Müller, Marlene; Rönz, Bernd

Working Paper

Credit scoring using semiparametric methods

SFB 373 Discussion Paper, No. 1999,93

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Müller, Marlene; Rönz, Bernd (1999) : Credit scoring using semiparametric methods, SFB 373 Discussion Paper, No. 1999,93, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10046812

This Version is available at:
https://hdl.handle.net/10419/61709

Credit Scoring using Semiparametric Methods

Marlene Müller and Bernd Rönz

November 10, 1999

Credit scoring methods aim to assess credit worthiness of potential borrowers to keep the risk of credit loss low and to minimize the costs of failure over risk groups. Standard parametric approaches as logistic discrimination analysis assume that the probability of belonging to the group of "bad" clients is given by $P(Y = 1|X) = F(\beta^T X)$, with $Y = 1$ indicating a "bad" client and X denoting the vector of explanatory variables.

We consider a semiparametric approach here, that generalizes the linear argument in the probability $P(Y = 1|X)$ to a partial linear argument. This model is a special case of the Generalized Partial Linear Model $E(Y|X, T) = G\{\beta^T X + m(T)\}$ (GPLM) which allows to model the influence of a part T of the explanatory variables in a nonparametric way. Here, $G(\bullet)$ is a known function, β is an unknown parameter vector, and $m(\bullet)$ is an unknown function. The parametric component β and the nonparametric function $m(\bullet)$ can be estimated by the quasilikelihood method proposed in Severini & Staniswalis (1994).

We apply the GPLM estimator mainly as an exploratory tool in a practical credit scoring situation. Credit scoring data usually provide various discrete and continuous explanatory variables which makes the application of a GPLM interesting here. We estimate and compare different variations of the semiparametric model in order to see how the several explanatory variables influence credit worthiness. In contrast to more general nonparametric approaches, the estimated GPLM models allow an easy visualization and interpretation of the results. The estimated curves indicate in which direction the logistic discriminant should be improved to obtain a better separation of "good" and "bad" clients.

The research for this paper was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" at Humboldt University, Berlin (Germany). Address for correspondence: Marlene Müller, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin, Germany. email: marlene@wiwi.hu-berlin.de

1 Introduction

One of the most important tasks of a bank is to assess credit worthiness of potential borrowers. The aim of this assessment is to keep the risk of a credit loss low and to minimize costs of failure over risk groups.

Typical methods which are used for the statistical classification of credit applicants are linear or quadratic discriminant analysis and logistic discriminant analysis. These methods can be seen to be based on scores which depend on the explanatory variables in a predefined form (usually linear). Recent methods that allow a more flexible modeling are neural networks and classification trees (see e.g. Arminger, Enache & Bonne, 1997) as well as nonparametric approaches (see e.g. Henley & Hand, 1996).

In the following sections we discuss for real credit scoring data, how the given explanatory variables influence credit worthiness. The following Section 2 gives a short data description. Section 3 presents the results of a logistic discrimination analysis. Section 4 describes the semiparametric extension to the logistic discrimination analysis. We estimated and compared different variations of the semiparametric model in order to see how the several explanatory variables influence credit worthiness. Section 5 compares the semiparametric fits the classic logistic analysis. Finally, Section 6 discusses the estimated models with respect to misclassification and performance curves.

2 Data Description

The analyzed data in this paper have been provided by a French bank. The given full estimation sample (denoted as **data set A** in the following) consists of 6672 cases (clients) and 24 variables:

- Response variable Y (credit worthiness, 0="good", 1="bad"). The number of "bad" clients is relatively small (400 "bad" versus 6272 "good" clients in the estimation sample).
- Metric explanatory variables X2 to X9. All of them have (right) skewed distributions. Variables X6 to X9 in particular have one realization which covers a majority of observations.
- Categorical explanatory variables X10 to X24. Six of them are dichotomous. The others have three to eleven categories which are not ordered. Hence, these variables need to be categorized into dummies for the estimation and validation.

Figure 1 shows kernel density estimates (using rule-of-thumb bandwidths) of the metric explanatory variables X2 to X9. All density estimates show the existence of outliers, in particular in the upper tails. For this reason we restricted our analysis to only those observations with X2, ... , X9 $\in [-3, 3]$. We denote the resulting data set of 6180 cases as **data set B**. The kernel density estimates for this smaller sample are shown in Figure 2.

Figure 3 shows some bivariate scatterplots of the metric variables X2 to X9. It can be clearly seen that the variables X6 to X9 are of quasi-discrete structure. We will therefore concentrate on variables X2 to X5 for the nonparametric part of semiparametric model.

In addition to the estimation sample, the bank provided us with a validation data set of 2158 cases. We denote this validation data set as **data set C** in the following. Table 1 summarizes the percentage of "good" and "bad" clients in each subsample.

	Estimation (full) **data set A**		Estimation (used) **data set B**		Validation **data set C**	
0 ("good")	6272	(94.0%)	5808	(94.0%)	2045	(94.8%)
1 ("bad")	400	(6.0%)	372	(6.0%)	113	(5.2%)
total	6672		6180		2158	

Table 1. Responses in data sets A, B and C.

3 Logistic Credit Scoring

The logit model (logistic discriminant analysis) assumes that the probability of belonging to the group of "bad" clients is given by

$$P(Y = 1|X) = F\left(\sum_{j=2}^{24} \beta_j^T X_j + \beta_0\right) \tag{1}$$

where

$$F(u) = \frac{1}{1 + \exp(-u)}$$

is the logistic (cumulative) distribution function. X_j denotes the j-th variable if Xj is metric ($j \in \{2, \ldots, 9\}$) and the vector of dummies if Xj is categorical ($j \in \{10, \ldots, 24\}$). For all categorical variables we used the first category as reference.

The logit model is estimated by maximum–likelihood. Table 2 shows the estimation results for this model. It turns out, that in fact all variables contribute more or less to the explanation of the response. The modeling for the categorical variables cannot be further improved, since by using dummies one considers all possible effects. Concerning the continuous variables, we observe nonsignificant coefficients for some regressors. The continuous variables get more attention by using semiparametric models.

4 Semiparametric Credit Scoring

The logit model (1) is a special case of the generalized linear model (GLM, see McCullagh & Nelder, 1989) which is given by

$$E(Y|X) = G(\beta^T X).$$

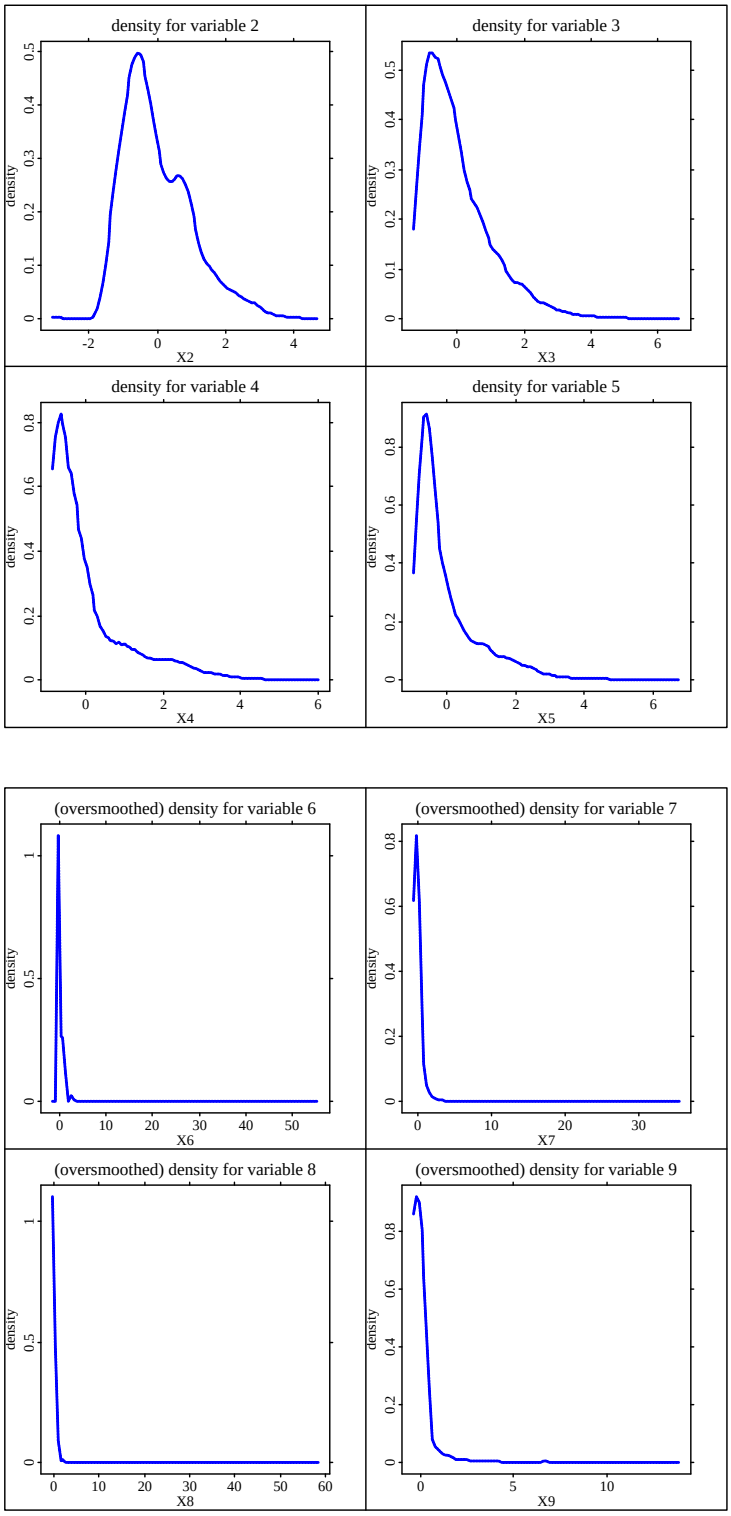


Figure 1. Kernel density estimates, variables X2 to X9, estimation data set A.

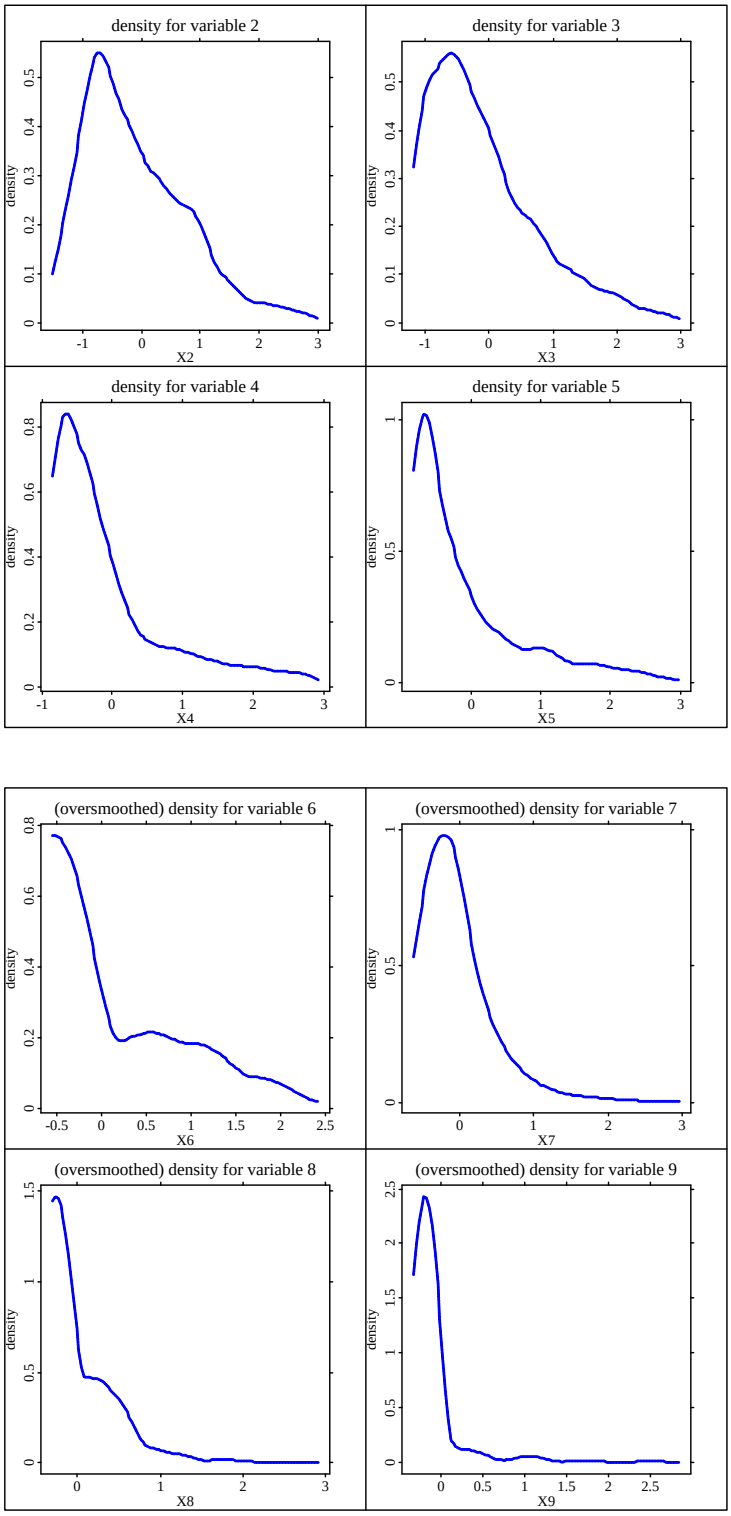
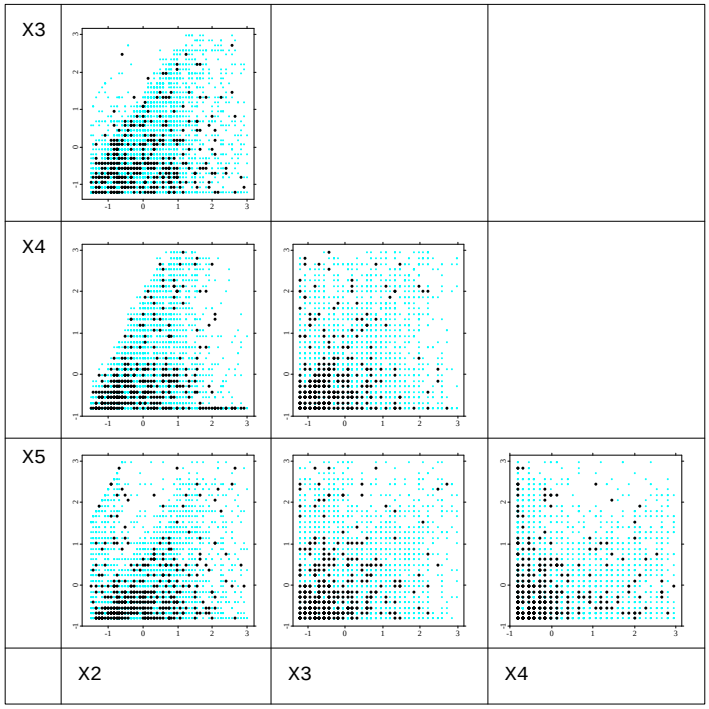


Figure 2. Kernel density estimates, variables X2 to X9, estimation data set B.



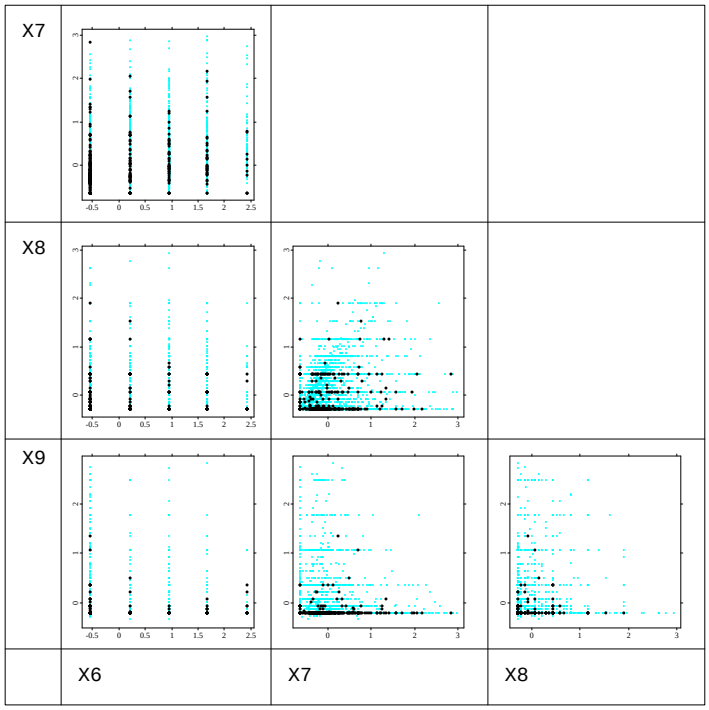


Figure 3. Scatterplots, variables X2 to X5 (upper plot) and X6 to X9 (lower plot), estimation data set B. Observations corresponding to Y=1 are emphasized in black.

In the special case of a binary response we have

$$E(Y|X) = P(Y = 1|X).$$

Variable	Coefficient	S.E.	t-value	Variable	Coefficient	S.E.	t-value
X0 (const.)	**-2.605280**	0.5890	-4.42	X19#2	-0.086954	0.3082	-0.28
X2	**0.246641**	0.1047	2.35	X19#3	0.272517	0.2506	1.09
X3	**-0.417068**	0.0817	-5.10	X19#4	-0.253440	0.4244	-0.60
X4	-0.062019	0.0849	-0.73	X19#5	0.178965	0.3461	0.52
X5	-0.038428	0.0816	-0.47	X19#6	-0.174914	0.3619	-0.48
X6	**0.187872**	0.0907	2.07	X19#7	0.462114	0.3419	1.35
X7	-0.137850	0.1567	-0.88	X19#8	**-1.674337**	0.6378	-2.63
X8	**-0.789690**	0.1800	-4.39	X19#9	0.259195	0.4478	0.58
X9	**-1.214998**	0.3977	-3.06	X19#10	-0.051598	0.2812	-0.18
X10#2	-0.259297	0.1402	-1.85	X20#2	-0.224498	0.3093	-0.73
X11#2	**-0.811723**	0.1277	-6.36	X20#3	-0.147150	0.2269	-0.65
X12#2	-0.272002	0.1606	-1.69	X20#4	0.049020	0.1481	0.33
X13#2	0.239844	0.1332	1.80	X21#2	0.132399	0.3518	0.38
X14#2	-0.336682	0.2334	-1.44	X21#3	**0.397020**	0.1879	2.11
X15#2	**0.389509**	0.1935	2.01	X22#2	-0.338244	0.3170	-1.07
X15#3	0.332026	0.2362	1.41	X22#3	-0.211537	0.2760	-0.77
X15#4	**0.721355**	0.2580	2.80	X22#4	-0.026275	0.3479	-0.08
X15#5	0.492159	0.3305	1.49	X22#5	-0.230338	0.3462	-0.67
X15#6	**0.785610**	0.2258	3.48	X22#6	-0.244894	0.4859	-0.50
X16#2	**0.494780**	0.2480	2.00	X22#7	-0.021972	0.2959	-0.07
X16#3	-0.004237	0.2463	-0.02	X22#8	-0.009831	0.2802	-0.04
X16#4	0.315296	0.3006	1.05	X22#9	0.380940	0.2497	1.53
X16#5	-0.017512	0.2461	-0.07	X22#10	-1.699287	1.0450	-1.63
X16#6	0.198915	0.2575	0.77	X22#11	0.075720	0.2767	0.27
X17#2	-0.144418	0.2125	-0.68	X23#2	-0.000030	0.1727	-0.00
X17#3	**-1.070450**	0.2684	-3.99	X23#3	-0.255106	0.1989	-1.28
X17#4	-0.393934	0.2358	-1.67	X24#2	0.390693	0.2527	1.55
X17#5	**0.921013**	0.3223	2.86				
X17#6	**-1.027829**	0.1424	-7.22				
X18#2	0.165786	0.2715	0.61				
X18#3	0.415539	0.2193	1.89				
X18#4	**0.788624**	0.2145	3.68				
X18#5	**0.565867**	0.1944	2.91	df			6118
X18#6	0.463575	0.2399	1.93	Log-Lik.			-1199.6278
X18#7	**0.568302**	0.2579	2.20	Deviance			2399.2556

Table 2. Results of the Logit Estimation. Estimation data set B. Bold coefficients are significant at 5%.

The semiparametric logit model that we consider here generalizes the linear argument $\beta^T X$ to a partial linear argument:

$$E(Y|X,T) = G\{\beta^T X + m(T)\}$$

This generalized partial linear model (GPLM) allows us to describe the influence of a part T of the explanatory variables in a nonparametric way. Here, $G(\bullet)$ is a known function, β is an unknown parameter vector, and $m(\bullet)$ is an unknown function. The

parametric component β and the nonparametric function $m(\bullet)$ can be estimated by the quasilikelihood method proposed in Severini & Staniswalis (1994).

We will use the GPLM estimator mainly as an exploratory tool in our practical credit scoring situation. Therefore we consider the GPLM for several of the metric variables separately as well as for combinations of them. As said before, we only consider variables X2 to X5 to be used within a nonparametric function because of the quasi–discrete structure of X6 to X9. For instance, when we include variable X5 in a nonlinear way, the parametric logit model is modified to

$$P(Y = 1|X) = F\left(m_5(X_5) + \sum_{j=2, j\neq 5}^{24} \beta_j^T X_j\right)$$

where a possible intercept is contained in the function $m_5(\bullet)$.

Table 3 contains only the parametric coefficients for the parametric and semiparametric estimates for variables X2 to X9. The column headed by "Logit" repeats the parametric logit estimates for the for model with variables X2 to X24. The rest of the columns correspond to the semiparametric estimates where we fitted those variables nonparametrically which are heading the columns.

		Nonparametric in					
Variable	Logit	X2	X3	X4	X5	X4,X5	X2,X4,X5
constant	**-2.605**	–	–	–	–	–	–
X2	**0.247**	–	**0.243**	**0.241**	**0.243**	**0.228**	–
X3	**-0.417**	**-0.414**	–	**-0.414**	**-0.416**	**-0.408**	**-0.399**
X4	-0.062	-0.052	-0.063	–	-0.065	–	–
X5	-0.038	-0.051	-0.045	-0.034	–	–	–
X6	**0.188**	**0.223**	**0.193**	**0.190**	**0.177**	0.176	**0.188**
X7	-0.138	-0.138	-0.142	-0.131	-0.146	-0.135	-0.128
X8	**-0.790**	**-0.777**	**-0.800**	**-0.786**	**-0.796**	**-0.792**	**-0.796**
X9	**-1.215**	**-1.228**	**-1.213**	**-1.222**	**-1.216**	**-1.214**	**-1.215**

Table 3. Parametric coefficients in parametric and semiparametric logit, variables X2 to X9. Estimation data set B. Bold values are significant at 5%.

It turns out, that all coefficients vary little over the different estimates. This holds as well for their significance (determined by a t–test). Variables X4 and X5 are constantly insignificant over all estimates. Hence, they are interesting candidates for a nonparametric modeling: variables which are significant may already capture a lot of information on Y by the parametric inclusion into the model.

The semiparametric logit model is estimated by semiparametric maximum-likelihood, a combination of ordinary and smoothed maximum-likelihood. The fitted curves for the nonparametric components according to Table 3 can be found in Figures 4 for the marginal fits (variables X2 to X5 separately as the nonparametrical component) and Figure 6 for the bivariate surface (variables X4 and X5 jointly nonparametrically included). Additionally,

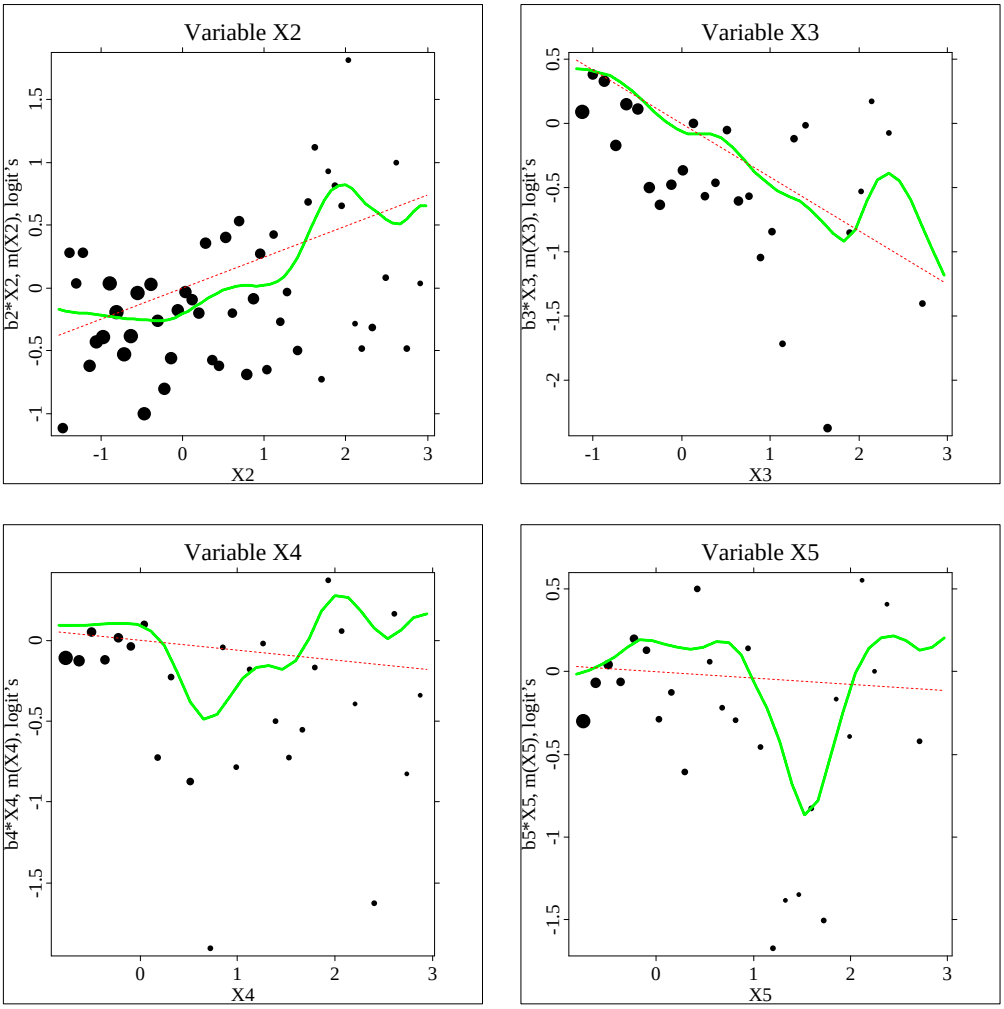


Figure 4. Marginal dependencies, variables X2 to X5. Thicker bullets correspond to more observations in a class. Parametric logit fits (thin dashed linear functions) and GPLM logit fits (thick solid curves).

Figures 4 and 5 reflect the actual dependence of the response Y on variables X2 to X9. We have plotted each variable restricted to [-3,3] (i.e. the data from sample B) versus the logits

$$\text{logit} = \log\left(\frac{\widehat{p}}{1-\widehat{p}}\right)$$

where $\widehat{p}$ are the relative frequencies for $Y = 1$. Essentially, these logits are obtained from classes of identical realizations. In case that $\widehat{p}$ was 0 or 1, several realizations have been summarized into one class. For all variables but X7 this only concerns single values.

The plots of the marginal dependencies for variables X6 to X9 show that the realizations essentially concentrate in one value. Hence we did not fit a nonparametric function here.

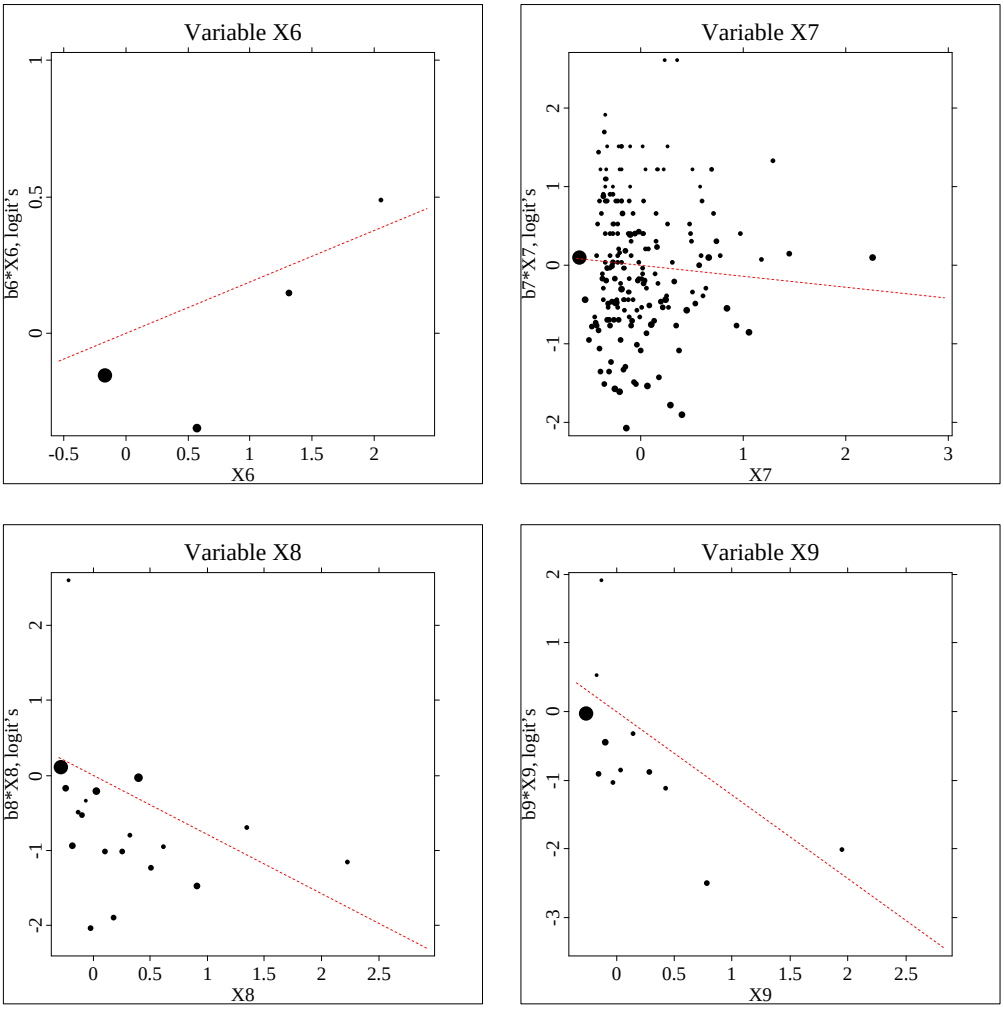


Figure 5. Marginal dependencies, variables X6 to X9. Thicker bullets correspond to more observations in a class. Parametric logit fits (thin dashed). Estimation data set B.

5 Testing the Semiparametric Model

To assess, whether the semiparametric fit outperforms the parametric logit or not, we have a number of statistical characteristics. For the above estimated models, they are summarized in Table 4.

The deviance is minus twice the estimated log–likelihood of the fitted model in our case. For the parametric case, the degrees of freedom just denote

$$df = n - k$$

where n is the sample size and k the number of estimated parameters. In the semiparametric case, a corresponding number of degrees of freedom can be approximated. Deviance and (approximate) degrees of freedom of the parametric and the semiparametric model can be used to construct a likelihood ratio test to compare both models (see Buja, Hastie

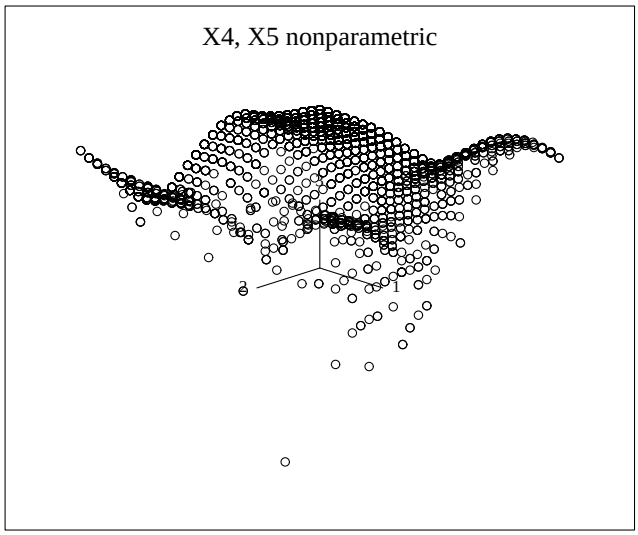


Figure 6. Bivariate nonparametric surface for variables X4, X5. Estimation data set B.

& Tibshirani, 1990; Müller, 1997). The obtained significance levels from these tests are denoted by α. Finally, we listed the pseudo R^2 values, an analog to the linear regression coefficient of determination.

It is obvious to see that models containing variable X5 in the nonparametric part considerably decrease the deviance and increase the coefficient of determination R^2. Accordingly, the significance level for the test of parametric versus nonparametric modeling decreases. In particular, it is below 5% for the both models including X5 alone and including X4, X5 jointly in a nonparametric way.

		Nonparametric in					
	Logit	X2	X3	X4	X5	X4,X5	X2,X4,X5
Deviance	2399.26	2393.16	2395.06	2391.17	2386.97	2381.49	2381.96
df	6118.00	6113.79	6113.45	6113.42	6113.36	6108.56	6107.17
α	–	0.212	0.459	0.130	**0.024**	**0.046**	0.094
pseudo R^2	14.68%	14.89%	14.82%	14.96%	15.11%	15.31%	15.29%

Table 4. Statistical characteristics in parametric and semiparametric logit fits. Estimation data set B. Bold values are significant at 5%.

6 Misclassification and Performance Curves

The different fits can be compared by looking at misclassification rates. For the validation, the provided data comprise a subsample (data set C) which was not included in the estimation. We use this validation sample to evaluate all estimators.

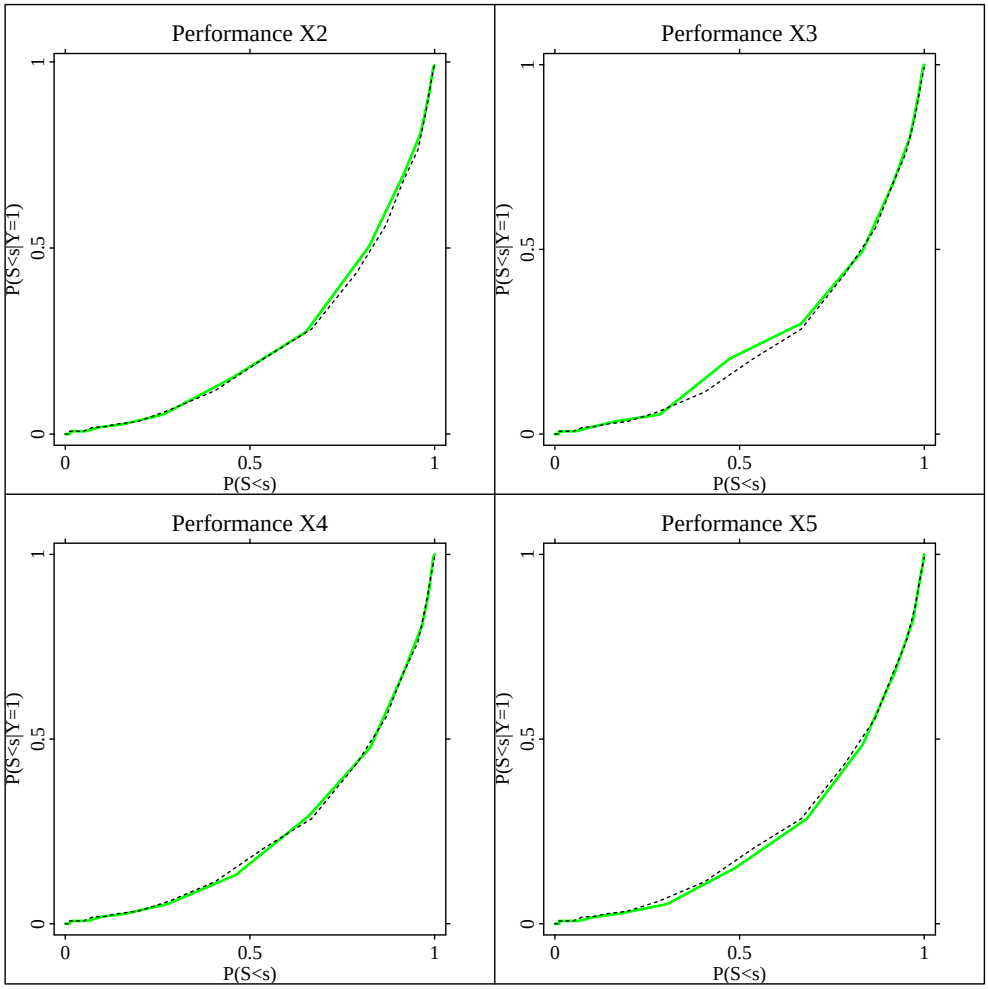


Figure 7. Performance curves, parametric logit (black dashed) and semiparametric logit models (thick grey), with variables X2 to X5 (separately) included nonparametrically. Validation data set C.

The misclassification rates can be pictured by performance curves (Lorenz curves). The performance curve is defined by plotting the probability of observations classified as "good"

$$P(S < s)$$

versus the conditional relative frequency of observations classified as "good" conditioned on "bad"

$$P(S < s|Y = 1).$$

Here, S denotes the score which equals in the parametric logit model

$$S = \sum_{j=2}^{24} \beta_j^T X_j + \beta_0$$

and in the semiparametric logit model

$$S = m_5(X_5) + \sum_{j=2, j\neq 5}^{24} \beta_j X_j$$

when fitting X5 nonparametrically, for instance.

The probability value $P(S < s|Y = 1)$ is a measure for misclassification and thus to be minimized. Hence, one performance curve is to be preferred to another, when it is more downwards shaped.

In practice, the probability $P(S < s)$ is replaced by the relative frequency of classifications $Y = 0$ ("good") given a threshold s. The analog is done for $P(S < s|Y = 1)$. We have computed performance curves for both the estimation data set B and the validation data set C.

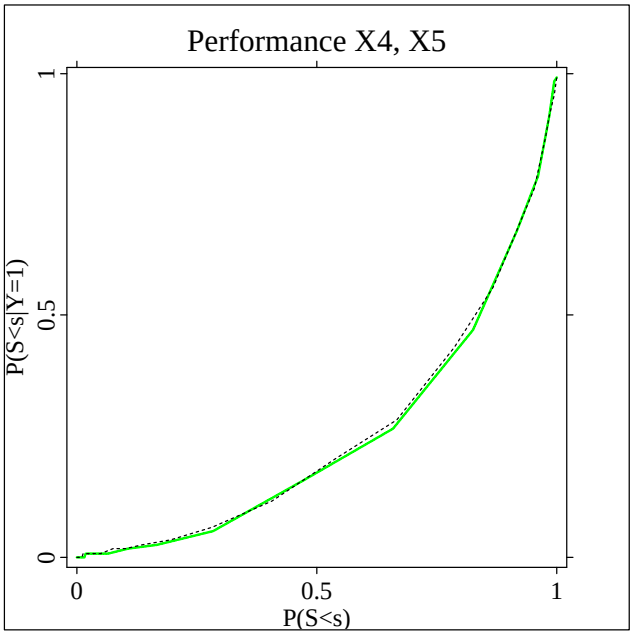


Figure 8. Performance curves, parametric logit (black dashed) and semiparametric logit model (thick grey), with variables X4, X5 (jointly) included nonparametrically. Validation data set C.

Figure 7 compares the performance of the parametric logit fit and the semiparametric logit fit obtained by separately including X2 to X5 nonparametrically. Indeed, the semiparametric model for the influence of X5 improves the performance with respect to the parametric model. The semiparametric models for the influence of X2 to X4 do not improve the performance with respect to the parametric model, though.

Figure 8 compares the performance of the parametric logit fit and the semiparametric logit fit obtained by jointly including X4, X5 nonparametrically. This performance curve improves versus nonparametrically fitting only X4, but shows less power versus fitting

only X5. Hence, the improvement of using both variables jointly may be explained by the influence of X5 only.

7 References

Arminger, G.; Enache, D.; Bonne, T. (1997), Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks. *Computational Statistics* **12**, Sonderband "10 Jahre AG GLM", 293–310.

Buja, A.; Hastie, T.; Tibshirani, R. (1989), Linear smoothers and additive models (with discussion), *Annals of Statistics* **17**, 453–555.

Gourieroux, C. (1994), Credit Scoring, Script.

Hand, D. J.; Henley, W. E. (1997), Statistical Classification Methods in Consumer Credit Scoring: a Review. *J. R. Statist. Soc. A* **160**, 523–541.

Härdle, W.; Mammen, E.; Müller, M. (1998), Testing Parametric versus Semiparametric Modelling in Generalized Linear Models. *Journal of the American Statistical Association* **93**, 1461–1474.

Henley, W. E.; Hand, D. J. (1996), A k-nearest-neighbor classifier for assessing consumer credit risk. *Statistician* **45**, 77–95.

McCullagh, P.; Nelder, J. A. (1989), *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, 2 edn, Chapman and Hall, London.

Müller, M.; Rönz, B.; Härdle, W. (1997), Computerassisted Semiparametric Generalized Linear Models. *Computational Statistics* **12**, Sonderband "10 Jahre AG GLM", 153–172.

Müller M. (1997), Computer–assisted Generalized Partial Linear Models, in: *Proceedings of Interface'97* **29/1**, Houston, Texas, 221–230.

Severini, T. A.; Staniswalis, J. G. (1994), Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**, 501–511.