ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Jan Żółtowski**

APPLICATION OF PROBIT MODELS AND SELECTED DISCRIMINATION ANALYSIS METHODS FOR CREDIT DECISION EVALUATION

Abstract

Retail banking deals with servicing consumer credits and it constitutes one of the major banking activities. A customer applying for the credit fills in the application which is basis to evaluated of his creditworthiness.

The paper considers the problem of evaluation to which of the two groups the person applying for a credit should be assigned to: a) those who possess the creditworthiness; b) those who do not possess the creditworthiness. It analyses the possibility of applying the probit models and the discrimination analysis methods using the quadratic and linear discrimination function. An evaluation of the correctness of the classification based on the real data from a commercial bank is conducted.

Key words: Bayes discrimination methods, quadratic discrimination function, classification function, probit model.

I. INTRODUCTION

Among various types of activities performed by banks, retail banking is one, which deals with the issue of consumer credits. Each bank acts according to previously established regulations regarding credit granting and repaying. A client applying for a consumer credit fills out a credit application, which constitutes a basis for the client's creditworthiness evaluation. Data from the credit application are processed into scoring, which allows to assign the applicant to one of the two groups: a) able to repay a credit, b) unable to repay a credit.

---

* Ph.D., student, Chair of Statistical Methods, University of Łódź.

Therefore, a problem arises whether, and if so, how we can predict which of the two groups the credit applicant will be assigned to, based on the statistical data pertaining to credit granting and on the information about the client. Also, how can we establish which values of the socio-economic client characteristics assure an appropriate scoring level?

A credit decision made by a bank can be described by a binary variable:

$$Y = \begin{cases} 1, & \text{when a credit was granted} \\ 0, & \text{when a credit was not granted} \end{cases}. \tag{1}$$

Regression models are commonly used in the causality relationship analysis. One of them is the following linear regression model:

$$Y_t = \mathbf{x}_t^T \boldsymbol{\alpha} + \varepsilon_t, \quad \text{for} \quad t = 1, ..., T, \tag{2}$$

where: $\mathbf{x}_t$ is a vector of exogenous variables, $\boldsymbol{\alpha}$ – a vector of parameters, ε_t – an error term with the expected value of 0.

Let us consider a case, in which the endogenous variable Y is binary with probability distribution function given by:

$$\mathrm{P}(Y_t = 1) = \pi_t, \ \mathrm{P}(Y_t = 0) = 1 - \pi_t \quad \text{and} \quad \pi_t \in (0, 1). \tag{3}$$

Hence $\mathrm{E}(Y_t) = \pi_t$. Moreover, based on the assumptions and the model specification $\mathrm{E}(Y_t) = \mathbf{x}_t^T \boldsymbol{\alpha}$. The existence of a binary endogenous variable in the regression model causes a particular interpretation of the theoretical values $\hat{Y}_t = \mathbf{x}_t^T \hat{\boldsymbol{\alpha}}$ obtained from model (2). Specifically, they are not unbiased estimators of probabilities $\mathrm{P}(Y_t = 1) = \pi_t$, if $\mathrm{E}(\hat{\boldsymbol{\alpha}}) = \boldsymbol{\alpha}$. As a result, it is necessary to select a method, which while estimating the parameters of model (2) satisfies the following condition: $\hat{Y}_t = \mathbf{x}_t^T \hat{\boldsymbol{\alpha}} \in (0, 1)$. A probit model is one of such methods. After having estimated its parameters, one can estimate the probability $\mathrm{P}(Y_t = 1)$ also for other values of the exogenous variables.

The problem of a bank decision prediction analysed above can also be considered as classification issue. A population Π of credit applicants can be divided into two sub-populations Π_0 i Π_1. Assigning an applicant to the sub-population Π_0 is equivalent to denying a credit, while assigning him or her to the sub-population Π_1 corresponds to granting a credit. A bank decision is made after the analysis of the client's ability to repay the credit. The assignment to one of the two described above groups is based on values of m statistical characteristics describing client's socio-economic situation. A vector $\mathbf{x} \in R^m$ will represent them. The space of values of the characteristics can be divided (*based on their values for the elements of the learning set*) into two disjoint regions $\mathcal{X}_0$ and $\mathcal{X}_1 = R^m \backslash \mathcal{X}_0$. A situation in which vector x belongs to the region $\mathcal{X}_0$ is equivalent to assigning a credit applicant to the sub-population Π_0.

This study examines an application of both approaches in the prediction of credit granting decisions based on the example of a branch of a certain bank.

II. PROBIT MODELS[1]

Let's assume that we have a large sample (obtained from an independent sampling) and that we divide the set of observations into M subsets. For each of the subsets we can derive the frequency of the variable Y taking a value of one. Let each k-th subset $(i = 1, 2, ..., M)$ with n_k elements have m_k number of ones. Then the empirical probability can be computed as frequency $\frac{m_k}{n_k}$. We assume that with the accuracy of the error ε_k, it is equal to the theoretical probability π_k, which can be interpreted as the value of the cumulative distribution function of a certain distribution, i.e.:

$$\pi_k = F(\mathbf{x}_k^T \alpha).$$

Therefore:

$$\frac{m_k}{n_k} = \pi_k + \varepsilon_k, \tag{4}$$

where:

$$E(\varepsilon_k) = 0, \ D^2(\varepsilon_k) = \frac{\pi_k(1 - \pi_k)}{n_k}.$$

Hence,

$$F^{-1}\left(\frac{m_k}{n_k}\right) = F^{-1}(\pi_k + \varepsilon_k). \tag{5}$$

[1] Models with discrete exogenous variable are discussed by Jajuga in chapter 8 of works by S. Bartosiewicz (1990).

After having expanded the function F^{-1} into Taylor series about the point π_k we obtain the following model:

$$F^{-1}\left(\frac{m_k}{n_k}\right) = \mathbf{x}_k^T \alpha + \eta_k, \tag{6}$$

where:

$$E(\eta_k) = E\left(\frac{\varepsilon_i}{f(\mathbf{x}_k^T \alpha)}\right) = 0, \ D^2(\eta_k) = \frac{\pi_i(1-\pi_i)}{f^2(\mathbf{x}_i^T \alpha) n_i}.$$

Model (6) is called a probit model[2] and it is a model, in which the error term is heteroscedastic. Such a model can be estimated with the generalised least squares method or with the maximum likelihood method.

III. SELECTED BAYES DISCRIMINATION METHODS

A selection of the discrimination method based on the theory of statistical decision functions and a procedure in the case, in which there exist two sets of elements Π_0 and Π_1, depend on the information regarding the prior probabilities p_0 and p_1 of a certain element belonging to a particular set and of the distribution of the variables $\mathbf{X} = [X_1, X_2, ..., X_m]^T$ characterising the elements of the population[3]. Applying Bayes classification rule, we can choose one of the alternative decisions regarding whether the element belongs to a certain sub-population.

Let,

$$f_i(\mathbf{x}) = (2\pi)^{-\frac{m}{2}}(\det \ \mathbf{\Sigma}_i)^{-0.5}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right] \tag{7}$$

be the probability density function of the random variable $\mathbf{X}$, when the analysed element $0 \in \Pi_i$ dla $i = 0, 1$.

$S_i^*(\mathbf{x}) = p_i f_i(\mathbf{x})$, $i = 0, 1$ can be used as the classification function provided that the loss is constant when an element is misclassified. More than one

[2] Interesting examples of the application of probit analysis can be found for example in publication by: Wiśniewski (1986), Pruska (2001).

[3] Methods suggested in such cases were gathered by K. Jajuga (1990), p. 40–41 in Table 1.

particular classification function can be chosen, since the classification will not be altered, when function $S_i(\mathbf{x})$, is replaced with:

$$S_i(\mathbf{x}) = g(S_i^*(\mathbf{x})), \tag{8}$$

where g is any increasing function.
$S_i^*(\mathbf{x}) = \ln(p_i f_i(\mathbf{x}))$ may be applied as the classification function, i.e.

$$S_i^*(\mathbf{x}) = \frac{-m}{2}\ln(2\pi) - \frac{1}{2}(\det \boldsymbol{\Sigma}_i) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p_i. \tag{9}$$

Since the first element in the formula (9) is constant with respect to i we can ignore it and the equivalent classification function is as follows:

$$S_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}(\det \boldsymbol{\Sigma}_i) + \ln p_i, \quad \text{for } i = 0, 1. \tag{10}$$

The function (10) contains a quadratic form of a vector $(\mathbf{x}-\boldsymbol{\mu}_i)$, and as a result it is called a quadratic classification function. Its value for a given $\mathbf{x}$ depends upon the prior probability p_i and upon the parameters of the distribution of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$.

Applying Bayes classification rule with respect to a prior distribution (p_0, p_1), we include an observation $\mathbf{x}$ in the population Π_i, for which the classification function $S_i(\mathbf{x})$ takes the biggest value for $i = 0, 1$. Classification regions are determined using the Bayes rule and take the following form:

$$\mathcal{X}_0 = \{\mathbf{x} : S_0(\mathbf{x}) \geqslant S_1(\mathbf{x})\}. \tag{11}$$
$$\mathcal{X}_1 = R^m \backslash \mathcal{X}_0$$

The inequality in formula (11) can be substituted with the following equivalent inequality:

$$(S_0(\mathbf{x}) - \ln \mathrm{p}_0) - S_1(\mathbf{x}) - \ln \mathrm{p}_1) \geqslant \ln \frac{\mathrm{p}_1}{\mathrm{p}_0}. \tag{12}$$

Denoting the left-hand side of the inequality (12) by $S_{01}(\mathbf{x})$ and taking into consideration formula (10) we receive the following function:

$$S_{01}(\mathbf{x}) = \frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) + \ln\frac{\det \boldsymbol{\Sigma}_1}{\det \boldsymbol{\Sigma}_0}], \tag{13}$$

which is independent of the prior probability p_i and called a quadratic discrimination function.

Quasi-Bayesian estimator is a consistent estimator of the quadratic discrimination function (13). It is obtained based on the normal distribution probability density function estimator of the following form (see: Krzyśko, 1990: 53):

$$\hat{S}_{01}(\mathbf{x}) = \frac{N_1}{2}\ln[1 + \frac{N_1}{N_1^2 - 1} D_1^2(\mathbf{x})] - \frac{N_0}{2}\ln[1 + \frac{N_0}{N_0^2 - 1} D_0^2(\mathbf{x})] + \ln\frac{c_0}{c_1}, \tag{14}$$

where:

$$c_1 = \frac{N_i}{\pi(N_i^2 - 1)} \frac{\Gamma(\frac{N_i}{2})}{\Gamma(\frac{N_i - 2}{2})\sqrt{\det\hat{\Sigma}_i}} \quad \text{and} \quad \operatorname*{plim}_{N_i, N_j \to \infty} \frac{c_0}{c_1} = \ln\frac{|\Sigma_1|}{|\Sigma_0|},$$

$$D_i^2(\mathbf{x}) = (\mathbf{x} - \overline{\mathbf{x}}_i)^{\mathrm{T}} \hat{\Sigma}_i^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i), \quad \text{for} \quad i = 0, 1. \tag{15}$$

Statistics from the sample are usually used as estimators of the parameters in the formula:

$$\hat{\mu}_i = \overline{\mathbf{x}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{x}_{ij}, \quad \hat{\Sigma}_i = \frac{1}{N_i - 1}\mathbf{W}_i, \quad \mathbf{W}_i = \sum_{j=1}^{N_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^{\mathrm{T}}. \tag{16}$$

Employing the estimator $\hat{S}_{01}(\mathbf{x})$ of a quadratic discrimination function given by the formula (15), we assign an observation $\mathbf{x}$ to the sub-population Π_0 according to the Bayes classification rule when $\hat{S}_{01}(\mathbf{x}) \geqslant \ln\frac{\hat{p}_1}{\hat{p}_0}$ where $\hat{p}_0$ and $\hat{p}_1$ are prior probabilities estimators.

Also, in the discrimination analysis one considers the problem of a reduction of the number of variables characterising elements subject to classification. The set of the original variables $X_1, X_2, ..., X_m$ is divided into disjoint subsets and a new variable, called a discrimination variable, is assigned to each of the subsets. The discrimination variable constitutes a linear combination of the variables contained in a particular subset. Searching for the discrimination variables, one should aim at $U_1, U_2, ..., U_r$ which are not mutually correlated, which have unit variances and maximise the selected distribution measure[4].

Let us assume, just like we did previously, that Π_0, Π_1 are sub-populations of the general population Π and that $\mathbf{x} = [x_1, x_2, ..., x_m]^{\mathrm{T}}$, whose distribution

[4] This issue is discussed for example by Krzyśko (1990), Chapter 3.

is multivariate normal, is the realisation of a random vector $\mathbf{X} = [X_1, X_2, ..., X_m]^{\mathrm{T}}$ in the sample.
Let $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_r$ be the largest roots of the equation:

$$\det(\mathbf{B} - \hat{\lambda}\,\mathbf{W}) = 0, \tag{17}$$

and $\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_r$ vectors of length 1 satisfying the following matrix equation:

$$(\mathbf{B} - \hat{\lambda}_j \mathbf{W})\hat{\mathbf{l}} = \mathbf{0}, \tag{18}$$

respectively for $j = 1, 2, ..., r$
where

$$\mathbf{W} = (\mathbf{W}_0 + \mathbf{W}_1),$$

$$\mathbf{B} = N_0(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}})^{\mathrm{T}} + N_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})^{\mathrm{T}},$$

$$\bar{\mathbf{x}} = \frac{N_0\bar{\mathbf{x}}_0 + N_1\bar{\mathbf{x}}_1}{N_0 + N_1}$$

The discriminatory variable $\hat{U}_j$ can be estimated from the sample as:

$$\hat{U}_j = \hat{\mathbf{l}}_j^{\mathrm{T}}\,\mathbf{x}. \tag{19}$$

Denoting by $\hat{\mathbf{U}} = [\hat{U}_1, \hat{U}_2, ..., \hat{U}_r]^{\mathrm{T}}$ and $\mathbf{v}_i = [\hat{\mathbf{l}}_1 \vdots \hat{\mathbf{l}}_2 \vdots ... \vdots \hat{\mathbf{l}}_r]^{\mathrm{T}}\bar{\mathbf{x}}_i$ for $i = 0, 1$ we obtain the following form of the classification function:

$$S_i(\hat{\mathbf{U}}) = -\frac{1}{2}(\hat{\mathbf{U}} - \mathbf{v}_i)^{\mathrm{T}}(\hat{\mathbf{U}} - \mathbf{v}_i) + \ln \hat{\mathrm{p}}_i. \tag{20}$$

Observation $\mathbf{x}$ is assigned to the sub-population Π_0, when $S_0(\hat{\mathbf{U}}) \geqslant S_1(\hat{\mathbf{U}})$.

IV. EMPIRICAL EXAMPLE

In his or her credit application a client provides basic data (such as personal information, address, net income, additional sources of income, housing and other stable monthly expenses, potential obligations to serve in the army) and supplemental data (regarding his or her housing situation, marital status, number of members of the household, type of employer and years worked for that employer, finally regarding the number of credits taken or guaranteed).

The data contained in the application are transformed into scoring, which constitutes a basis for assigning the applicant to one of the two groups:

1) with the ability to repay a credit;
2) without the ability to repay.

The second group is sometimes divided into two sections: applicants who will be denied a credit and those who will be further considered in the credit decision after having supplied an additional collateral.

Data concerning received credit applications and bank decisions about granting or denying a credit over a period of six consecutive months in 2001 were gathered in one of the branches of a commercial bank. It was established at that time that, as a general rule, a credit was denied if a client has not fulfilled his army obligations. Therefore, all the applications in which this was the case were removed and as a result a set of 239 observations was obtained.

Those applications were divided into two groups. The first group was created from the applications received during the first 5 months and was treated as a *learning* set. This group consisted of 203 applications (including 131 cases followed by a negative decision – credit denial, and 72 cases followed by a positive decision). Applications received in June (36 applications, including 24 cases followed by a negative decision) made up the second group (which was treated as the examined set), which was used to predict a credit decision. This enabled us to evaluate the accuracy (*fitness*) of the applied methods.

Based on the applications the following variables characterising a credit applicant were singled out:

1) **quantitative variables:**

X_1 – primary monthly net income [in PLZ],
X_2 – supplemental monthly net income [in PLZ],
X_3 – stable monthly housing expenses [in PLZ],
X_4 – other stable monthly expenses [in PLZ],
X_5 – number of household members,
X_6 – period worked with the current employer [in years],
X_7 – number of taken or guaranteed credits,
X_8 – monthly income of the co-applicant if there is one.

2) **qualitative variables:**

X_9 – variable specifying whether the applicant rents/owns an apartment (a house) ($X_9 = 1$, when a client rents or owns an apartment (a house) and $X_9 = 0$ otherwise),

X_{10} variable specifying the applicant's marital status ($X_{10} = 1$, if the applicant is married and $X_{10} = 0$ if the applicant is single),

X_{11} – variable specifying the applicant's employment status ($X_{11} = 1$, if the applicant works for a governmental company, public administration, owns a proprietorship or is a partner in a partnership and $X_{11} = 0$ otherwise).

A credit application decision made by a bank can be described by a binary variable:

$$Y = \begin{cases} 1, & \text{when a credit was granted} \\ 0, & \text{when a credit was nokt granted} \end{cases}.$$

The amount of credit requested in the application [in PLZ] is an additional variable: Y_2.

Two new variables were derived:

X_{12} – net discretionary income (a sum of primary and supplemental net income after deducting stable monthly expenses, $X_{12} = (X_1 + X_2) - (X_2 + X_4)$),

and

X_{13} – disposable gross income (the sum of the net income of the applicant and the co-applicant $X_{13} = X_{12} + X_8$).

Let us consider the problem of predicting, which of the two groups a client will be assigned to based on the decisions made in the *learning* set and on the data regarding the new client. We will utilise probit models and Bayes discrimination analysis method to examine this problem.

In order to compare results of the client classification obtained with different methods described above, we had to select variables, which can be employed by all methods. In particular, normal distribution of all utilised variables was assumed in Bayes discrimination. We verified that X_{12} and $\frac{X_{13}}{Y_2}$ can be treated as variables with a logarithmic-normal distribution. Therefore, basic variables used in all examined models were: $w_1 = \ln X_{12}$ and $w_2 = \ln X_{13} - \ln Y_2$.

Three types of probit model were analysed:

$$\Phi^{-1}\left(\frac{m_k}{n_k}\right) = \alpha_0 + \alpha_1 w_1 + \alpha_2 w_2 + \eta_1, \tag{21}$$

$$\Phi^{-1}\left(\frac{m_k}{n_k}\right) = \beta_0 + \beta_1 w_1 + \beta_2 w_2 + \beta_3 X_9 + \eta_2, \tag{22}$$

$$\Phi^{-1}\left(\frac{m_k}{n_k}\right) = \gamma_0 + \gamma_1 w_1 + \gamma_2 w_2 + \gamma_3 X_9 + \gamma_4 X_{10} + \eta_3, \tag{23}$$

where:

$w_1 = \ln X_{12}$ – logarithm of net discretionary income,

$w_2 = \ln \frac{X_{13}}{Y_2}$ – logarithm of gross disposable income and credit amount,

X_9 – binary variable accepting the value of 1 if an applicant rents or owns an apartment (house),

X_{10} – binary variable accepting the value of 1 if an applicant is married.

After an application of a probit analysis and estimation[5] of appropriate probit models parameters (based on the data from the *learning* set) we have obtained the following results:

Table 1. Accuracy of credit applications classification based on probit models

Model	Observed value Y_1	Results for the *learning* set			Results for the examined set		
		Predicted value Y_1		% accurate class.	Predicted value Y_1		% accurate class.
		0	1		0	1	
(21)	0	117	14	89.3	19	5	79.2
	1	20	52	72.2	3	9	75.0
(22)	0	121	10	92.4	22	2	91.7
	1	11	61	84.7	2	10	83.3
(23)	0	124	7	94.7	20	4	83.3
	1	7	65	90.3	1	11	91.7

Source: Author's computations.

While analysing the results, we note that for the *learning* set, the percentage of accurate classification based on model (21) is relatively high (89% and 72%). However, classification based on probit models (22) and (23) is more accurate (the number of correctly predicted credit decisions increases). Models "perform better" in terms of identifying the cases of credit denial in the *learning* sample. The percentage of an accurate prediction of a credit denial is 17 points higher than the percentage of an accurate prediction of a credit granting decision (for model (21)). For the examined sample (unfortunately, not very numerous) the general situation regarding the accuracy of prediction is similar.

Using the estimator $\hat{S}_{01}(\mathbf{x})$ of a quadratic discrimination function given by the formula (14), observation x is assigned to sub-population $\Pi_0(Y_t = 0)$, according to the Bayes classification rule if $\hat{S}_{01}(\mathbf{x}) \geqslant \ln \frac{\hat{q}_1}{\hat{q}_0}$.

[5] Necessary calculations were performed in STATISTICA-5.0.

The mean values, variances and a covariance of variables w_1, w_2 for both sub-populations were derived from the *learning* set (including 203 observations from the first 5 months of 2001) and $\mathbf{x} = \mathbf{w}^{(j)} = [w_1^{(j)} w_2^{(j)}]^T$ was substituted in formula (14). As a result, we have arrived at the following form of an estimator of a discrimination function for an i-th observation:

$$\hat{S}_{01}(\mathbf{w}^{(j)}) = 36 \ln [1 + 0.01389\, D_1^2(\mathbf{w}^{(j)})] - 65.5 \ln [1 + 0.00763\, D_0^2(\mathbf{w}^{(j)})] - 1.0376 \tag{24}$$

where:

$$D_0^2(\mathbf{w}^{(j)}) = 11.6419(w_1^{(j)} - 6.7118)^2 + 4.78541(w_2^{(j)} + 1.1622)^2 + - 4.5013(w_1^{(j)} - 6.7118)(w_2^{(j)} + 1.1622)$$

$$D_1^2(\mathbf{w}^{(j)}) = 10.8785(w_1^{(j)} - 7.2581)^2 + 12.1320(w_2^{(j)} + 0.5310)^2 + - 7.2689(w_1^{(j)} - 7.2581)(w_2^{(j)} + 0.5310)$$

For each element j of the *learning* set ($j = 1, 2, ..., 203$) and the examined set ($j = 204, ..., 239$) (credit applicant), the values of a discrimination function $\hat{S}_{01}(\mathbf{w}^{(j)})$ were computed[6] based on the estimated elements of vector $\mathbf{w}^{(j)} = \begin{bmatrix} w_1^{(j)} \\ w_2^{(j)} \end{bmatrix}$. Then credit applications were assigned to sub-population Π_0, namely to the set of applications followed by a credit denial, when

$$\hat{S}_{01}(\mathbf{w}^{(j)}) \geqslant -0.5985$$

and to sub-population Π_1 of applications followed by a credit granting decision otherwise. The following classifications of credit applications have been received:

Table 2. Accuracy of credit applications classification based on the value of the quadratic discrimination function estimator

Observed value Y_1	Results for the *learning* set			Results for the examined set		
	Predicted value Y_1		% accurate class.	Predicted value Y_1		% accurate class.
	0	1		0	1	
0	117	14	89.3	19	5	79.2
1	14	58	80.6	2	10	83.3

Source: Author's computations.

[6] Computations obtained in Excel 5.0.

While analysing the results we note that, the percentage of accurate classifications for the *learning* set is 89% and 80%. The classification obtained with the estimator of a quadratic discrimination function (24) "performed better" in predicting credit denial. The percentage of accurate prediction of credit denial is 9 points higher than that of a credit granting decision. For the examined sample (unfortunately not very numerous) the general situation regarding the accuracy of prediction is similar. However, the percentage of a correct prediction of a credit granting decision increased (by 3 points) and the percentage of a correct prediction of credit denial decreased (by 10 points).

Variables used in the discrimination were (just like above) variables w_1, w_2. Having computed their mean values, variances and their covariance for both sub-populations we received: $\overline{\mathbf{x}}_1$, $\hat{\mathbf{\Sigma}}_0$, $\mathbf{W}_0$, $\overline{\mathbf{x}}_1$, $\hat{\mathbf{\Sigma}}_1$, $\mathbf{W}_1$, and then derived for the entire *learning* set $\overline{\mathbf{x}}$, $\mathbf{W}$, $\mathbf{B}$.

In order to approximate a discrimination variable from the sample $\hat{u}$, $\hat{\lambda} = \max\{\lambda_1, \lambda_2\}$ was introduced, where λ_1, λ_2 symbolise roots of the quadratic equation (17), and was estimated as $\hat{\lambda} = 0.934984$. Vector $\hat{\mathbf{l}}$ satisfying equation (18) turned out to have the following elements $\hat{\mathbf{l}} = [0.8805\ 0.4741]^T$.

Therefore, the following linear combination of the variables w_1, w_2 is a discrimination variable form the sample:

$$\hat{u} = 0.8805\, w_1 + 0.4741\, w_2. \tag{25}$$

Having computed constants v_1 (for the sub-population Π_0) and v_2 (for the sub-population Π_1) we received the explicit forms of estimators of both classification functions[7].

$$S_0(\hat{u}) = -\frac{1}{2}(0.8805\, w_1 + 0.4741\, w_2 - 5.3581)^2 + \ln \hat{p}_0,$$

$$S_1(\hat{u}) = -\frac{1}{2}(0.8805\, w_1 + 0.4741\, w_2 - 6.139)^2 + \ln \hat{p}_1.$$

On their basis the following credit applications classification was obtained[8]:

[7] Let us note that the discrimination function estimator derived from both classification functions would take the following form: $\tilde{S}_{01}(\hat{u}) = -0.6877w_1 - 0.3702w_2 + 4.5963$.

[8] Computations obtained from Excel 5.0.

Table 3. Accuracy of credit applications classification based on the value of the discrimination variable

Scoring $\hat{p}_1$	Observed value Y_1	Results for the *learning* set			Results for the examined set		
		Predicted value Y_1		% accurate class.	Predicted value Y_1		% accurate class.
		0	1		0	1	
0.50	0	110	21	84.0	19	5	79.2
	1	13	59	81.9	2	10	83.3
0.55	0	121	10	92.4	20	4	83.3
	1	28	44	61.1	3	9	75.0
0.60	0	130	1	99.2	23	1	95.8
	1	44	28	38.9	9	3	25.0

Source: Author's computations.

Classification obtained with the discrimination variable $\hat{u}$ leads to similar conclusions as the one obtained with properly constructed estimator of the quadratic discrimination function $\hat{S}_{01}$, if one assumes a prior probability of 0.5. The estimation of this probability obtained from the frequency of a credit denial decision in the *learning* sample amounted to 0.64. If we take values higher than 0.5 for $\hat{p}_0$ we observe an increase in the percentage of correct classification of the credit denial decision for both samples (over 90%). However, this increase is accompanied by a rapid decrease in the correctly classified credit granting decisions.

V. FINAL CONCLUSIONS

The results obtained from the probit model (utilising the same variables as Bayes methods) are similar to the ones received from Bayes discrimination, although the percentage of correct classification is slightly lower in the probit model. The results provided by the extended models (22) and (23) are better as the percentage of correctly classified, both accepted and denied, credit applications increases. In conclusion, additional exogenous variables are relevant for the process of accurate classification. It would be interesting to utilise the same variables in Bayes analysis. However doing so is not trivial since we assumed that the variables used in this model are continuously distributed, while additional variables are binary.

REFERENCES

Bartosiewicz S. et al. (1990), *Estymacja modeli ekonometrycznych*, PWE, Warszawa.
Jajuga K. (1990), *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa.
Krzyśko M. (1990), *Analiza dyskryminacyjna*, Wyd. Naukowo-Techniczne, Warszawa.
Krzyśko M. (1998), *Statystyka matematyczna*, Wyd. Nauk. UAM, Poznań.
Pruska K. (2001), Modele probitowe i logitowe w programach nauczania studiów ekonomicznych, [in]: *Metody analizy cech jakościowych w procesie podejmowania decyzji* (working papers), Wyd. UŁ, Łódź, 89–98.

Jan Żółtowski

ZASTOSOWANIE MODELI PROBITOWYCH I WYBRANYCH METOD ANALIZY DYSKRYMINACYJNEJ DO PRZEWIDYWANIA DECYZJI KREDYTOWEJ

Streszczenie

Obsługa kredytów konsumpcyjnych jest jednym z rodzajów działalności banków. Zdolność kredytowa klienta jest oceniana na podstawie złożonego przez niego wniosku.

W pracy rozważany jest problem przewidywania, do której z dwóch grup klientów, posiadających zdolność kredytową lub nie (w ocenie banku), zostanie zaliczona osoba ubiegająca się o kredyt. Analizowane są tu możliwości zastosowania modeli probitowych oraz metod analizy dyskryminacyjnej wykorzystujących kwadratową funkcję dyskryminacyjną i zmienną dyskryminacyjną z próby. Przeprowadzona jest także ocena poprawności klasyfikacji danych z pewnego banku.