“Partial Least Squares Regression in Payment Default Prediction”

AUTHORS	Erkki K. Laitinen
ARTICLE INFO	Erkki K. Laitinen (2006). Partial Least Squares Regression in Payment Default Prediction. *Investment Management and Financial Innovations*, *3*(1)
RELEASED ON	Wednesday, 01 March 2006
JOURNAL	"Investment Management and Financial Innovations"
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES **0**

NUMBER OF FIGURES **0**

NUMBER OF TABLES **0**

Partial Least Squares Regression in Payment Default Prediction

Erkki K. Laitinen

Abstract

Payment default is affected by many interrelated factors. When concentrating on financial data, payment default can be predicted by profitability, growth, liquidity, solidity, and size variables. Usually, these financial variables are many, strongly correlated and non-normally distributed. These difficulties can be reduced by the (orthogonal) factor analysis, which identifies independent latent variables (factors) explaining a greater part of variation in predictors. However, the partial least squares (PLS) regression finds a few independent factors that most efficiently explain variation in both predictors and response. The purpose of the study is to analyse the performance of the PLS in payment default prediction. The data consist of eight financial variables from 1500 default and 1500 non-default firms. The original financial variables, Varimax-rotated factors, and PLS-factors are used in the logistic regression models to predict payment default one year prior to the event. It is showed that three Varimax-rotated factors or only two PLS-factors can effectively substitute eight original financial ratios as predictors. Each of the three models will lead to performance equal in terms of classification accuracy. When the sample size is remarkably reduced, the efficiency of the PLS-factors will become more obvious.

Key words: Payment default prediction, financial ratios, Partial Least Squares regression.
JEL classification: M Business Administration and Business Economics Marketing; Accounting, M4 Accounting, M41 Accounting.

1. Introduction

Typical statistical methods in payment default (failure) prediction include regression analysis, linear discriminant analysis, logit analysis, recursive partitioning, and neutral networks (for reviews see Zavgren 1983; Jones, 1987; Laitinen and Kankaanpää, 1999; and LeClere, 2000). Irrespective of the statistical method, the main difficulties in prediction are due to that payment default is affected by many interrelated, non-normally distributed financial variables (see Richardson & Davidson, 1983 and Karels & Prakash, 1987). Many researchers have used the factor analysis to solve these problems (see Pinches, Mingo & Caruthers, 1973; Taffler, 1982; and Skogsvik, 1990). Factor analysis identifies latent variables (factors) explaining most efficiently the variation in *predictors*. However, it can lead to a large number of latent variables that are difficult to interpret. Skogsvik (1990), for example, applied the factor analysis separately to 71 standard financial ratios and found seventeen factors. However, the Partial Least Squares (PLS) regression is a method that finds a few factors that most efficiently explain the variation in both *predictors* and *response*. Thus, the resulted latent factors may be fewer and easier to interpret. *The purpose of the study is to analyse the performance of the PLS in payment default prediction.*

The partial least squares method was originally developed in the 1960s by the econometrician Herman Wold (1966) for modelling paths of causal relation between any numbers of blocks of variables. It became popular first in chemo metrics (see Wold & Dunn, 1983). Nowadays it is popular also in social sciences (Martens, 2001; and Abdi, 2003). *However, PLS is not applied in payment default prediction.* First of all, PLS is a method for constructing predictive models, when the factors are many and highly collinear (Geladi and Kowalski, 1986). PLS may be the least restrictive of the various multivariate extensions of the multiple linear regression models. Thus it can be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than predictor variables. Furthermore, PLS can be used as an exploratory analysis tool to select suitable predictor variables. The algorithm used by PLS exam-

ines both independent and dependent variable data and extracts factors, which are directly relevant to both sets of variables. These are extracted in decreasing order of relevance. So, to form a model, the most important thing is to extract the correct number of factors to model relevant underlying effects.

The objective of this paper is thus to demonstrate the use of the factor analysis, especially of PLS, in predicting payment default in Finnish data. The financial data base has been obtained from Suomen Asiakastieto Oy (Finska Ltd, see http://www.asiakastieto.fi). It includes financial ratios from 1500 default firms and 1500, randomly selected non-default firms. On a basis of a hypothetical model, eight financial variables are selected for predicting payment default. All the statistical analyses are made by the SAS package. First, payment default is predicted by the logistic regression analysis (LRA) using the original eight variables to give a benchmark. Secondly, the factor analysis with a Varimax (orthogonal) rotation is applied to the eight variables to reduce dimensions in prediction. The extracted three factors are used as independent variables to predict payment default in the LRA (Figure 1). Thirdly, PLS is used to find relevant factors which are applied by the LRA. The results show that extracted three Varimax-rotated factors or only two PLS-factors as predictors lead to an equal classification accuracy as the eight original variables which are highly correlated. *Thus, especially the PLS provides us with a powerful method to reduce dimensions in default prediction*. The study is organized as follows. The second section describes the selection of the eight original variables, the data, and the LRA results for the eight variables. The third sector presents the results for the factor analysis and the associated LRA. The fourth section shows similar results for the PLS. Finally, the fifth section shortly summarizes the study.

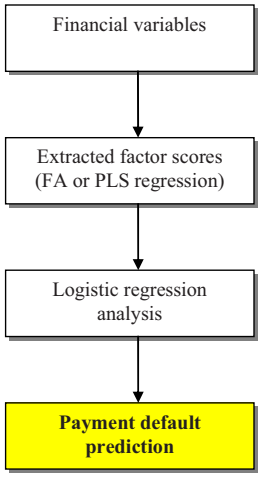

Fig. 1. Prediction of the payment default in this study

2. Variables, data, and logistic regression analysis

2.1. Choice of financial variables

Payment default is usually a result of a multi-year process leading to financial difficulties. Although there are many types of payment default, this phenomenon can in general terms be defined as the inability of the firm to pay its financial obligations when they come due (see for example Beaver, 1966 and Altman, 1968). At this stage of insolvency the firm has not financial resources enough and is unable to get such resources immediately, to pay the mature obligations in time. The reasons for the start of the process are often associated with the relationship between growth and profitability (see Laitinen, 1991). The higher the level of the annual cash flow (before interest and taxes) is, the higher is the profitability of the firm *ceteris paribus*. In addition, the lower this flow is; the higher is the rate of growth *ceteris paribus*. Therefore the process may start when there is an exceptionally large positive difference between growth and profitability (high growth rate & low profitability). This

may be due to a fast growth strategy or to a diminished profitability, or to both. Consequently, the cash flow as a measure of revenue finance will be low and the firm is not able to pay taxes and interest expenses without outside financing that is typically debt.

Thus the firm will take more debt. If the cash flow (revenue finance) continues to be at a low level, the firm is running into a vicious circle. It needs more debt to pay its taxes and interest expenses which leads to a deeper indebtedness and higher interest expenses. When approaching the moment of default, the firm may be so indebted that it will not get long-term debt due to the lack of securities. Thus, at the final stages of the default process, the firm tries to get more current debt to avoid a default of payments. Finally, its financial assets become very scarce (critical) because they have been used to pay financial obligations. Simultaneously, if the firm does not get any additional current debt, it has no financial assets enough, or possibilities to get more debt, to pay the mature obligations. This situation obviously leads to a default of payments. The size of the firm may affect the process at least at the final stages because larger firms have more resources to avoid default in this situation. This default process is outlined in Figure 2.

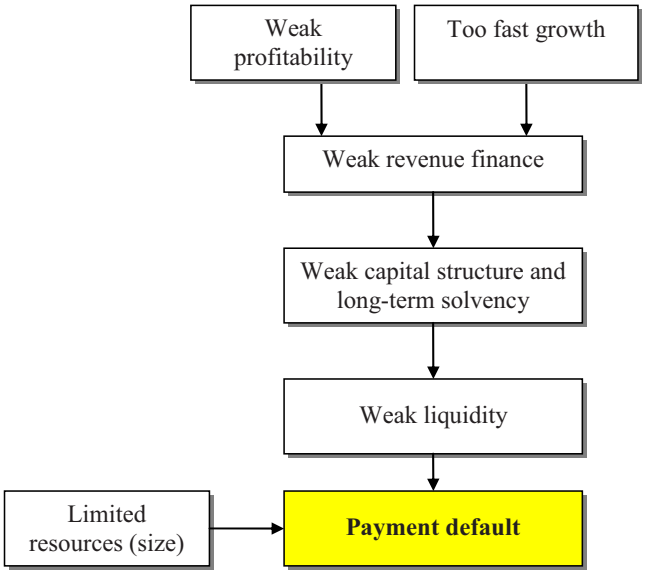

Fig. 2. Payment default process of a firm

The financial variables to be used to develop the default prediction model are chosen on the basis of the default process described above. All in all, eight financial variables are chosen. The *growth* of the firm is measured only by the percentage annual change in net sales while there are two measures for the *profitability*: the return on investment and the net profit to net sales ratios. Quick ratio is employed as the measure of traditional short-term *liquidity*. Two different traditional cash flow measures are applied, that is the traditional cash flow to net sales and the traditional cash flow to total debt ratios. The first of these measures refers to *revenue finance* and the second one to *long-term solvency*. Finally, the equity ratio (shareholder capital to total debt ratio) is applied to measure the *solidity* (indebtedness, capital structure) of the firm whereas the logarithmic net sales represent for the *size*. The chosen variables are largely comparable with the variables used in previous failure studies (see Mossman, Bell, Swartz, and Turtle, 1998 and Turetsky and McEwen, 2001: 325-326).

2.2. The data of the study

The empirical data which have been used in the study contain the eight financial variables from altogether 1500 default and 1500 non-default firms. The payment defaults of the firms have

taken place during the years 1998-2003. A firm is regarded as a default firm when even one payment disturbance has been officially registered during time period of 1998-2003. The non-default firms will not have registered payment disturbances by the end of the year 2003. The eight variables that have been calculated of the financial statements are used as predictor variables. The observation material contains the financial data of firms from the years 1997-2001. The values of the variables have been calculated from the financial statements of the year which precedes the payment disturbance. Thus, there are 0-12 months of time to the payment disturbance. *This means that the prediction model will be based on the values of the predictors at the final stage of the default process.* The distributions of the variables have been truncated so that the natural upper limits and lower limits have been set on them. This procedure diminishes the effect of outliers and improves the normality of the variables.

Table 1 shows the averages and standard deviations of all eight variables in default and non-default firms. The default firms are smaller on average, grow faster, and show weaker profitability, and their solidity and liquidity are weak compared to non-default firms. *The average growth rate of the default firms is as much as 25.2% even though the average return on investment is only 10.6%.* This result refers to the difference between growth and profitability as a source for default process. Their net profit ratio is negative (-3.8%) on average and the quick ratio is below unity (0.91) while the equity ratio is near zero (5.6%). In the non-default firms the average return on investment is 23.7% which exceeds the growth rate (12.6%) distinctly. Their average quick the ratio is 2.5 and the equity ratio over 40 (42.8%). The differences between the groups are statistically extremely significant on every variable measured with a T test. This is indeed expected because of the large sample size. The clearest differences are in the equity ratio and in the cash flow to debt ratio. For every variable, the normality assumption can be rejected on the basis of the Kolmogorov-Smirnov D test. However, this test is very sensitive to small deviations from normality due to the large sample.

Table 2 shows the correlation coefficients between the predictor variables. The table also includes the binary default status as a variable so that 0 refers to a non-default firm and 1 to a default firm. *This status variable has a statistically significant correlation to all eight predictors*. The highest correlations are to the equity ratio (-0.44) and to the cash-flow to debt ratio (-0.28). *In addition, as expected, there exist several high correlations between the predictors.* The correlations of the logarithmic net sales to other variables are not high. However, it seems to depend on the net profit to net sales ratio (0.11). The growth rate has the highest correlation to the return on investment ratio (0.14). The return on investment ratio has several high correlations to other predictors which shows the importance of profitability to the economic performance of the firm. It has the highest correlations to the cash-flow to debt ratio (0.55), to the net profit to net sales ratio (0.51) and to the cash-flow to net sales ratio (0.44). The net profit to net sales ratio shows still higher correlations to the cash-flow to debt ratio (0.56) and to the cash-flow to net sales ratio (0.89).

The liquidity measure, quick ratio, depends on the cash-flow to debt ratio (0.34) and the equity ratio (0.34) which respectively measure the long-term solvency and solidity of the firm. The cash-flow to net sales ratio strongly depends on the net profit to net sales ratio (0.89), on the cash-flow to debt ratio (0.55) and on the return on investment ratio (0.44). The equity ratio has the highest correlation expectedly to the cash-flow to debt ratio (0.51) but other strong dependences also are found. The cash-flow to debt ratio exceptionally strongly depends on the net profit to net sales ratio (0.56), on the cash-flow to net sales ratio (0.55), on the return on investment ratio (0.55), and on the equity ratio (0.51). *Thus the correlations between the predictor variables deviate from zero statistically extremely significantly and they cannot be considered independent of each other as several statistical methods suppose.*

Table1

Descriptive statistics of the original financial variables (N = 1500 + 1500)

	Non-default firms:				Default firms:				Comparison of groups:	
	Mean	Standard deviation	Kolmogorov-Smirnov D test	Probability level of D	Mean	Standard deviation	Kolmogorov-Smirnov D test	Probability level of D	T Statistic	Probability level of T
Logarithmic net sales	6.341	1.886	0.0286	<0.0100	5.789	1.726	0.0489	<0.0100	8.37	<.0001
Growth in net sales (%)	12.563	45.635	0.1813	<0.0100	25.207	65.384	0.1842	<0.0100	-6.14	<.0001
Return on investment ratio (%)	23.678	41.035	0.1355	<0.0100	10.603	58.174	0.1257	<0.0100	7.11	<.0001
Net profit to net sales (%)	4.134	21.288	0.2235	<0.0100	-3.800	21.441	0.1952	<0.0100	10.17	<.0001
Quick ratio	2.536	5.866	0.3328	<0.0100	0.907	1.608	0.2864	<0.0100	10.37	<.0001
Cash flow to net sales (%)	10.487	23.449	0.2010	<0.0100	1.186	21.889	0.1830	<0.0100	11.23	<.0001
Equity ratio (%)	42.827	35.464	0.0586	<0.0100	5.602	41.455	0.1610	<0.0100	26.43	<.0001
Cash flow to debt (%)	42.871	67.833	0.1634	<0.0100	9.731	43.687	0.1501	<0.0100	15.91	<.0001

Table 2

Pearson correlation coefficients between the original financial variables and their significance levels

	Default status	Logarithmic net sales	Growth in net sales (%)	Return on investment ratio (%)	Net profit to net sales (%)	Quick ratio	Cash flow to net sales (%)	Equity ratio (%)	Cash flow to debt (%)
Default status	1.0000	-0.1510	0.1115	-0.1288	-0.1826	-0.1861	-0.2009	-0.4347	-0.2790
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Logarithmic net sales	-0.1510	1.0000	0.0562	0.0907	0.1091	-0.0803	0.0037	0.0942	0.0284
	<.0001		0.0021	<.0001	<.0001	<.0001	0.8381	<.0001	0.1198
Growth in net sales (%)	0.1115	0.0562	1.0000	0.1399	0.0397	-0.0752	0.0288	-0.0530	0.0303
	<.0001	0.0021		<.0001	0.0299	<.0001	0.1147	0.0037	0.0971
Return on investment ratio (%)	-0.1288	0.0907	0.1399	1.0000	0.5119	0.0424	0.4407	0.2960	0.5468
	<.0001	<.0001	<.0001		<.0001	0.0201	<.0001	<.0001	<.0001
Net profit to net sales (%)	-0.1826	0.1091	0.0397	0.5119	1.0000	0.1763	0.8943	0.3235	0.5623
	<.0001	<.0001	0.0299	<.0001		<.0001	<.0001	<.0001	<.0001
Quick ratio	-0.1861	-0.0803	-0.0752	0.0424	0.1763	1.0000	0.1585	0.3384	0.3443
	<.0001	<.0001	<.0001	0.0201	<.0001		<.0001	<.0001	<.0001
Cash flow to net sales (%)	-0.2009	0.0037	0.0288	0.4407	0.8943	0.1585	1.0000	0.3203	0.5479
	<.0001	0.8381	0.1147	<.0001	<.0001	<.0001		<.0001	<.0001
Equity ratio (%)	-0.4347	0.0942	-0.0530	0.2960	0.3235	0.3384	0.3203	1.0000	0.5065
	<.0001	<.0001	0.0037	<.0001	<.0001	<.0001	<.0001		<.0001
Cash flow to debt (%)	-0.2790	0.0284	0.0303	0.5468	0.5623	0.3443	0.5479	0.5065	1.0000
	<.0001	0.1198	0.0971	<.0001	<.0001	<.0001	<.0001	<.0001	

2.2. Logistic regression analysis based on original variables

There are many different statistical methods which can be applied in developing a default prediction model on the basis of the eight predictors. The method applied here should be as simple as possible to show the importance of factorization. The ordinary linear discriminant or regression analyses are not very recommendable here because they require normality and independence of the predictors and linearity in relation to the default risk. Therefore, the logistic regression analysis in which an attempt is made to predict the probability of non-default is used in default prediction. This method does not require the normality of the variables and not for example homoscedasticity of the model. However, the logistic regression analysis requires in the same way as the linear models that the independent predictor variables do not depend on each other. *The model supposes thus that multicollinearity will not appear*. In this case it cannot be avoided if all the predictor variables are included in the same model. Technically, the dependence between the variables should not affect the estimates of the coefficients of the model but it weakens their reliability (it increases the standard deviation of the estimates).

With the help of the logistic regression model, the following non-default probability can be estimated to the firm *i*:

$$P(\textit{non-default, } i) = 1 / [1+\exp(-Z(i))], \tag{1}$$

where $Z(i)$ is a linear logit estimated to the firm *i*. This logit will be estimated as follows

$$Z(i) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots \text{ or} \tag{2a}$$

$$Z = X b, \tag{2b}$$

where b_j (j=0,1, ... , m) are the parameters of the logistic regression model and X_{ji} is the value of the predictor variable j for the firm i (j=1,2, ... , m and i=1,2, ... , n). In (2b), Z is a $n \times 1$ logit vector, X is a $n \times (m+1)$ value matrix where the first column is a unit vector, and b is a $(m+1) \times 1$ coefficient vector. For the present data, m = 8 and n = 3000.

Table 3 shows the logistic regression model in which all the original eight variables are included. This solution is based on the maximum likelihood estimation and is calculated by the PROC LOGISTIC of the SAS (http://www.sas.com/). In this model only the estimate of the coefficient for the return on investment ratio does not differ statistically significantly from zero. The distinctly most significant factor in the model is the equity ratio with the Chi-Square test statistic of 230.3. The logarithmic net sales have the next highest value of the test statistic but it remains already distinctly smaller (70.9). The good significance of the coefficients of the LR model is expected because of the large data. The concordance coefficient of the model is 79.7 so that the goodness of the fit of the model to the data is quite good. *In spite of the high goodness of fit the interpretation and use of the LR model are not unproblematic.* The estimation gave negative coefficients for the return on investment ratio and for the net profit to net sales ratio even though it is intuitively clear that their effect on the probability of non-default is positive. When the estimated model is used in practice, it may happen according to the model that when a target firm improves its profitability (*ceteris paribus*), it at the same time increases its default risk.

The LR model is extracted from a large data base and only includes eight predictor variables. Thus, it is possible that the model can be generalized in spite of the dependences between the predictors. This can be tested by validation. Table 3 also shows the validated classification results for the estimated LR model, when the Lachenbruch cross-validation (jackknife) method is applied. The Lachenbruch "leave-one-out" method uses n-1 observations, develops the LR model, and then classifies the remaining one. Because the present analysis contains an equal number of default and non-default firms, the theoretically plausible critical value of the non-default probability is near 0.50. In this case the validated classification accuracy of the model rises to about 72-73% when the critical value is located around 0.50. *The best total classification accuracy is obtained with critical value 0.46 with which altogether 73.0% of the firms is correctly classified*. The LR model correctly classifies 78% of non-default firms and 68.1% of default firms. Because incor-

rectly classified default firms are more costly to the potential users of the model than such non-default firms, the most useful critical value exceeds 0.50. For example, the model correctly classifies 79.1% of the default firms and 65.9% of the non-default firms, when a critical value of 0.54 is applied. *In this case the total classification accuracy is thus 72.5%.*

Table 3

Results of the logistic regression analysis based on the eight original variables

Panel 1. Test statistics

Test statistic	Chi-Square	Significance level
Likelihood ratio	798.082	<.0001
Score	671.868	<.0001
Wald	499.446	<.0001

Somers' D	0.595
Gamma	0.596
Tau-a	0.298
Concordance coefficient	79.7

Panel 2. Logistic regression model

Variable	Estimate	Standard error	Wald Chi-Square	Significance level
Constant	-2.061	0.173	141.804	<.0001
Logarithmic net sales	0.207	0.025	70.945	<.0001
Growth in net sales (%)	-0.004	0.001	28.376	<.0001
Return on investment ratio (%)	-0.001	0.001	1.508	0.2194
Net profit to net sales (%)	-0.020	0.005	13.552	0.0002
Quick ratio	0.059	0.023	6.825	0.0090
Cash flow to net sales (%)	0.023	0.005	23.369	<.0001
Equity ratio (%)	0.024	0.002	**230.302**	<.0001
Cash flow to debt (%)	0.003	0.001	5.810	0.0159

Panel 3. Classification accuracy of the model

Correct classifications (%):

Critical value	All firms	Non-default firms	Default firms
0.44	72.4	80.7	64.1
0.46	73.0	78.0	68.1
0.48	72.8	74.6	70.9
0.50	**72.3**	**71.5**	**73.2**
0.52	72.5	68.9	76.1
0.54	72.5	65.9	79.1
0.56	72.1	62.6	81.7

3. Factor analysis and LRA based on extracted factors

3.1. Factor analysis

Factor (principal component) analysis was carried out by the PROC FACTOR of the SAS (http://www.sas.com/). In the principal component analysis the factors are chosen to explain as much as possible the variance of X where X describes here the 3000×8 value matrix of independ-

ent variables. The number of components extracted can be decided on the basis of the eigenvalue of the successive components. This eigenvalue determines the coefficient of determination for a factor solution. When this eigenvalue is divided with the number of the variables (here 8), the coefficient of determination for the factor will be obtained as a result. This coefficient refers to how many per cent the extracted factors of the solution (the latent variables) account for the variation of the eight original variables. The more they account for this variation the better the new latent variables include the information contained by them. Table 4 shows the eigenvalue and the coefficient of determination of the (principal component) model for each number of factors. The first factor alone explains already 39.2% of the variation of the original variables. The eigenvalue of the third factor is about unity which often is used for the model as a critical value to the including of a factor. *When a model consists of three factors, its coefficient of determination will be 67.3% which is quite good.* The following fourth factor raises the explanation degree still 11.7%. However, the model is here marked off to three factors on the basis of the eigenvalue criterion.

Table 4

Results of the factor (principal component) analysis

Panel 1. Eigenvalues of the factors

Factor	Eigenvalue	Difference	Rate of determination	Cumulative rate of determination
1	3.1346	1.8865	0.3918	0.3918
2	1.2481	0.2483	0.1560	0.5478
3	0.9998	0.0630	0.1250	0.6728
4	0.9368	0.2398	0.1171	0.7899
5	0.6970	0.1548	0.0871	0.8770
6	0.5422	0.1955	0.0678	0.9448
7	0.3467	0.2517	0.0433	0.9881
8	0.0950		0.0119	1.0000

Panel 2. Loadings of the variables on factors (Varimax-rotated)

	Factor 1	Factor 2	Factor 3
Logarithmic net sales	0.0122	-0.0837	**0.9455**
Growth in net sales (%)	0.2061	**-0.5213**	0.1589
Return on investment ratio (%)	0.7357	-0.0809	0.1623
Net profit to net sales (%)	**0.9006**	0.0342	0.0093
Quick ratio	0.1564	**0.7707**	-0.0526
Cash flow to net sales (%)	**0.8852**	0.0401	-0.1063
Equity ratio (%)	0.4135	**0.6124**	0.3019
Cash flow to debt (%)	0.7446	0.3706	0.0934
Eigenvalue	2.9283	1.3945	1.0596

Table 4 also shows the correlations of the factors to the original eight variables which can be used in identifying the object of measurement for the factors. These correlations are called the loadings of the factors on the variables. The factors are rotated by the Varimax method which makes the resulted factors linearly independent of each other. Thus the coefficient of correlation between them is zero and there is no multicollinearity when these factors are employed as predictors of default. These factors are also normalized so that their mean is zero and standard deviation is unity. On the basis of the highest loadings the first latent variable (Factor 1) mainly refers to *profitability*, *revenue finance*, and *long-term solvency* (cash flows). However, the second factor (Factor 2) deals with *growth*, *liquidity*, and *capital structure*. The loading on growth is negative hence referring to that quick growth will decrease the value of the factor (increase the estimate for

the non-default probability). The third latent variable (Factor 3) measures above all the *size* of the firm. When normality is tested by the D test, it is rejected for all the latent variables (due to the sensitivity of the test to the sample size).

3.2. Logistic regression analysis

The original variables in the LR model can be replaced by the extracted orthogonal factors (factor scores). As for any such method, this replacement essentially improves the usability of the estimated model. In this version of model the eight interdependent variables are reduced to three latent variables which are normalized and linearly independent of each other. Table 5 shows the solution of the LR model when it has been estimated for the three latent variables as predictors of default. The coefficient of concordance for the LR model is now 78.1 referring to high goodness of fit. The estimated coefficients for all three factors are positive and extremely significant statistically (due to the large sample size). *The coefficient of the second latent variable (Factor 2) which measures growth, liquidity and capital structure, has the highest value of the Chi-Square statistic (309.4).* It also has the highest weight on the value of predicted non-default probability. In this case the regression coefficient directly gives the idea of the weight of the variable because the scales of the factors are normalized.

Table 5 also shows the validated (Lachenbruch) classification results for the LR model based on the factor scores. The best total classification accuracy is obtained by the critical value 0.52 of the non-default probability in which case the model classifies correctly altogether 72.3% of the observations. When paying attention to the classification costs, a better result may however be given by the critical value 0.50 which correctly classifies 70.8% of the non-default firms and 72.5% of the default firms. Thus it correctly classifies altogether 71.7% of all the firms. Thus the validated classification accuracy of the LR analysis which is based on the extracted three factors, is about on the same level as in the analysis which is based on eight original variables. *Thus the factorization and rotation carried out, do not destroy information needed in the classification of default and non-default firms but they improve the statistical usability of the variables essentially.*

Table 5

Results of the logistic regression analysis based on the three factors

Panel 1. Test statistics

Test statistic	Chi-Square	Significance level
Likelihood ratio	723.008	<.0001
Score	567.582	<.0001
Wald	461.993	<.0001

Somers' D	0.564
Gamma	0.565
Tau-a	0.282
Concordance coefficient	78.1

Panel 2. Logistic regression model

Variable	Estimate	Standard error	Wald Chi-Square	Significance level
Constant	0.056	0.042	1.834	0.1756
Factor 1	0.601	0.048	158.701	<.0001
Factor 2	1.177	0.067	**309.369**	<.0001
Factor 3	0.488	0.044	123.880	<.0001

Table 5 (continuous)

Panel 3. Classification accuracy of the model

Correct classifications (%):

Critical value	All firms	Non-default firms	Default firms
0.44	70.4	78.8	62.1
0.46	70.6	75.9	65.3
0.48	71.7	73.7	69.7
0.50	**71.7**	**70.8**	**72.5**
0.52	72.3	67.7	76.9
0.54	72.0	64.3	79.6
0.56	71.1	60.1	82.1

4. PLS regression analysis and LRA based on extracted factors

4.1. PLS regression analysis

In the factor (principal component) analysis the factors are chosen to explain as much as possible the variance of *X*. However, it is not guaranteed that the extracted factors are relevant for *Y*, which here describes the *3000 × 1* value vector of the dependent variable. By contrast, the partial least squares (PLS) regression analysis searches for a set of components (latent vectors) that performs a simultaneous decomposition of *X* and *Y* with the constraint that these components explain as much as possible of the covariance between *X* and *Y* (see Abdi 2003: 2). *Thus in this case the PLS method does not extract a general factor solution (for X) as above but looks for the solution which predicts the payment default (Y) in the most efficient way.* The PLS regression analysis was carried out by the PROC PLS of the SAS. All the PLS estimations are based on the usual iterative NIPALS algorithm (http://www.sas.com/). Table 6 shows the validation test (PRESS, for predicted residual sum of squares) which is based on random subset selection (CV = RANDOM) and can be used to define the optimal number of PLS factors when predicting payment default. The number of the factors should be as small as possible but anyway sufficient. According to this test only *two* efficient latent variables (PLS factors) are extracted from the eight original variables. The table also shows how much these PLS factors account for the variance of the original variables and for the variance of the binary default status. *The extracted two PLS factors account altogether for 52.0% of the eight original variables and 21.6% of the default status.* Panel 3 of the table shows that the factor scores are not normally distributed, according to the D test. There are statistically significant differences in the mean of the scores between default and non-default firms (see T test).

Table 6

Results of the PLS regression analysis

Panel 1. Test of the number of the PLS factors

Factor	PRESS	Significance level
0	1.171019	<.0001
1	1.078913	<.0001
2	**1.029207**	**1.000000**
3	1.039696	<.0001
4	1.038167	<.0001
5	1.034346	<.0001
6	1.034304	<.0001
7	1.033952	<.0001
8	1.033933	<.0001

Table 6 (continuous)

Panel 2. Pearson correlation coefficients between the original variables and the PLS factor scores

	PLS factor 1	PLS factor 2
Default status	0.4081	0.2225
	<.0001	<.0001
Logarithmic net sales	-0.2017	-0.3082
	<.0001	<.0001
Growth in net sales (%)	0.0961	**0.4240**
	<.0001	<.0001
Return on investment ratio (%)	-0.5716	0.4590
	<.0001	<.0001
Net profit to net sales (%)	**-0.7365**	**0.5081**
	<.0001	<.0001
Quick ratio	-0.4731	-0.1622
	<.0001	<.0001
Cash flow to net sales (%)	-0.7072	**0.5062**
	<.0001	<.0001
Equity ratio (%)	**-0.7843**	**-0.4253**
	<.0001	<.0001
Cash flow to debt (%)	**-0.8099**	0.1980
	<.0001	<.0001

The PLS factors that have been extracted with the help of the PLS analysis are linearly independent of each other in the same way as the three principal component scores above. The PLS scores have also been normalized (centralized) so that their average will be zero. Panel 2 of Table 6 shows the Pearson correlations of the two PLS factors to the original eight variables (that is, loadings). The first latent variable (PLS factor 1) correlates strongly above all to long-term solvency, capital structure, profitability, and revenue finance. The second PLS factor, however, depends especially on growth, profitability and revenue finance. The correlations of this factor to profitability and revenue finance are positive whereas they are negative for the PLS factor 1. *Thus the PLS factor 2 efficiently accounts for the residual that is not accounted for by the PLS factor 1.* This is also the technical idea of the solution for the successive factors (see Abdi, 2003).

4.2. Logistic regression analysis

The PLS method also contains a regression analysis and estimates the coefficients of the regression model directly when extracting the factors. In this context the extracted PLS factors (scores), however, are adapted as separate variables in the logistic regression analysis in the same way as above. Table 7 shows the results of the LR analysis when the extracted PLS factors have been used as independent variables. Even though there are only two independent variables in the LR model, its statistical properties are extremely good. The goodness of fit for the model is represented by the high concordance coefficient which in this case is 79.0. The both PLS factors have statistically very significant estimates for the coefficients. The first PLS factor has got a Chi-Squared test statistic of 420.9 which distinctly exceeds the values which have appeared in earlier models. The model correctly classifies 72.6% of all the firms (with critical values 0.52 and 0.54 for the non-default probability) in the Lachenbruch validation test. Paying attention to the classification error costs, a good result is also obtained by the theoretically plausible critical value (0.50). When using it, the LR model correctly classifies 72.9% of the default firms and 71.7% of the non-default firms (72.3% of all the firms). *Thus, the PLS regression analysis is able to extract two factors which provides us with the LR model that reach about the same level of classification accuracy as the eight original variables or the three principal component factors.*

Table 7

Results of the logistic regression analysis based on the two PLS factors

Panel 1. Test statistics

Test statistic	Chi-Square	Significance level
Likelihood ratio	770.333	<.0001
Score	648.126	<.0001
Wald	494.698	<.0001

Somers' D	0.581
Gamma	0.582
Tau-a	0.291
Concordance coefficient	79.0

Panel 2. Logistic regression model

Variable	Estimate	Standard error	Wald Chi-Square	Significance level
Constant	-0.006	0.042	0.021	0.8854
PLS factor 1	-0.689	0.034	**420.853**	<.0001
PLS factor 2	-0.551	0.042	171.263	<.0001

Panel 3. Classification accuracy of the model

	Correct classifications (%)		
Critical value	All firms	Non-default firms	Default firms
0.44	71.3	79.5	63.0
0.46	71.9	77.1	66.7
0.48	72.0	74.4	69.6
0.50	**72.3**	**71.7**	**72.9**
0.52	72.6	69.3	75.9
0.54	72.6	66.2	78.9
0.56	71.9	62.7	81.2

5. Summary

There are many background and financial variables which affect the payment default risk of the firm and which strongly correlate with each other. This multicollinearity problem seriously weakens the reliability and generalization of default prediction models. The ordinary factor analysis is useful in these models to reduce the number of predictor variables and to make them linearly independent of each other. In addition, the factor analysis may result in factors which follow the normal distribution better than the original variables. The purpose of this study was to show how the ordinary factor analysis and especially the PLS regression analysis can be used to reduce the number of independent variables in default prediction. The PLS regression analysis is an efficient method for constructing predictive models, when the variables are many and highly collinear. *It has been widely adopted also in social sciences and economics but is not adopted in payment default prediction.*

In this study the factor analysis and the PLS regression analysis were used in the financial data from 3000 firms to extract factors employed in the LR analysis. The benchmark LR model was based on eight financial variables which led to a total classification accuracy of 73.0% in a Lachenbruch validation test. An ordinary factor (principal component) analysis with an orthogonal (Varimax) solution made it possible to extract three latent variables (factors) which are linearly independent of each other. These factor scores correctly classified altogether 72.3% of the firm.

However, with the aid of the PLS regression analysis, it was possible to extract only two relevant factors which in the LR analysis correctly classified 72.6% of the firms. *Thus, the analyses showed that the factor analysis is in payment default prediction an efficient statistical method to reduce the number of predictor variables and to eliminate the dependence between those variables without any essential impairment in the classification accuracy. In this respect the PLS regression analysis is even more efficient than the ordinary factor analysis.*

The flexibility of the PLS regression makes it advantageous to use it in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than predictor variables. The benefits of the PLS are thus obvious, when small samples are considered. Table 8 presents Lachenbruch validated classification results when the LR analysis is applied for alternative sample sizes randomly selected from the present data. The critical value in this table has been chosen to give the best classification result provided that the classification accuracy for the default firms is at least 70% (if possible). The total classification accuracy of the LR based on the eight original variables reduces only slightly when the sample size is diminished from 3000 to 200 and to 100. However, for the sample size 50 the model does not work at all. *The LR model based on the PLS-factor score shows the highest classification accuracy for the sample sizes 200 and 100.* When the sample size is 50, the accuracy is remarkably diminished but is still at the level of 68.0%. It evidently gives more reliable results than the model based on the original variables.

Table 8

Classification accuracy for alternative sample sizes

Panel 1. Sample size = 1500+1500

	Percent of correct classifications:			
	Critical value	All firms	Non-default firms	Default firms
a) Original variables (8)	0.48	72.8	74.6	70.9
b) Factor scores (3)	0.52	72.3	67.7	76.9
c) PLS factor scores (2)	0.52	72.6	69.3	75.9

Panel 2. Sample size = 100+100

	Percent of correct classifications:			
	Critical value	All firms	Non-default firms	Default firms
a) Original variables (8)	0.46	71.5	73.0	71.0
b) Factor scores (3)	0.46	73.5	77.0	70.0
c) PLS factor scores (2)	0.44	75.0	79.0	71.0

Panel 3. Sample size = 50+50

	Percent of correct classifications:			
	Critical value	All firms	Non-default firms	Default firms
a) Original variables (8)	0.50	71.0	66.0	76.0
b) Factor scores (3)	0.54	67.0	62.0	72.0
c) PLS factor scores (2)	0.52	74.0	66.0	82.0

Panel 4. Sample size = 25+25

	Percent of correct classifications:			
	Critical value	All firms	Non-default firms	Default firms
a) Original variables (8)	0.50	56.0	56.0	56.0
b) Factor scores (3)	0.52	68.0	64.0	72.0
c) PLS factor scores (2)	0.48	68.0	64.0	72.0

References

1. Abdi, H. (2003). Partial Least Squares (PLS) regression. In M. Lewis-Beck, A. Bryman and T. Futing (Eds.). *Encyclopedia of Social Sciences Research Methods*. Sage. Thousands Oaks (CA).
2. Altman, E.I. (1968). Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy. *The Journal of Finance* 23, 4, 589-609.
3. Beaver, W.H. (1966). Financial Ratios as Predictors of Failure. Empirical Research in Accounting: Selected Studies. *Journal of Accounting Research*. Supplement to Vol. 4, 71-127.
4. de Jong, S. (1993). Simpls. An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* 18, 251-253.
5. Garthwaite, P.H. (1994). An interpretation of partial least squares. *Statistical Journal of America's Statistical Association* 89, 122-127.
6. Geladi, P. & B. Kowalski (1986). Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta* 185, 1-17.
7. Haenlein, M. & A.M. Kaplan (2004). A Beginner's Guide to Partial Least Squares Analysis. *Understanding Statistics* 3, 4, 283-297.
8. Helland, I. (1988). There is a the structure of partial least squares. *Communication in Statistics, Simulation and Computation* 17, 581-607.
9. Jones, F. (1987). Current Techniques in Bankruptcy Predicting. *Journal of Accounting Literature* 6, 131-164.
10. Karels, G.V. & A.J. Prakash (1987). Multivariate Normality and Forecasting of Business Bankruptcy. *Journal of Business Finance & Accounting*, Winter, 573-592.
11. Laitinen, E.K. (1991). Financial Ratios and Different Failure Processes. *Journal of Business Finance and Accounting* 18, 3, 649-674.
12. Laitinen, T. & M. Kankaanpää (1999). Comparative Analysis of Failure Prediction Methods: The Finnish Case. *The European Accounting Review* 8, 1, 67-92.
13. LeClere, M. (2000). The Occurrence and Timing of Events: Survival Analysis Applied to the Study of Financial Distress. *Journal of Accounting Literature* 19, 158-189.
14. Martens, H. (2001). Reliable and relevant modelling of real world data: a personal account of the development of PLS regression. *Chemometrics and Intelligent Laboratory Systems* 58, 85-95.
15. Mossman, C.E., G.G. Bell, L.M. Swartz, & H. Turtle (1998). An Empirical Comparison of Bankruptcy Models. *The Financial Review* 33, 35-54.
16. Pinches, G.E., K.A. Mingo, & J.K. Caruthers (1973). The Stability of Financial Patterns in Industrial Organizations. *Journal of Finance*, May, 389-396.
17. Richardson, F.M. & L.F. Davidson (1983). An Exploration into Bankruptcy Discriminant Model Sensitivity. *Journal of Business & Accounting*, 2, 195-208.
18. Skogsvik, K. (1990). Current Cost Accounting Ratios as Predictors of Business Failure: The Swedish Case. *Journal of Business Finance & Accounting*, 1, 137-160.
19. Stone, M. & R.J. Brooks (1990). Continuum regression: Cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *Journal of Royal Statistical Society* 52, 237-269.
20. Taffler, R.J. (1982). Forecasting Company Failure in the UK using Discriminant Analysis and Financial Ratio Data. *Journal of Royal Statistical Society*, 3, 342-358.
21. Turetsky, H.F. & R.A. McEwen, R.A. (2001). An Empirical Investigation of Firm Longevity: A Model of the Ex Ante Predictors of Financial Distress. *Review of Quantitative Finance and Accounting* 16, 323-343.
22. Wold, P. & W.J. Dunn (1983). Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability. *Journal of Chemical Information and Computer Sciences* 23, 1, 6-13.
23. Zavgren, C.V. (1983). The Prediction of Corporate Failure: The State of the Art. *Journal of Accounting Literature*, 1, 1-38.