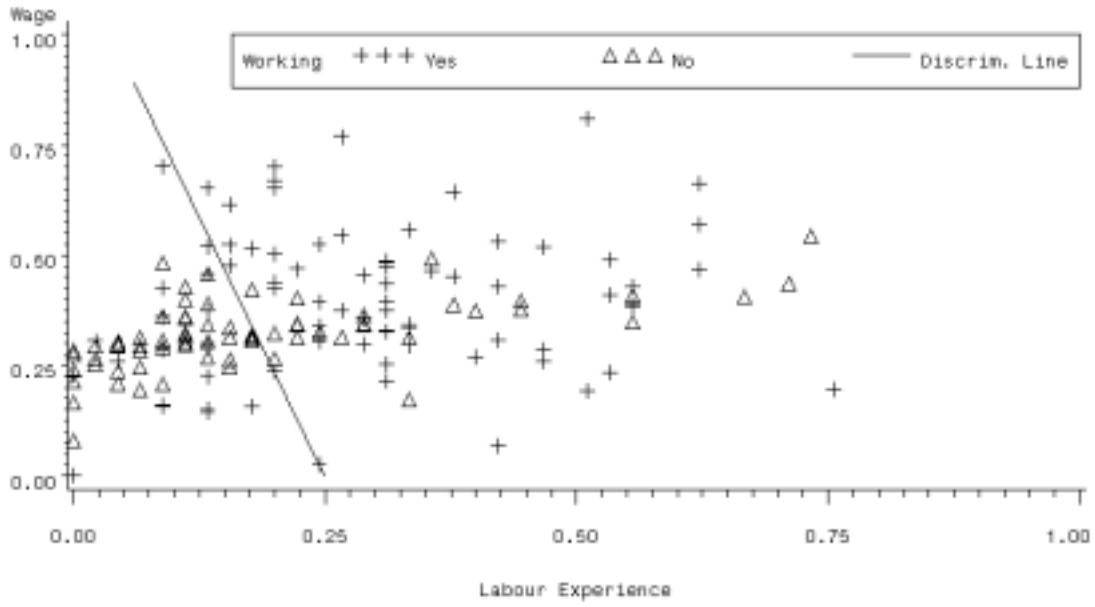

Figure 4: The Logit model's discrimination line. Observations above the line are classified as workers and observations below as non-workers.

The simple logistic regression model managed to classify 69.28% correctly on the test set. Comparing this to the 80.39% that was achieved using all eleven variables shows that the wage rate and previous labour experience are indeed important determinants for the labour force participation decision.

5.3 Extended Logistic regression

To capture some of the nonlinearity of the problem the simple logistic regression models were extended to incorporate nonlinear terms. These new terms included squares, cubes and the cross-products of the two exogenous variables generating an extended

logistic regression model.[9] The resulting parameter estimates gave the following extended logistic regression:

$$p_i = \Lambda\left(9.96AX_i - 17.26LWW_i + 55.05LWW_i^3 - 8.67AX_i^2 + 1.63\right). \quad (14)$$

As with the simple logistic regression model, the maximum probability rule can be used to classify a woman as a worker if

$$\Pr(W) = \Lambda\left(9.96AX_i - 17.26LWW_i + 55.05LWW_i^3 - 8.67AX_i^2 + 1.63\right) > 0.5. \quad (15)$$

By simplifying and taking logs equation 15 becomes

$$9.96AX_i - 17.26LWW_i + 55.05LWW_i^3 - 8.67AX_i^2 + 1.63 > 0. \quad (16)$$

This inequality defines an area in the LWW-AX space containing only working women with decision boundaries given in figure 5.

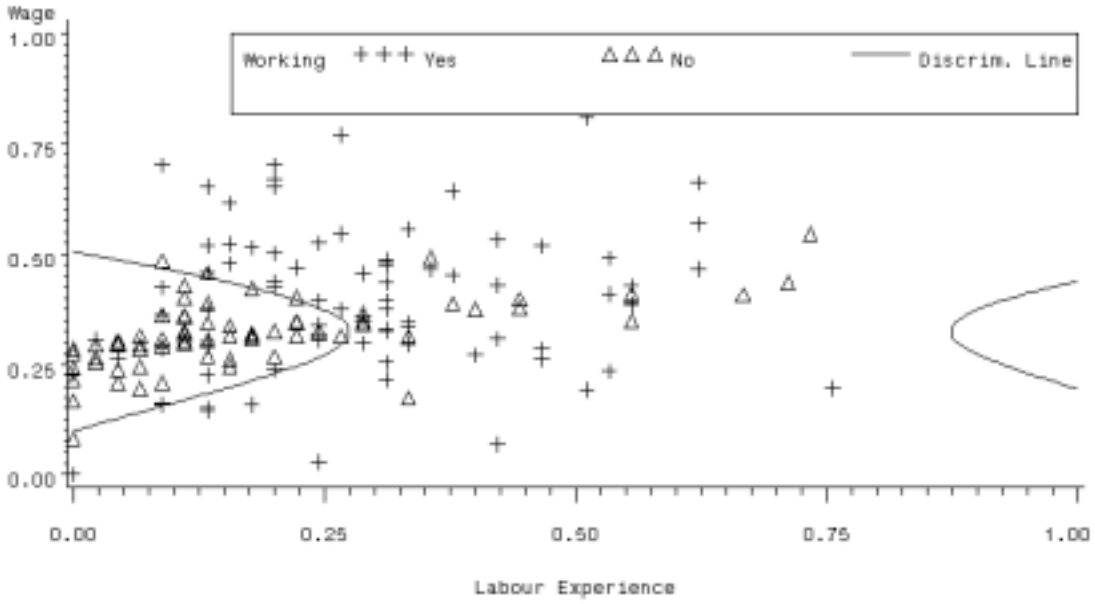

Figure 5: The optimal extended Logit model's discrimination lines. Observations inside the semi ellipse are classified as non–workers and observations outside as workers.

The discrimination line is very different from the simple logistic regression model's above. It is no longer a straight line but closer to half an ellipse. At the right-hand side there is a second discrimination line, which is due to the restricted functional form of the logistic regression model and does not correspond to any training data. The nonlinear discrimination of the extended logistic regression model classified 77.12% of the test set correctly which is about eight percentage points better than the simple logistic regression model and very close to the simple logistic regression model that

[9]Insignificant variables were eliminated using Wald tests as with the 11-variable problem.

used all variables. Therefore, even in this simplified labour participation problem, there seems to be a gain from including nonlinear terms in the logistic regression model.

5.4 Neural Networks

One of the criticisms of NN models is that they cannot be easily interpreted and tend to be used as "black boxes" which might be an argument against using them since it is undesirable to base policy conclusions on an incomprehensible model. The traditional way of interpreting econometric models is to investigate if there are any clear relationships between the exogenous variables and the endogenous variable by simply looking at the signs of the estimated parameters. This method cannot be applied here, due to the interaction between the hidden nodes and the complexity of the resulting model. The hidden nodes, i.e., internal logistic functions in this case, are not themselves modelling labour force participation but instead internally transform the data to make the classification problem easier for the main logistic function [27]. So giving an economic interpretation to the parameters of the hidden nodes is problematic. Instead we are examining graphically the effect of two important variables as done above for C4.5 and the logistic regression models.

The NN models were also reestimated for the reduced labour supply problem which lead to the following model for data set A:[10]

$$p_i = \Lambda\left(-9.391H_1 + 10.329H_2 + 2.633\right), \tag{17}$$

where

$$H_1 = \Lambda\left(-3.808AX_i - 8.746LWW_i - 1.871\right) \tag{18}$$

and

$$H_2 = \Lambda\left(-1.409AX_i - 13.486LWW_i - 5.594\right). \tag{19}$$

The condition that a woman will be classified as a worker, again using the maximum probability rule, is given by:

$$\Lambda\left(-9.391H_1 + 10.329H_2 + 2.633\right) > 0.5, \tag{20}$$

which can be simplified to:

$$-9.391H_1 + 10.329H_2 + 2.633 > 0. \tag{21}$$

The decision boundary for the Network is given in figure 6.

The Neural Network's decision boundary is very similar to the extended logistic regression model's. They are both elliptical in shape but the Network's boundary being more narrow and less symmetric. The classification ability of the Network is similar to the extended logistic regression model's with 78.43% correctly classified

[10] For the simplified problem two hidden nodes were used.

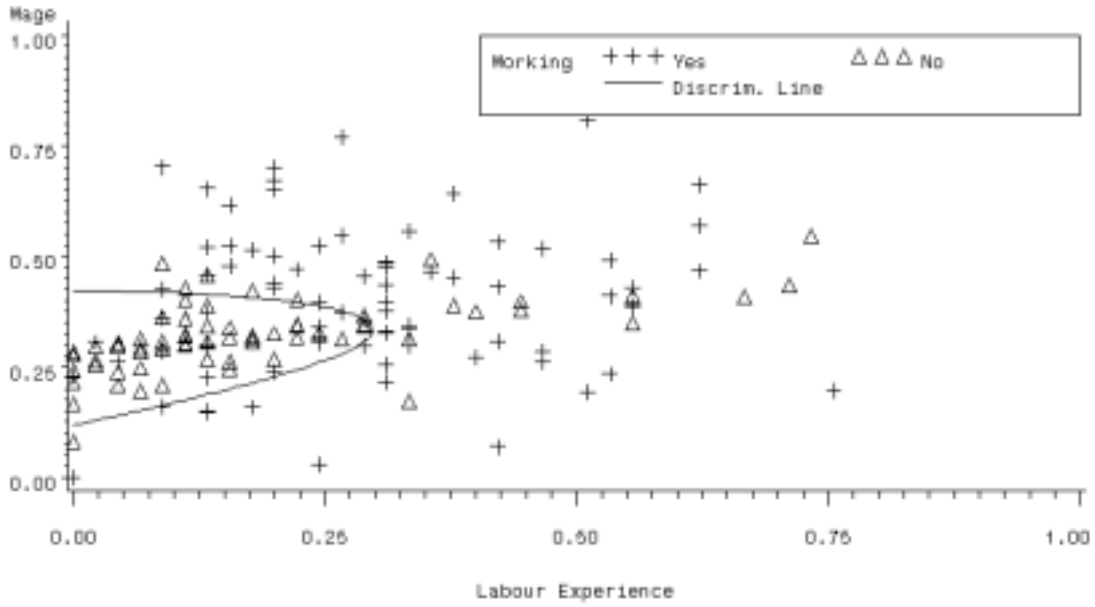

Figure 6: The Neural Network's discrimination line. Observations inside the semi ellipse are classified as non-workers. Observations outside as workers.

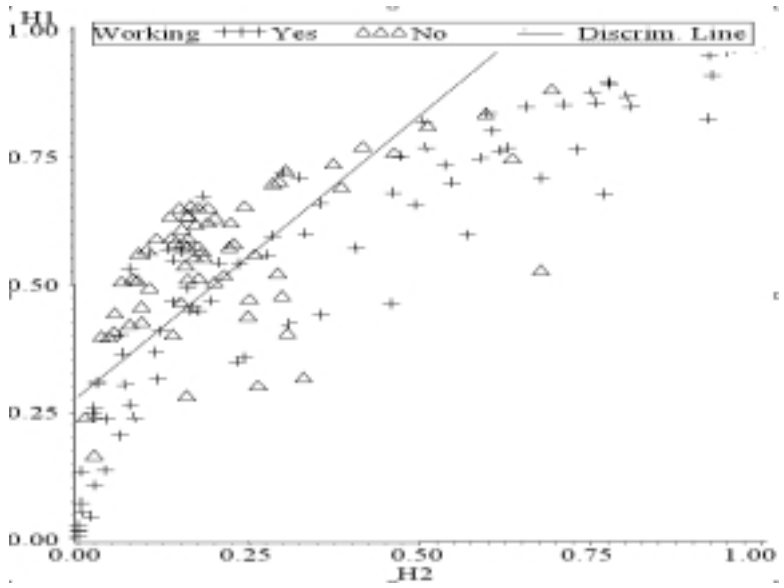

Figure 7: Main logistic function discrimination on the output from the two hidden nodes.

out–of–sample. In order to gain a better insight on the behaviour of the Network in this simplified model it is interesting to note that the main logistic function performs a linear discrimination in H_1-H_2 space. Since H_1 and H_2 are the outputs of logistic regression models themselves their values are confined to lie between zero and one. We can thus plot the output of H_2 against the output from H_1 and superimpose the discrimination line derived from equation 21. The graph is shown in figure 7. Comparing the data points in figure 7 with the initial scatterplot, it is clear that the hidden nodes have to some extent linearised the problem. The non-workers are now clustered on the upper left part of the graph, making it possible to separate most of the workers from the non-workers by a straight line. Figure 7 also shows why the hidden nodes sometimes are referred to as intermediate classifiers [27].

6 Elasticity of Labour Supply

One of the principal considerations in the analysis of micro–econometric models is their elasticity estimates, which give a measure of the effect of changes in the exogenous variables. When the dependent variable is a binary variable, a measure such as the quasi–elasticity $\eta_{ik} = \partial(p_i)/\partial(\log x_{ik})$, which measures the percentage point change in p_i given a percentage change in the value of the kth variable for the ith observation, is often used. This measure will clearly vary among individuals in the data set, depending on their characteristics.

Continuing the graphical analysis of the simplified two–variable problem, an illustration of the elasticities predicted by the logistic regression and the NN models can be achieved by plotting a 3-D graph of the underlying probability surface.

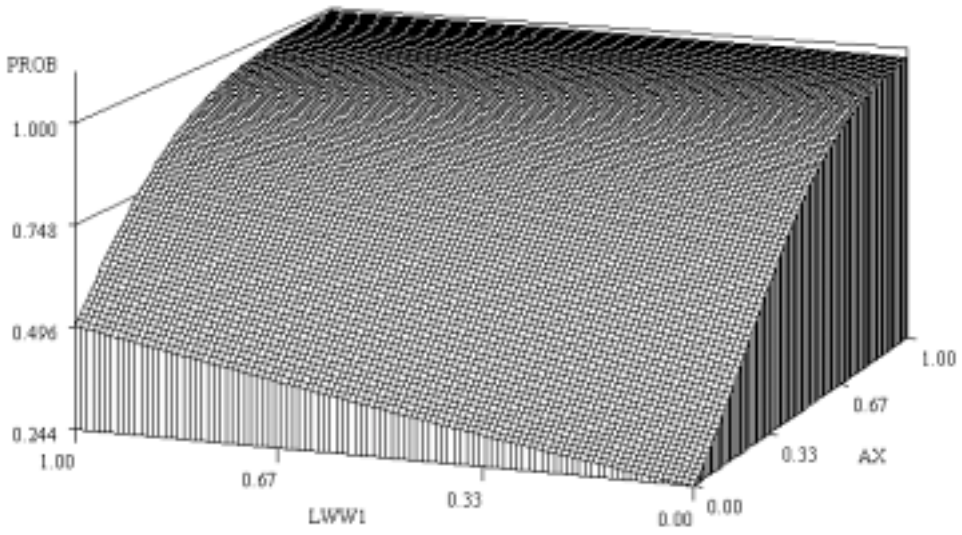

Figure 8: Probability surface for the Logit model.

Figure 8 shows the probability that a woman will be working for different combi-

nations of wage (LWW) and labour experience (AX) in the simple logistic regression model. The logistic surface gives a smooth transition from low to high probabilities. The graph shows that the wage elasticity is positive so that for every level of labour experience, higher wages will increase the probability of working and lower wages will decrease the probability which is what classical labour economic theory would suggest.

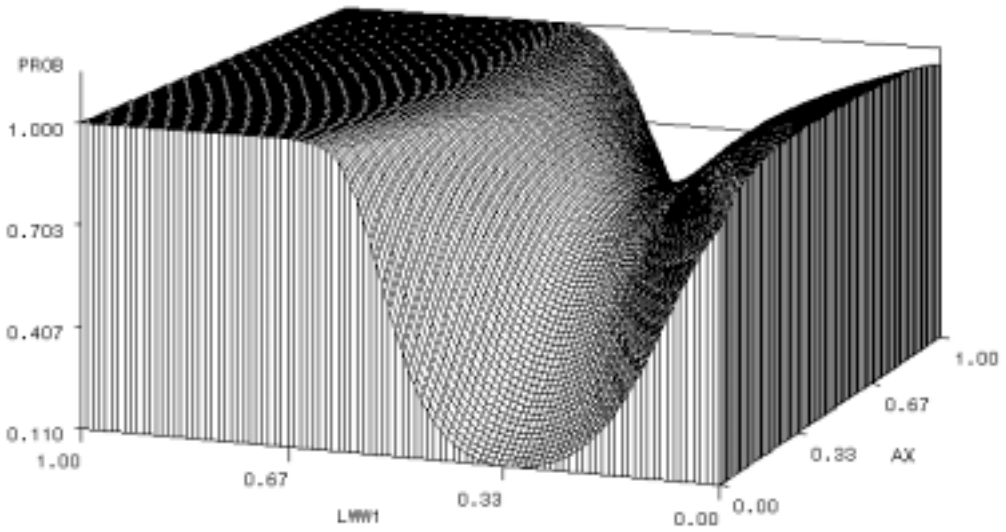

Figure 9: Probability surface for the extended Logit model.

As figure 6 showed, the discrimination line for the extended logistic regression model was different from the simple logistic regression model's and the inclusion of nonlinearities produced a semi-ellipsis which more accurately separated the data points. Figure 9 shows that instead of a smooth transition from low to high probabilities the probability surface now changes shape more abruptly and provides a much more localised fit to the data. Thus, small changes in expected wages for non-working women can have a large effect on the probability of working which implies very different elasticity estimates from the simple logistic regression model's. For every value of labour experience (AX), it is possible to increase the probability of working by either increasing or decreasing wages sufficiently. Such a response is not predicted by standard economic theory and seems counter-intuitive. This result, combined with the higher out-of-sample classification accuracy of the nonlinear model, suggests that there is an unusual nonlinear pattern to the data that cannot be captured by the simple logistic regression. The reason for the unexpected surface is again due to the wage distribution. As illustrated in the discussion of the Decision Tree there are a number of women working for very low wages. These women are often low skilled and some are working a substantial number of hours. The typical non-worker is a person with average earnings but low experience. The simple logistic regression model always generates strictly monotonic probability surfaces due to its functional form

and is therefore not able to identify differences in working patterns for subgroups in the data. It 'merges' such differences which in this case meant assigning all low experienced low earners to be non-workers contrary to what was observed in the data. This is obviously problematic if the aim of the analysis is to understand labour participation elasticities.

The 3-dimensional surface for the NN is similar to the surface for the extended logistic regression model but with slightly steeper slopes. The surface is shown in figure 10.

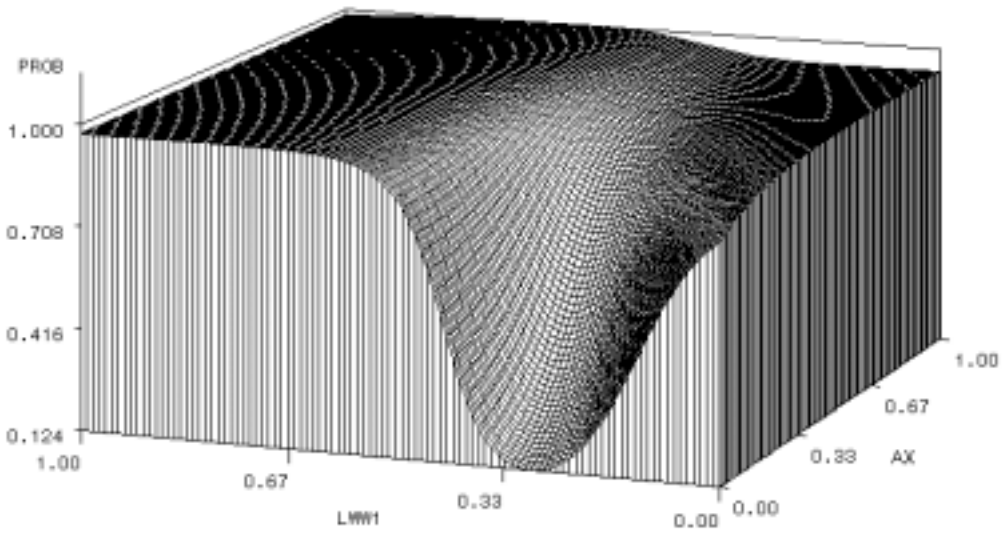

Figure 10: Probability surface for the Neural Network.

For the Neural Network no explicit nonlinear data transformations were needed, as the model itself performs a nonlinear approximation of the data. The type and extent of nonlinearity that can be achieved by the Network depends on the complexity of its architecture and the transfer functions in the hidden and output nodes. Despite the very simple architecture that we used in this analysis, the same extent of nonlinearity as the nonlinear logistic regression model was achieved.

Finally, C4.5 generates rules that correspond to localised threshold functions. This characteristic does not allow elasticity interpretation which is not surprising as the algorithm was not intended for function approximation but for classification only.

7 Conclusion

The decision to participate in the labour market is highly complex and often involves nonlinearities, but most econometric work still uses a linear or semi–linear framework to model this choice. In recent years there has been growing interest in Machine Learning techniques and many new exiting models have been developed that are capable of modelling binary choice variables nonlinearly. In this paper, two types of

Machine Learning techniques (Decision Trees and NNs) were compared with a standard participation model (logistic regression) and it was found that the NNs were able to classify significantly better out–of–sample than the other methods. It was further shown that if enough nonlinear terms are included in the logistic regression model it is possible to get the same performance as for the NNs. However, because the growth in these terms is exponential it quickly becomes impractical to estimate a highly nonlinear logistic regression model. For a NN the nonlinearity comes from a more complex functional form which makes it more suitable for nonlinear approximation.

Using the simple but standard logistic regression model is equivalent to imposing strict assumptions on the probability surfaces and thereby the elasticities. The economic example showed that a subgroup of women with low earnings would effectively be ignored by a linear specification. Whether this is desirable depends on the purpose of the analysis. It might sometimes be preferable to focus on general patterns in the data but if the interest is in a close approximation of elasticities it would be beneficial to use a more flexible functional form that would allow for non–monotonic probability surfaces.

The analysis has focused on a graphical interpretation which works well in low dimensions. For high-dimensional problems one approach would be to investigate wage effects for typical or average individuals by keeping all their characteristics fixed and letting the wage vary. Such a 'pertubation' analysis is simple and essentially equivalent to the graphical analysis used here.

We believe that since these new Machine Learning methods are relatively easy to use they should be seen as at least complementary to a standard logistic regression analysis. Nonlinear models are essential for capturing complex real–word phenomena, like the labour force participation decision, and these methods have proved to serve this purpose very efficiently in a variety of disciplines. The nonparametric McNemar test can be used as an indicator of potential misspecification and if significant the standard logistic regression results should be interpreted with care. We would also suggest that Decision Trees be used, not only as classification models, but as a tool to help economic theory construction as the plain text rules are easy to interpret and could spark new theoretical ideas.

Appendix A

If the classification performance of two models on the same data set is compared the information can be summarised in a contingency table

	Model A		
Model B	True	False	Total
True	a	b	$a+b$
False	c	d	$c+d$
Total	$a+c$	$b+d$	N

Table 4: Data on two outcomes from matched pairs

The McNemar [15] test is a variation of the standard sign test [26] used when $b+c$ is large ($b+c>20$). If Model A classify better than Model B then b would be less than c and the McNemar test statistic for the hypothesis

$$\begin{aligned} H_0 &: \ b=c \\ H_1 &: \ b<c \end{aligned} \tag{22}$$

becomes:

$$T=\frac{b-c}{\sqrt{b+c}}\sim Z\left(0,1\right). \tag{23}$$

For $n<20$ the standard sign test is used instead with $p=1/2$ and $n=b+c$.

References

[1] T. Amemiya. *Advanced econometric*. Blackwell, 1985.

[2] D. A. Belsley. Modelling and forecasting reliability. *International Journal of Forecasting*, 4:427–447, 1988.

[3] E. Berndt. *The Practice of Econometrics: Classic and Contemporary*. Addison-Wessley, 1991.

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[5] J. Cramer. *The Logit Model*. Edward Arnold, 1991.

[6] G. Cybenko. Approximation by superposition of a sigmoidal function. *Math. Controls Signals Systems*, 2:303–314, 1989.

[7] R. B. Freeman. The minimum wage as a redistributive tool. *Economic Journal*, 106:639–649, 1996.

[8] K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2:183–192, 1989.

[9] G. Garson. A comparison of neural network and expert systems algorithms with common multivariate procedures for analysis of social science data. *Social Science Computer Review*, 9:399–435, 1991.

[10] W. H. Greene. *Econometric Analysis*. Macmillan Publishing Company, 1993.

[11] M. H. Hassoun. *Fundamentals of Artificial Neural Networks*. MIT Press, 1995.

[12] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[13] M. R. Killingsworth. *Labor supply*. Cambridge University Press, 1983.

[14] J. F. Kolen and J. B. Pollack. Backpropagation is sensitive to initial conditions. *Complex Systems*, 4:269–280, 1990.

[15] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947.

[16] T. Mroz. The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. *Econometrica*, 55(4):765–799, 1987.

[17] A. C. Nakamura, M. Nakamura, and H. O. Duleep. Alternative approaches to model choice. *Journal of Economic Behavior and Organization*, 14:97–125, 1990.

[18] T. Niblett and I. Bratko. Learning Decision Rules in Noisy Domains. In M. Bramer, editor, *Research and Development in Expert Systems*, volume 3, pages 25–34. Cambridge: Cambridge University Press, 1987.

[19] L. Oxley and C. Roberts. Introduction. In L. Oxley, editor, *Surveys in Econometrics*. Blackwell, 1995.

[20] F. Pau and L. Gianotti. *Economic and financial knowledge-based processing.* Springer-Verlag, 1993.

[21] J. Quinlan. Learning Efficient Classification Procedures and Their Application to Chess End Games. In R. Michalski, J. Carbonell, and T. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*, pages 463–482. Morgan Kaufmann, 1983.

[22] J. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, 1993.

[23] P. Refenes. *Neural Networks in the Capital Markets.* John Willey and Sons, 1995.

[24] W. S. Sarle. The tnn macros for feedforward neural networks. Technical report, SAS Institute, 1996.

[25] P. Schrodt. Prediction of interstate conflict outcomes using a Neural Network. *Social Science Computer Review*, 9:359–380, 1991.

[26] P. Sprent. *Applied nonparametric statistical methods.* Chapman and Hall, 2nd edition, 1993.

[27] A. Webb and D. Lowe. The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis. *Neural Networks*, 3:367–375, 1990.

[28] S. M. Weiss and C. A. Kulikowski. *Computer systems that learn.* Morgan Kaufmann, San Mateo, CA, 1990.

[29] H. White. *Estimation, inference and specification analysis.* Cambridge University Press, 1992.