Journal of the Operational Research Society (2005) 56, 1063–1071

www.palgrave-journals.com/jors

Hurdle models of loan default

PG Moffatt*

University of East Anglia, Norwich, UK

Some models of loan default are binary, simply modelling the probability of default, while others go further and model the extent of default (eg number of outstanding payments; amount of arrears). The double-hurdle model, originally due to Cragg (*Econometrica*, 1971), and conventionally applied to household consumption or labour supply decisions, contains two equations, one which determines whether or not a customer is a potential defaulter (the 'first hurdle'), and the other which determines the extent of default. In separating these two processes, the model recognizes that there exists a subset of the observed non-defaulters who would never default whatever their circumstances. A Box-Cox transformation applied to the dependent variable is a useful generalization to the model. Estimation is relatively easy using the Maximum Likelihood routine available in STATA. The model is applied to a sample of 2515 loan applicants for whom loans were approved, a sizeable proportion of whom defaulted in varying degrees. The dependent variables used are amount in arrears and number of days in arrears. The value of the hurdle approach is confirmed by finding that certain key explanatory variables have very different effects between the two equations. Most notably, the effect of loan amount is strongly positive on arrears, while being U-shaped on the probability of default. The former effect is seriously under-estimated when the first hurdle is ignored.

Journal of the Operational Research Society (2005) **56,** 1063–1071. doi:10.1057/palgrave.jors.2601922
Published online 12 January 2005

Keywords: double hurdle model; credit scoring; loan default

Introduction

A feature of many models of loan default, for example straightforward binary or censored data models, is that the process which results in non-default is assumed to be the same as that which determines the extent of default. Thus, for example, if a particular borrower characteristic is known to have a positive effect on the extent of default, then a very high value of this characteristic would inevitably lead to the prediction of default for such a borrower. While such assumptions may turn out to hold, there is no reason to expect this *a priori*. One reason why such an assumption might fail is that there may exist a proportion of the population of borrowers who would, out of principle, never default under any circumstances.

Such considerations lead us to a class of model in which the event of a borrower being a potential defaulter, and the extent of default by that borrower, are treated separately. This type of model is known as the 'double-hurdle' model and is originally due to Cragg.[1] As the name suggests, the model assumes that a borrower must cross two hurdles in order to be a defaulter. Those who fall at the fist hurdle are the borrowers to whom we refer in this paper as 'never-defaulters'. Passing the first hurdle places a borrower in the class of 'potential defaulter'. Whether a potential defaulter actually defaults then depends on their current circumstances; if they do default, we say that they have crossed the second hurdle. Both hurdles have equations associated with them, incorporating the effects of borrower characteristics and circumstances. Such explanatory variables may appear in both equations or only in one. Most importantly, a variable appearing in both equations may have opposite effects in the two equations.

The double-hurdle model has been applied at least once in the credit scoring literature, by Dionne *et al*,[2] whose dependent variable is the number of non-payments. The model has been applied in a variety of other contexts such as cigarette consumption by individuals,[3] where it is assumed, justifiably, that a proportion of the population would never smoke whatever circumstances they found themselves in.

The model is heavily parametric in character, the error terms of both equations typically being assumed to be normally distributed. Such assumptions may be costly when the data does not fit, resulting in inconsistent estimation. Finding ways of accommodating these assumptions is therefore paramount. Transforming the dependent variable is one possibility. The logarithmic transformation is clearly inappropriate since the dependent variable contains zeros, but the Box-Cox transformation, which in fact includes the log transformation as a limiting case, is feasible. The Box-Cox double-hurdle model was introduced recently by Jones and Yen,[4] and the same generalization is usefully applied in this paper.

Another direction in which the model could be generalized is by allowing a non-zero correlation between the error terms

*Correspondence: PG Moffatt, School of Economics, University of East Anglia, Norwich NR4 7TJ, UK.
E-mail: p.moffatt@uea.ac.uk

of the two equations. This leads to the 'double-hurdle model with dependence', which has been analysed in some detail by Smith.[5] This approach has not been followed here because Smith's[5] findings are to the effect that the correlation parameter is poorly identified even if the parameter is large in magnitude, and that assuming this parameter is zero allows more profitable generalisations in different directions.

Estimation of the double-hurdle model and its variants is possible using the ML routine available in the econometric software STATA.[a] Indirect evidence of the recent popularity of the model is McDowell's[6] advice 'from the (STATA) help desk' on the programming required to estimate models of this sort. Two STATA programmes used to estimate the most general of the models described in this paper are presented in the Appendix.

The next section is concerned with the theory underlying the double-hurdle and related models, while the subsequent section describes the data sample, which consists of 2515 loan applicants for whom loans were approved, a sizeable proportion of whom defaulted in varying degrees. Further section reports model estimates and interprets the results, in particular deducing estimates of the proportion of the population who are in the 'never-default' category, and the last section concludes.

Double-hurdle model and variants

Tobit

First, consider the linear specification:

$$\begin{aligned} y_i^* &= x_i'\beta + u_i, \quad i = 1, \ldots, n \\ u_i &\sim N(0, \sigma^2) \end{aligned} \tag{1}$$

where y_i* is the latent variable representing borrower i's propensity to default, x_i is the vector of borrower characteristics relevant in explaining the extent of default, β is the corresponding vector of parameters to be estimated, and u_i is the homoscedastic, normally distributed error term. Let y_i be the actual default (eg amount in arrears). Since actual default cannot be negative, the relationship between y_i* and y_i is

$$y_i = \max(y_i^*, 0) \tag{2}$$

Equation (2) gives rise to the standard censored regression ('tobit') model estimation of which is routinely available in econometric software packages. The log-likelihood function for the tobit model is

$$\begin{aligned} \text{Log}\, L = &\sum_0 \ln\left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \\ &+ \sum_+ \ln\left[\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right] \end{aligned} \tag{3}$$

in which '0' indicates summation over the zero observations in the sample, while '+' indicates summation over positive observations. $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cdf and pdf, respectively.

p-*tobit*

A possibly over-restrictive feature of the tobit model described in the previous section is that it only allows one type of zero observation, and the implicit assumption is that zeros arise as a result of borrower circumstances. The generalization to the tobit model, which is of interest in this paper, assumes the existence of an additional class of borrower who, perhaps as a point of principle, would never default whatever their circumstances.

In the first instance, let us simply assume that the proportion of the population who are potential defaulters is p, so that the proportion of the population who would never default is $(1-p)$. For the former group, the tobit model applies, while for the latter group, the extent of default is automatically zero.

This assumption leads to the *p-tobit* model, originally proposed by Deaton and Irish[7] in the context of household consumption decisions, where they were essentially allowing for a class of 'abstinent' consumers for each good modelled. The log likelihood function for the p-tobit model is

$$\begin{aligned} \text{Log}\, L = &\sum_0 \ln\left[1 - p\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \\ &+ \sum_+ \ln\left[p\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right]. \end{aligned} \tag{4}$$

Maximizing (4) returns an estimate of the parameter p, in addition to those of β and σ obtained under tobit.

Double hurdle

Since the class of borrowers who would never default is the focus of this analysis, it is desirable to investigate which types of borrower are most likely to appear in this class. With this in mind, we assume that the probability of a borrower being in the said class depends on a set of borrower characteristics. In other words, we shall generalize the p-tobit model of the previous section by allowing the parameter p to vary according to borrower characteristics. This generalization leads us to the 'double-hurdle' model.

As the model name suggests, borrowers must cross two hurdles in order to be default. The 'first hurdle' needs to be crossed in order to be a potential defaulter. Given that the borrower is a potential defaulter, their current circumstances then dictate whether or not they do in fact default—this is the 'second hurdle'.

The double-hurdle model contains two equations. We write:

$$\begin{aligned} d_i^* &= z_i'\alpha + \varepsilon_i \\ y_i^{**} &= x_i'\beta + u_i \\ \begin{pmatrix}\varepsilon_i \\ u_i\end{pmatrix} &\sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\ 0 & \sigma^2\end{pmatrix}\right] \end{aligned} \tag{5}$$

Note from the diagonality of the covariance matrix that the two error terms are assumed to be independently distributed.

The first hurdle is then represented by

$$\begin{aligned} d_i &= 1 \text{ if } d_i^* > 0 \\ d_i &= 0 \text{ if } d_i^* \leqslant 0 \end{aligned} \tag{6}$$

The second hurdle closely resembles the tobit model (2):

$$y_i^* = \max(y_i^{**}, 0) \tag{7}$$

Finally, the observed variable, y_i, is determined as

$$y_i = d_i y_i^* \tag{8}$$

The log-likelihood function for the double-hurdle model is

$$\begin{aligned} \text{Log}\, L = &\sum_0 \ln\left[1 - \Phi(z_i'\alpha)\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \\ &+ \sum_+ \ln\left[\Phi(z_i'\alpha)\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right] \end{aligned} \tag{9}$$

Figure 1 is useful for understanding the model defined in (5)–(8). The concentric circles are contours of the joint distribution of the latent variables d^* and y^{**}; they are circles (rather than ellipses) as a consequence of the assumed independence between the two error terms. These circles are centred on the point $(z_i'\alpha,\ x_i'\beta)$, so that the whole distribution moves around with changes in the values taken by the explanatory variables. The likelihood contribution associated with non-default (ie the first term in square brackets in (9)) is represented by the probability mass under the L-shaped region comprising the north-west, south-west and south-east quadrants of the graph; the contribution associated with a default (the second bracketed term in (9)) is represented by a thin strip of the probability mass within the north-east quadrant at the value of the observed default (two such values are depicted in the diagram).

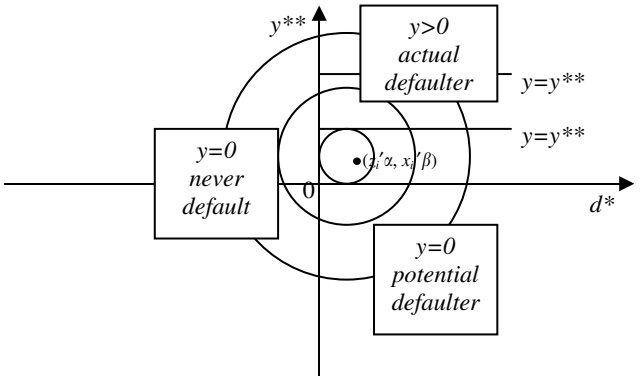

Figure 1 The relationship between latent (d^* and y^{**}) and observed (y) variables in the double-hurdle model.

Box-Cox double hurdle

Often the dependent variable under analysis shows a strong positive skew. In this situation, it is tempting to apply the logarithmic transformation. This is partly because all of the models outlined in this section rely heavily on the assumption of normality in the error terms: without normality the property of consistency of the MLE fails to hold. However, the logarithmic transformation is clearly inappropriate due to the presence of the zero observations in the sample, especially in the present situation in which the zeros are the focus of the analysis.

Instead, we follow Jones and Yen[4] by applying the Box-Cox transformation, defined as

$$y^{\text{T}} = \frac{y^\lambda - 1}{\lambda}, \quad 0 < \lambda \leqslant 1 \tag{10}$$

Note that the Box-Cox transformation (10) includes as special cases a straightforward linear transformation ($\lambda = 1$), and the logarithmic transformation ($\lambda \to 0$), but normally we would expect the parameter λ to be somewhere between these limits.

The transformation (10) can be applied to any of the models previously outlined in this section. When it is applied to the dependent variable in the double-hurdle model, we obtain the Box-Cox double-hurdle model, defined as follows (where the latent variables d^* and y^{**} are defined as in (5) above):

First hurdle:

$$\begin{aligned} d_i &= 1 \quad \text{if } d_i^* > 0 \\ d_i &= 0 \quad \text{if } d_i^* \leqslant 0 \end{aligned} \tag{11}$$

Second hurdle:

$$y_i^{*\text{T}} = \max\left(y_i^{**\text{T}}, -\frac{1}{\lambda}\right) \tag{12}$$

Observed y^{T}:

$$\begin{aligned} y_i^{\text{T}} &= y_i^{*\text{T}} \quad \text{if } d_i = 1 \\ y_i^{\text{T}} &= -\frac{1}{\lambda} \quad \text{if } d_i = 0 \end{aligned} \tag{13}$$

Note that the lower limit of the transformed variable is $-1/\lambda$ rather than zero.

The log-likelihood function for the Box-Cox double-hurdle model is

$$\begin{aligned}\mathrm{Log}\,L = &\sum_{0} \ln\left[1-\Phi(z_i'\alpha)\Phi\left(\frac{x_i'\beta+1/\lambda}{\sigma}\right)\right] \\ &+\sum_{+} \ln\left[\Phi(z_i'\alpha)y_i^{\lambda-1}\frac{1}{\sigma}\phi\left(\frac{y_i^{\mathrm{T}}-x_i'\beta}{\sigma}\right)\right]\end{aligned} \tag{14}$$

Note that (14) is not very different from the log-likelihood function for the double-hurdle model (9). One important difference is that the use of y^{T} in place of y in the final term requires a Jacobian term $y^{\lambda-1}$ to be included.

The STATA code required to maximize the log-likelihood function (14) is given in Appendix A.

Modelling interval data

The models described in the previous sections may be used in situations where the extent of default variable is observed exactly. It is sometimes the case that the variable is observed in interval form. An example that we encounter in the Results is the number of days in arrears, which is only observed up to 30-day intervals (0–30 days; 31–60 days; 61–90 days; …; more than 180 days).

The manner in which we deal with data in this form resembles Stewart's[8] grouped normal regression model. Let $I_j=[a_j, b_j]$ be the jth interval, $j=1, \ldots, J$. The nature of the data is that while we do not know the value of y_i for a given borrower i, we do know which of the J intervals contains y_i. The likelihood contribution corresponding to a defaulter is therefore the probability of falling in the observed interval, and the full log-likelihood for the Box-Cox double-hurdle model for interval data is:

$$\begin{aligned}\mathrm{Log}\,L = &\sum_{0} \ln\left[1-\Phi(z_i'\alpha)\Phi\left(\frac{x_i'\beta+1/\lambda}{\sigma}\right)\right] \\ &+\sum_{+} \ln\left[\Phi(z_i'\alpha)\sum_{j=1}^{J} I(y_i\in I_j)\left(\Phi\left(\frac{b_j^{\mathrm{T}}-x_i'\beta}{\sigma}\right)\right.\right. \\ &\left.\left.-\Phi\left(\frac{a_j^{\mathrm{T}}-x_i'\beta}{\sigma}\right)\right)\right]\end{aligned} \tag{15}$$

where $I(\cdot)$ is the indicator function, taking the value one if the statement in parentheses is true, zero otherwise, and a_j^{T} and b_j^{T} are the results of applying the Box-Cox transformation (10) to a_j and b_j, respectively.

The STATA code required to maximize the log-likelihood function (15) is given in Appendix B.

Data

The data set comprises 2515 loans approved between May and October 2000. The performance variables were created at the end of June 2002, representing an outcome period of around 2 years. Of the 2515 loans in the sample, 1188 were in arrears at this time, while the remaining 1327 were up to date.

An important issue needing to be addressed is that the sample selection criterion was not random: while 100% of defaulters appear in this sample, only 10% of non-defaulters appear. In order to remove the effects of this selection bias, data on the 1327 borrowers who were not in arrears was reproduced 10-fold, giving a sample size of 14 458. Of this expanded sample, only 8.2% are in arrears, which importantly corresponds to the proportion of all approved loans which are in arrears. The sample used in estimation is of size slightly less than 14 458, due to a small number of missing values in variables appearing in the models.

Although we shall be estimating our models for two different measures of the extent of arrears, we shall focus on one: amount in arrears. To give a feel for the distribution of this variable over the 1188 defaulters, a histogram is shown in Figure 2. As expected, this variable shows a strong positive skew.

As explained previously the very long tail to the right brings into doubt the validity of the assumption of normality of the error term, which is necessary for consistency of the MLE in each of those models. We then suggested the use of the Box-Cox transformation to address this problem. Figure 3 shows the distribution of arrears after applying the Box-Cox transformation (10) with the parameter λ set to 0.79 (which is the estimate of this parameter in our final model—see next section). As expected, the transformation has the effect of considerably reducing the positive skew.

It is also useful at this stage to use non-parametric analysis to investigate the effects of selected explanatory variables.

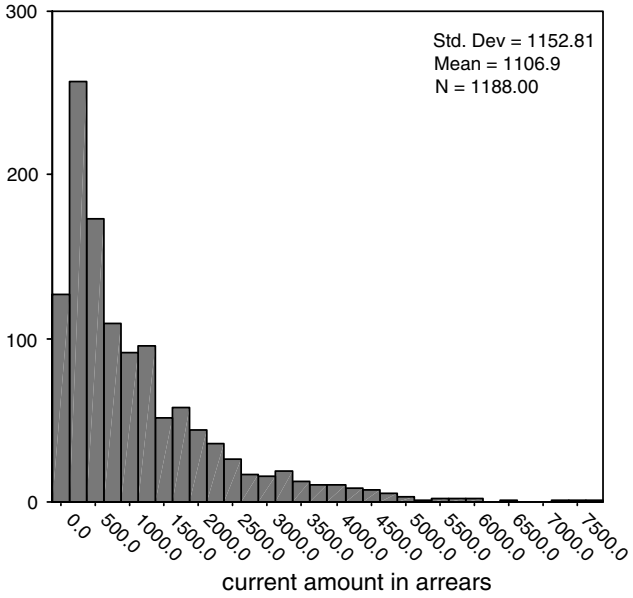

Figure 2 A histogram of current amount in arrears (£).

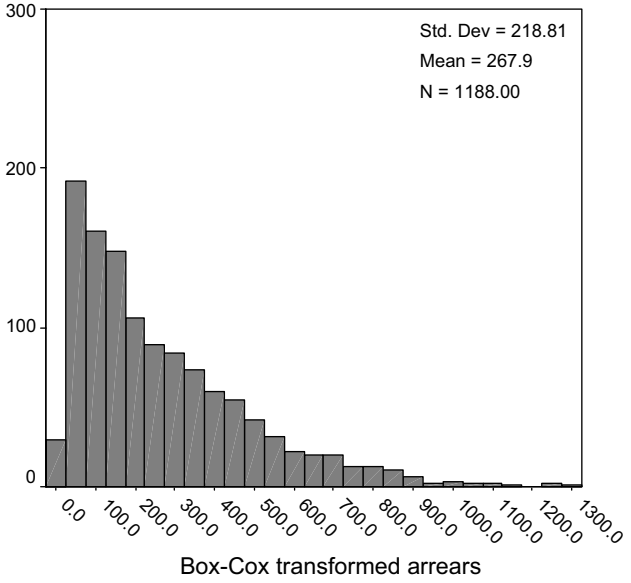

Figure 3 Histogram of arrears transformed using the Box-Cox transform with $\lambda = 0.79$.

Figure 4a shows a scatter of the binary variable representing default (1 = default; 0 = non-default) against loan amount. Clearly the scatter itself is not very informative in this situation, but a non-parametric regression (smooth) has also been included. The method used to obtain the smooth is 'lowess', originally due to Cleveland,[9] and available in recent versions of SPSS. The smooth essentially shows how the probability of default depends on loan amount, and we see that this relationship is negative over a considerable range. We compare this with Figure 4b, which shows a scatter and smooth of arrears against loan amount for defaulters only. Here, we see a clear positive effect: expected arrears rise with loan amount. It is the apparent contradiction between Figures 4a and b, which motivates the need for the double hurdle model since, having two separate equations, one for default and the other for arrears, this model allows the two effects to differ, and even to have opposite signs, as they appear to do in this case.

Figures 5a and b do the same as Figures 4a and b, but with age of borrower measured on the horizontal axis. In Figure 5a, we see a very strong U-shaped effect of age on default probability, calling for the use of both age and age-squared as explanatory variables in the first hurdle equation. However, in Figure 5b, we see that age appears to have no effect on arrears, so on this evidence, there is no reason to include age in the second hurdle equation (this is confirmed during the model selection procedure).

Finally, we consider the effect of gender. Table 1 shows that the proportion of females defaulting is higher than that of males, but that male defaulters are on average in arrears to a greater extent than females. Again, we see an apparent

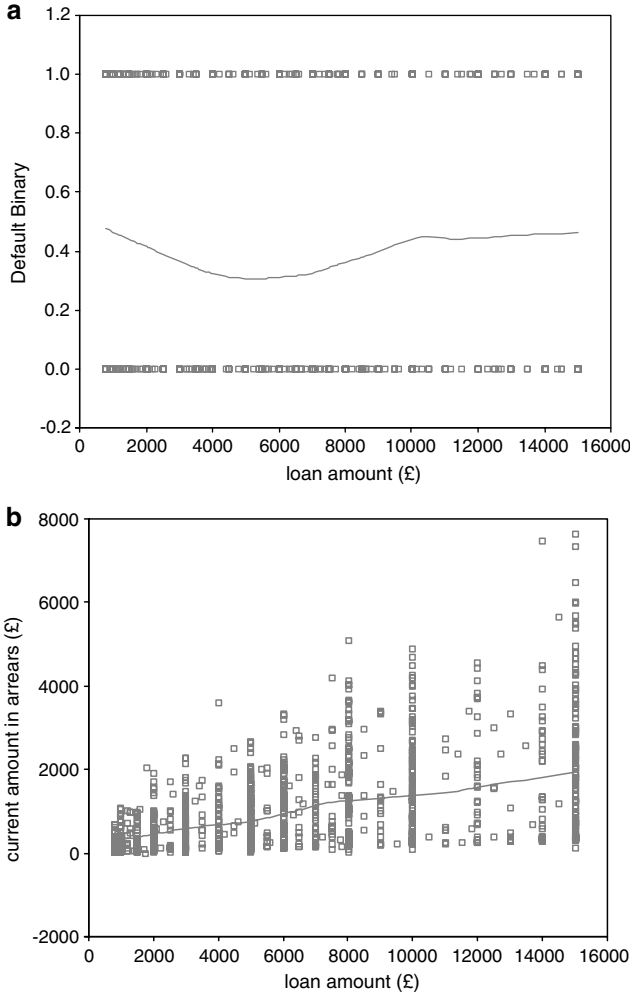

Figure 4 (a) Scatter of binary variable representing default, against loan amount, with smooth; complete sample. (b) Scatter of arrears against loan amount, with smooth; defaulters only.

contradiction, which is dealt with by including gender in both of the equations of the double-hurdle model.

Results

The results from three models are reported in Table 2 for amount in arrears. The sample size used in the estimation of each model is 14417. Recall from the previous section that this is a sample that has been artificially inflated in order to reflect the true population ratio of defaulters to non-defaulters. These three models are part of a lengthy model selection procedure, which started with straightforward tobit models and finished by trying out many different combinations of explanatory variables in the Box-Cox double-hurdle model. Results from the final model are reported in the final column of the table.

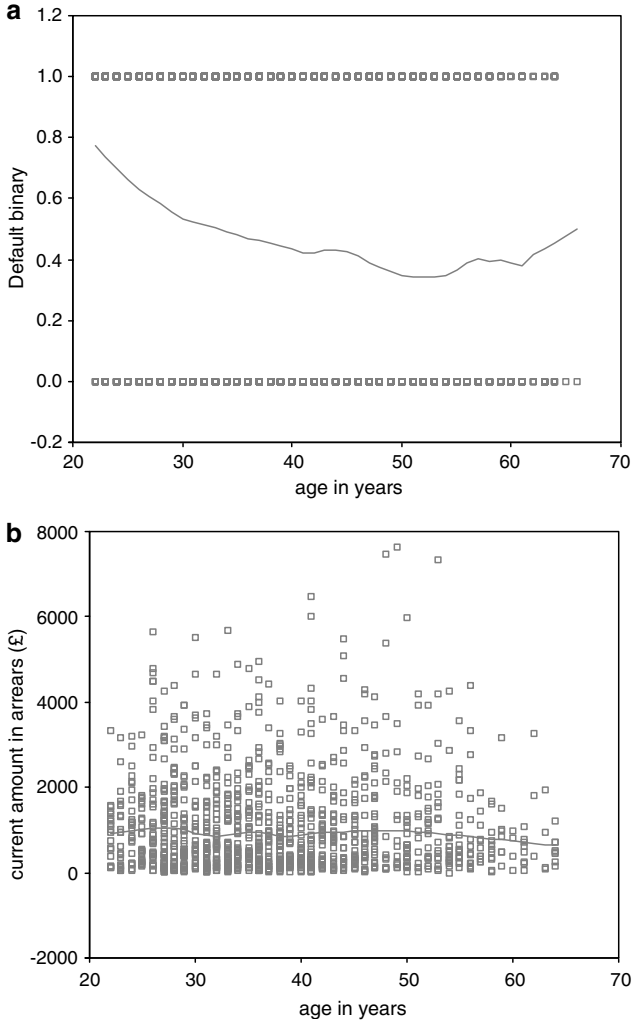

Figure 5 (a) Scatter of binary variable representing default, against age of borrower, with smooth; complete sample. (b) Scatter of arrears against age of borrower, with smooth; defaulters only.

Table 1 Proportion defaulting and mean arrears for defaulters, by gender

	Female	*Male*
Proportion defaulting	0.41	0.37
Mean arrears for defaulters	1064.0	1134.0

Statistically, the Box-Cox double-hurdle model appears vastly superior to the more restrictive models. We see this, for example, when testing the Box-Cox tobit model as a restricted version: the LR statistic is $2(5023.07-4728.60)=588.94$, which, when compared to the $\chi^2(16)$ distribution is seen to represent overwhelming evidence of the importance of the first hurdle, and hence the superiority of the double-hurdle model. For good measure, Akaike's information criterion (AIC) is included at the foot of each column. This is a model selection criterion which adjusts for the number of parameters. The model with the lowest AIC is preferred. This confirms the clear superiority of the Box-Cox double-hurdle specification.

Focusing on the effects of explanatory variables, we see that male borrowers are less likely to pass the first hurdle (ie less likely to be 'potential' defaulters) than females, but, conditional on default, males tend to have a higher level of arrears. This confirms the pattern observed in Table 1. Regarding age of borrower, we see that age indeed has a U-shaped effect on the probability of being a potential defaulter, with a minimum at age $0.06/(2\times 0.0006)=50.0$. This implies that borrowers aged 50 years are the most likely to be in the 'never-default' category.[b] Age is excluded from the second hurdle since it has no significant effect on arrears. Marriage appears to lower the probability of potential default, as does time in occupation and time at bank. Occupation appears to be important in the first hurdle, with office workers perhaps being the 'safest', while purpose of loan appears important in the second hurdle, with household loans being associated with the lowest levels of default. Credit history variables appear to have the expected signs in both equations.

Perhaps the most interesting effect is that of loan amount. This variable has a clear U-shaped effect on the probability of passing the first hurdle, with a minimum at a loan amount of £11 500, but a significantly positive effect on arrears. This was of course expected after the non-parametric analysis of this effect reported in the previous section. The apparent contradiction confirms the value of the hurdle approach. Furthermore, note that the coefficient of loan amount in the simpler Box-Cox tobit model (first column of Table 2) is 0.05. This is a serious under-estimate, being some 80% lower than the corresponding estimate of 0.24 in the Box-Cox double-hurdle model. This bias arises as a result of the invalid treatment of both hurdles as a single process.

The focus of interest in this paper is the borrowers in the 'never-default' category. It is interesting to deduce from the estimates in Table 2 the proportion of borrowers who are in this category. The Box-Cox *p*-tobit model estimates that the proportion passing the first hurdle is 0.71, implying that the proportion of 'never-defaulters' in the population is 0.29 or 29%. However, we prefer to address the same question with the results from the superior Box-Cox double-hurdle model. In this model, the probability of never-default obviously depends on borrower characteristics, according to:

$$\hat{P}(never\ default) = 1 - \Phi(z_i'\hat{\alpha}) \qquad (16)$$

where $\hat{\alpha}$ is the vector containing the first hurdle estimates. Equation (16) has been computed for each of the non-defaulters in the original sample, and the distribution of this predicted probability is shown in Figure 6. It is striking that this model is predicting such high probabilities

Table 2 MLEs for three models. (Dependent variable: amount in arrears (in thousands))

	Box-Cox tobit	*Box-Cox p-tobit*	*Box-Cox double hurdle*
First hurdle			
Constant			1.64 (0.41)
Male			−0.16 (0.06)**
Age			−0.06 (0.02)**
Age-squared			0.0006 (0.0002)**
Married			−0.23 (0.05)**
Time in occupation			−0.001 (0.0002)**
Time at bank			−0.002 (0.0002)**
Office worker			−0.12 (0.06)*
Social worker			0.29 (0.14)*
Supervisor			0.28 (0.10)**
Semi/unskilled			0.19 (0.07)**
# Credit searches			0.16 (0.01)**
# Settled CAIS a/cs			−0.04 (0.01)**
Term of loan			0.01 (0.002)**
Loan amount ('000)			−0.31 (0.03)**
Loan amount ('000) squared			0.013 (0.002)**
p (in p-tobit)		0.71 (0.06)	
Second hurdle			
Constant	−3.23 (0.19)	−2.72 (0.24)	−2.12 (0.19)
Male	−0.22 (0.08)**	−0.17 (0.08)*	0.19 (0.09)*
Homeowner	−1.50 (0.15)**	−1.48 (0.16)**	−0.54 (0.13)**
Tenant	−0.05 (0.18)	0.00 (0.18)	0.32 (0.14)*
Gross income ('000/month)	−0.30 (0.06)**	−0.34 (0.06)**	−0.32 (0.06)**
# Credit searches	0.42 (0.02)**	0.48 (0.02)**	0.20 (0.02)**
Loan amount ('000)	0.05 (0.01)**	0.06 (0.01)**	0.24 (0.01)**
Purpose: vehicle	−0.73 (0.11)**	−0.65 (0.11)**	−0.51 (0.10)**
Purpose: household	−0.90 (0.31)**	−0.83 (0.31)**	−0.72 (0.21)**
Purpose: one-off	0.34 (0.19)*	0.43 (0.19)*	0.29 (0.15)*
Purpose: consolidation	−0.09 (0.10)	−0.08 (0.10)	−0.09 (0.09)
σ	2.37 (0.06)	2.17 (0.07)	1.29 (0.05)
λ	0.87 (0.02)	0.86 (0.02)	0.79 (0.02)
Sample size (n)	14417	14417	14417
K	13	14	29
$\log L$	−5023.07	−5011.17	−4728.60
$AIC = (-\log L + k)/n$	0.349	0.348	0.330

Standard errors in parentheses.
*$P<0.05$, **$P<0.01$.

of never-default, with the vast majority in excess of 50%, and a mean of 83.4%. The clear message here is that the majority of actual non-defaulters are in fact 'never-defaulters'.

Table 3 shows the results for the most general model, the Box-Cox double-hurdle model, but with the number of days in arrears as the dependent variable. Recall that this variable is available only in interval form, so that the log-likelihood function (15) is required to estimate the model. The same set of explanatory variables is used as in Table 2.

Unsurprisingly, the results broadly similar to those of Table 2, with one notable exception. While loan amount once again appears to have a clear U-shaped effect on the probability of potential default (with a minimum at a loan amount of £10 000), the effect of loan amount on number of days in default, conditional on default, appears to be negative.

Conclusion

Casual observation would suggest there exist a subset of non-defaulters who would, on principle, never default under any circumstances. Given this, it is important to recognize the existence of this group in model construction, since their behaviour is clearly determined by a different process to that of the remainder of the population. The double-hurdle class of model has been applied in this paper with this important distinction in mind. Not only does the model allow a class of 'never-defaulters' to exist, but it allows the probability of being in this class to depend on borrower characteristics. The

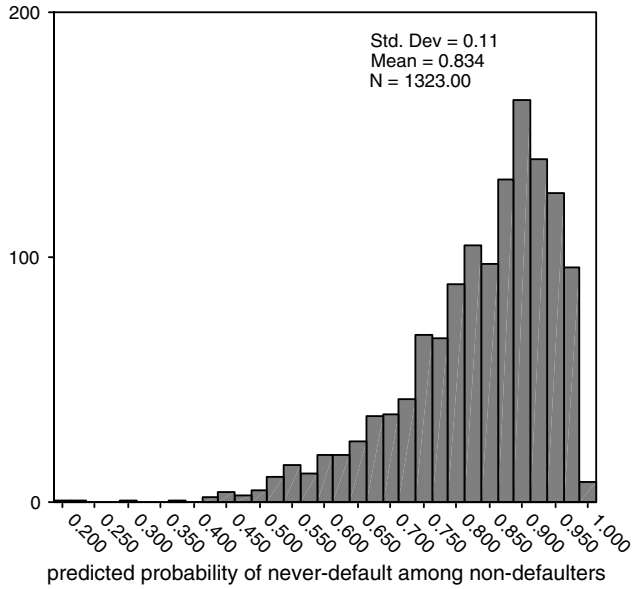

Figure 6 Histogram of predicted probability of 'never-default' for the sub-sample of non-defaulters, obtained from the estimates of the Box-Cox double-hurdle model.

models have been applied separately to two measures of the extent of loan default: amount in arrears; and number of days in arrears. Results are broadly similar for the two measures.

One aspect in which the results are interesting is the apparent differences in explanatory variable effects between the two hurdles. There is a broad separation of variables, with personal characteristics such as cohort, gender and occupation important in the first hurdle, and economic characteristics such as income and tenancy status in the second. More specifically, we have seen important differences between the effects of variables between the two hurdles, most notably loan amount: small borrowers are the most likely to be 'potential-defaulters', but large borrowers who do default, default by more than small borrowers.

The other aspect in which the results are interesting is that they enable us to obtain an estimate of the proportion of non-defaulters who are 'never-defaulters'. In the previous section, using the results of our final model, we estimated this proportion to be over 80%. This estimate seems very high and we recommend that some sensitivity analysis and out-of-sample predictions are carried out before practical use is made of these results.

Endnotes

[a]STATA version 8.0, Stata Corporation, College Station, Texas.

[b]This may, of course, be a 'cohort effect', with borrowers born around 1950 being 'safer' than other cohorts. In fact, given the nature of the model, it is logical to assume that it is a cohort effect, since age itself cannot affect the probability of *never* defaulting. To distinguish the cohort effect from the age effect would require additional observations taken in a different year.

Table 3 MLEs for the Box-Cox double-hurdle model for interval data ((15) used for estimation). (Dependent variable: number of days in arrears)

	Box-Cox double hurdle
First hurdle	
Constant	0.66 (0.48)
Male	0.09 (0.12)
Age	−0.04 (0.02)*
Age-squared	0.0003 (0.002)
Married	−0.20 (0.07)**
Time in occupation	−0.002 (0.0003)**
Time at bank	−0.002 (0.0004)**
Office worker	−0.16 (0.08)*
Social worker	0.24 (0.17)
Supervisor	0.20 (0.13)
Semi/unskilled	0.20 (0.09)*
# Credit searches	0.33 (0.05)**
# Settled CAIS a/cs	−0.06 (0.01)**
Term of loan	0.02 (0.002)**
Loan amount ('000)	−0.20 (0.05)**
Loan amount ('000) squared	0.01 (0.004)**
Second hurdle	
Constant	15.63 (10.21)
Male	−11.57 (6.41)*
Homeowner	−36.05 (7.70)**
Tenant	−16.75 (7.83)*
Gross income ('000/month)	−8.71 (2.39)**
# Credit searches	6.03 (1.43)**
Loan amount ('000)	−1.51 (0.90)*
Purpose: vehicle	−21.91 (4.93)**
Purpose: household	−41.55 (12.11)**
Purpose: one-off	10.38 (8.23)
Purpose: consolidation	−3.26 (3.82)
σ	64.57 (8.60)
λ	0.64 (0.03)
Sample size (n)	14417
K	29
Log L	−5679.03
$AIC = (-\text{Log}\, L + k)/n$	0.40

Standard errors in parentheses.
*$P<0.05$, **$P<0.01$.

References

1 Cragg JG (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* **39**: 829–844.
2 Dionne G, Artis M and Guillen M (1996). Count data models for a credit scoring system. *J Empirical Finance* **3**: 303–325.
3 Jones AM (1989). A double hurdle model of cigarette consumption. *J Appl Econom* **4**: 23–39.
4 Jones AM and Yen ST (2000). A Box-Cox double hurdle model. *The Manchester School* **68**: 203–221.
5 Smith MD (2002). On specifying double hurdle models. In: Ullah A, Wan A and Chaturvedi A (eds). *Handbook of Applied*

Econometrics and Statistical Inference. Marcel-Dekker, New York, pp 535–552.
6 McDowell A (2003). From the help desk: hurdle models. *Stata J* **3**: 178–184.
7 Deaton AS and Irish M (1984). Statistical models for zero expenditures in household budgets. *J Public Econom* **23**: 59–80.
8 Stewart MB (1983). On least squares estimation when the dependent variable is grouped. *Rev Econ Studies* **50**: 737–753.
9 Cleveland WS (1979). Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* **74**: 829–836.

Appendix A. STATA code for estimation of Box-Cox double-hurdle model

```
program define dh
  version 6
  args lnf theta1 theta2 theta3 theta4
  tempvar d p z p0 p1l yt
  quietly gen double 'd' = $ML_y1 > 0
  quietly gen double 'p' = normprob('theta3')
  quietly gen double 'l' = 'theta4'
  quietly gen double 'yt' = ($ML_y1^'l'-1)/'l'
  quietly gen double 'z' = ('yt'-'theta1')/('theta2')
  quietly gen double 'p0' = 1-('p'*normprob(-'z'))
  quietly gen double 'p1' = (($ML_y1 + (1-'d'))^
    ('l'-1))*'p'*normd('z')/'theta2'
  quietly replace 'lnf' = ln((1-'d')*'p0' + 'd'*'p1')
  end
ml model lf dh (y = 'listy') () (d = 'listd') ()
ml init b, copy
ml maximize
```

Notes: 'listy' is a previously defined list of variables appearing in the second hurdle; 'listd' contains the variables of the first hurdle. 'theta1' corresponds to $x_i'\beta$ in (14), 'theta2' to σ, 'theta3' to $z_i'\alpha$, and 'theta4' to λ. b is a vector of suitable starting values.

Appendix B. STATA code for estimation of Box-Cox double-hurdle model with interval data

```
program define intdh
  version 6
  args lnf theta1 theta2 theta3 theta4
  tempvar y d ppp p0 p1 p2 p3 p4 p5 p6 p7 p8 l
  quietly gen double 'y' = $ML_y1
  quietly gen double 'd' = 'y' > 0
  quietly gen double 'ppp' = normprob('theta3')
  quietly gen double 'l' = ('theta4')
  quietly gen double 'p0' = normprob
    ((-1/'l'-'theta1')/'theta2')
  quietly gen double 'p1' = normprob(((30^'l'-1)/
    'l'-'theta1')/'theta2')-
    normprob(((-1/'l')-'theta1')/'theta2')
  quietly gen double 'p2' =
    normprob(((60^'l'-1)/'l'-'theta1')/'theta2')-
    normprob(((30^'l'-1)/'l'-'theta1')/'theta2')
  quietly gen double 'p3' =
    normprob(((90^'l'-1)/'l'-'theta1')/'theta2')-
    normprob(((60^'l'-1)/'l'-'theta1')/'theta2')
  quietly gen double 'p4' =
    normprob(((120^'l'-1)/'l'-'theta1')/'theta2')-
    normprob(((90^'l'-1)/'l'-'theta1')/'theta2')
  quietly gen double 'p5' =
    normprob(((150^'l'-1)/'l'-'theta1')/'theta2')-
    normprob(((120^'l'-1)/'l'-'theta1')/'theta2')
  quietly gen double 'p6' =
    normprob(((180^'l'-1)/'l'-'theta1')/'theta2')-
    normprob(((150^'l'-1)/'l'-'theta1')/'theta2')
  quietly gen double 'p7' =
    normprob(((210^'l'-1)/'l'-'theta1')/'theta2')-
    normprob(((180^'l'-1)/'l'-'theta1')/'theta2')
  quietly gen double 'p8' = 1-'p0'-'p1'-'p2'-'p3'-
    'p4'-'p5'-'p6'-'p7'
  quietly replace 'lnf' =
    ln(('y' = = 0)*(1-('ppp'*(1-'p0'))) +
    'ppp'*(('y' = = 1)*'p1' + ('y' = = 2)*'p2'
    + ('y' = = 3)*'p3' + ('y' = = 4)*'p4'
    + ('y' = = 5)*'p5' + ('y' = = 6)*'p6'
    + ('y' = = 7)*'p7' + ('y' = = 8)*'p8'))
  end
ml model lf intdh (y = 'listy') () (d = 'listd') ()
ml init b, copy
ml maximize
```

Notes: The Notes to Appendix A also apply here. In addition, the $J = 8$ intervals assumed in the construction of this likelihood function are: 0–30; 30–60; 60–90; 90–120; 120–150; 150–180; 180–210; 210+. The probability associated with the final interval is computed as one minus the sum of the other seven.

Received January 2004;
accepted October 2004