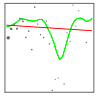
Semiparametric Credit Scoring

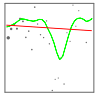
Marlene MÜLLER
Bernd RÖNZ
Wolfgang HÄRDLE

Center for Applied Statistics and Economics (CASE)



Plan

- Basel II & Probabilities of Default (PDs)
- Problem and Data Description
- Logistic Credit Scoring
- Semiparametric Credit Scoring
- Testing the Semiparametric Model
- Misclassification and Performance Curves



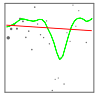
Credit Rating/Scoring

- new interest in this field because of Basel II:
  capital requirements of a bank are adapted to the individual credit portfolio
- possibilities for banks:
  - ⋆ ratings and PDs from external rating agencies
  - ⋆ internal ratings-based approach (IRB approach)
    $\longrightarrow$ better adaptation to bank-specific portfolio
- one of the key problems:
  estimation of PDs

Reference: The New Basel Capital Accord ("Basel II"), Bank for International Settlements

Probabilities of Default (PDs)

- basis: 1 year
- "grounded on historical experience, but forward looking"
- from 2007: historical period for observations of at least 5 years
- sufficient sample size
- validation: calibration, discriminatory power



From PDs ...

one determines

- ratings

  $$\text{AAA, AA+, AA, ..., BB, ..., D}$$

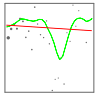
  corresponding e.g. to PDs

  $$0.01\%, 0.02\%, 0.03\%, ..., 1.17\%, ..., 100\%$$

- expected loss

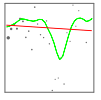
  $$\text{EL} = \text{PD} \cdot \text{EAD} \cdot \text{LGD}$$

  EAD = exposure at default, LGD = loss given default



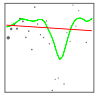
Methods for Estimating PDs

- discriminance analysis, classification
  $\rightarrow$ Scores
- categorical regression (logit/probit, panel, ordered categories)
  $\rightarrow$ Scores + PDs
- Merton approach (stock price as estimate for the market value)
  $\rightarrow$ PD by "distance to default"
- Jarrow, Lando, Turnbull (transition probabilites between different states = rating classes)



Specific Problems for Credit Data

- relatively small default frequencies
- sample selection: not all clients are creditworthy
- consumers
  - ⋆ in general no credit history
  - ⋆ many categorical covariates
- corporates
  - ⋆ not yet long enough data histories (Basel II: 5 years)
  - ⋆ in general not independent observations
- dependence on macroeconomic factors

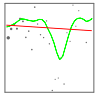


Categorical Regression

model and prediction for

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

- logit model (logistic discriminance analysis)
- modifications
  - probit model (different link function)
  - panel models
  - ordered responses
  - sample selection (Heckman estimator)

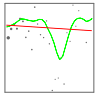


Data Examples

Sample for Cars

		Yes	No	(in %)	
Y	**Default**	26.4	73.6		
	previous loans OK	66.2	33.8		
	employed	73.2	26.8		
		Min	Max	Mean	S.E.
	duration (in months)	4	54	21.8	10.6
	amount (in DM)	428	14179	3902.3	2621.9
	age (in years)	19	75	34.2	10.8

References: Fahrmeir & Hamerle (1984), Fahrmeir & Tutz (1995)

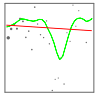


French Credit Data

- Response variable Y
  (credit status, 0="Non-Default", 1="Default")
- Metric variables X2 to X9.
- Categorical variables X10 to X24.

	Estimation data set	Validation data set
0 ("Non-Defaults")	5808 (94%)	1891 (94.6%)
1 ("Defaults")	372 (6%)	107 (5.4%)
total	6180	1998

Table 1: Responses.



Density Plots

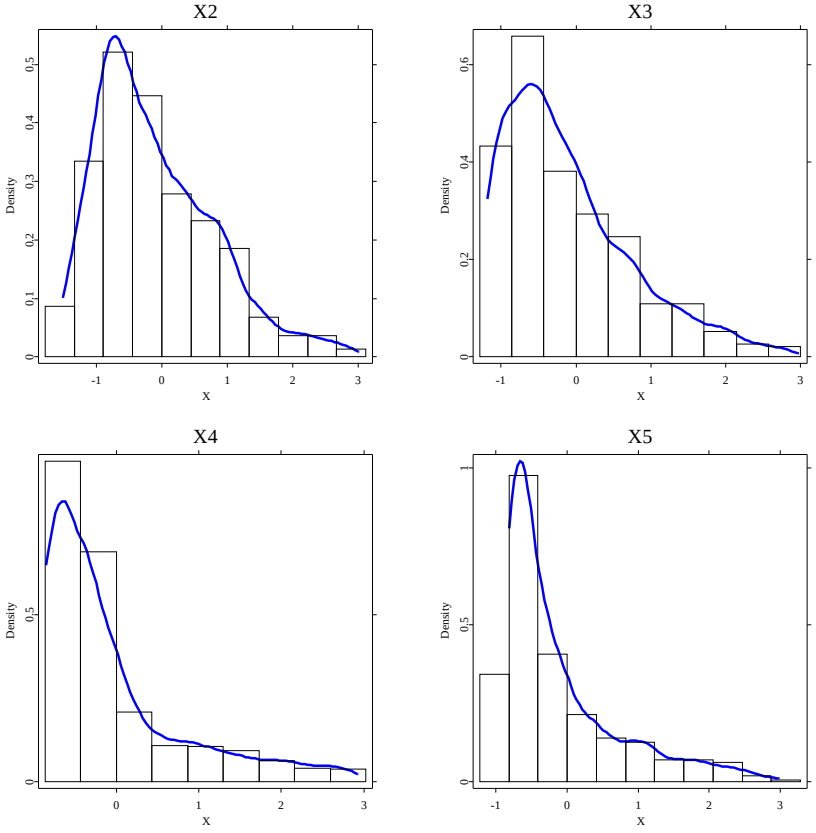
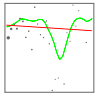


Figure 1: Kernel density estimates, variables X2 to X5.



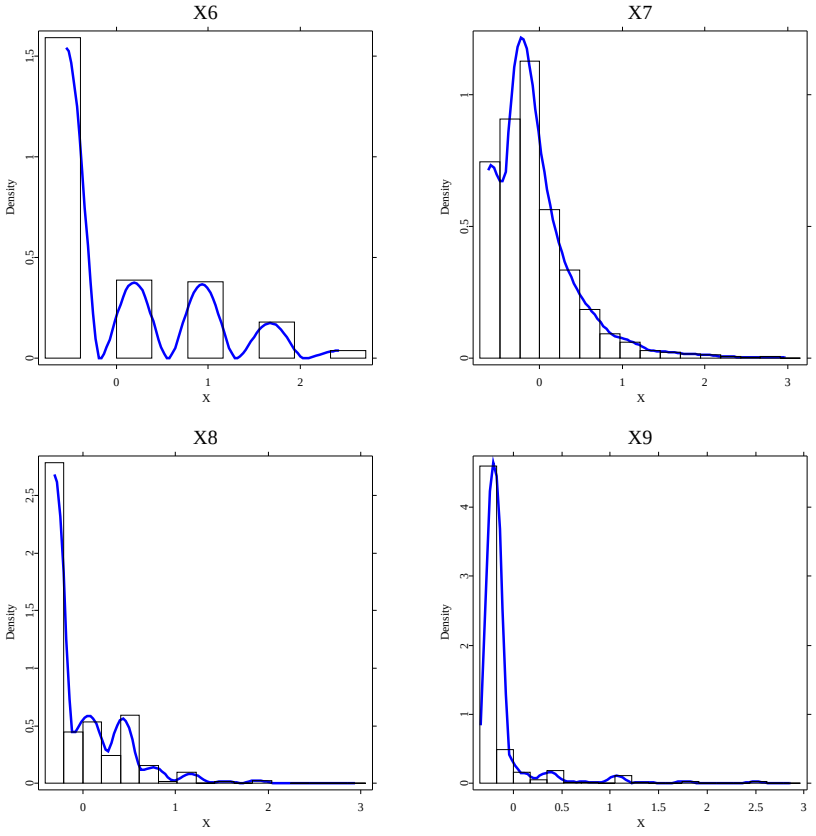
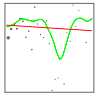


Figure 2: Kernel density estimates, variables X2 to X5.



Scatter Plots

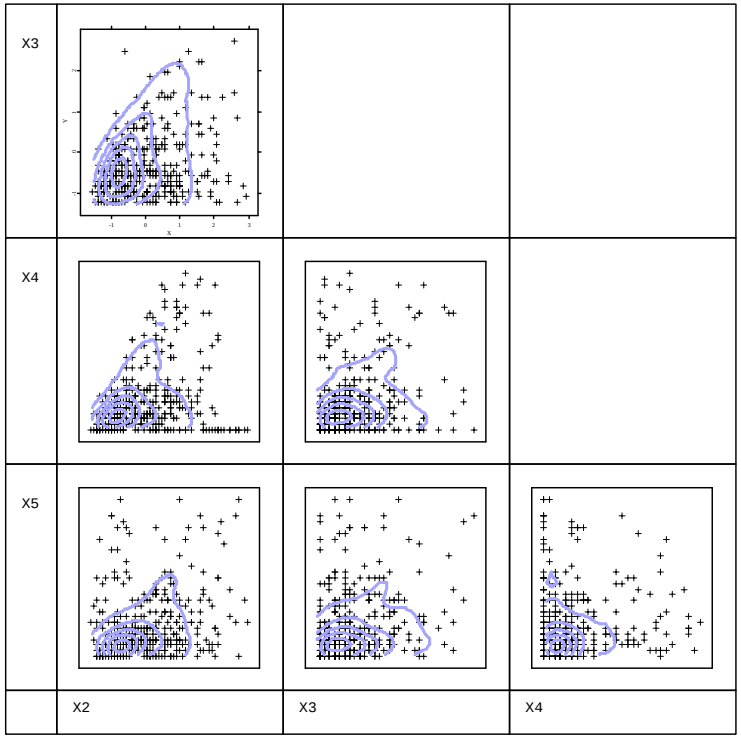
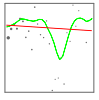


Figure 3: Scatter-contour-plots, variables X2 to X5. Observations corresponding to Y=1 are emphasized in black.



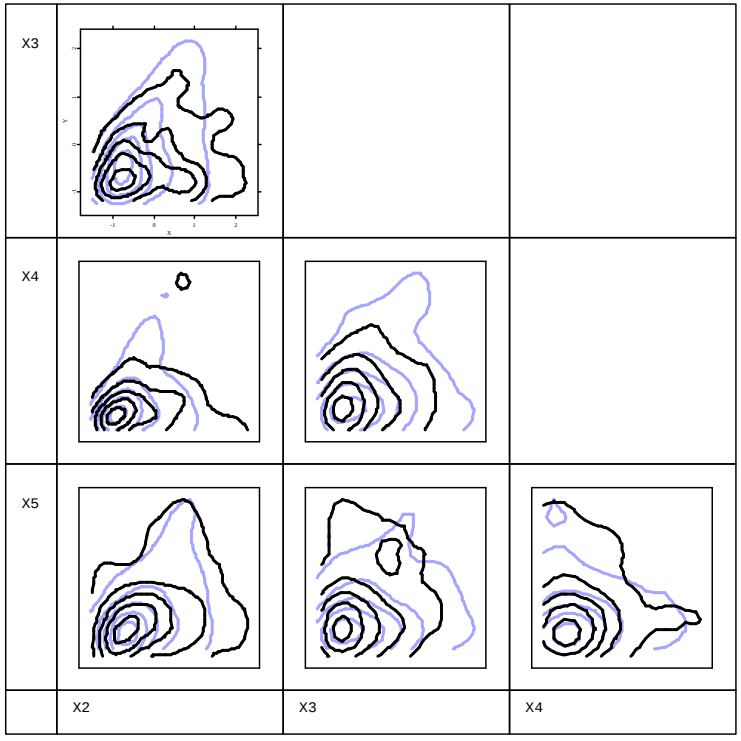


Figure 4: Contour-contour-plots, variables X2 to X5. Observations corresponding to Y=1 are emphasized in black.

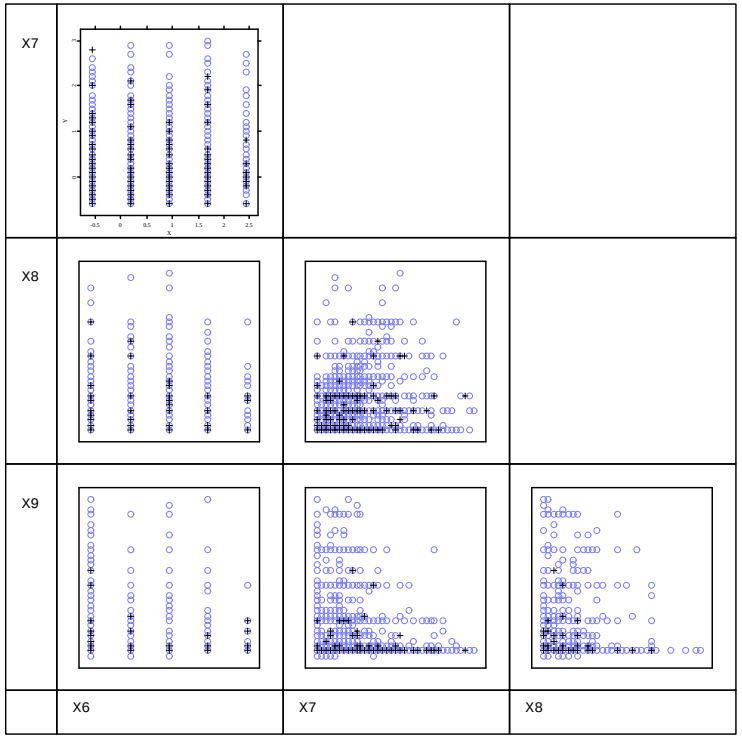
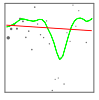


Figure 5: Scatterplots, variables X6 to X9. Observations corresponding to Y=1 are emphasized in black.



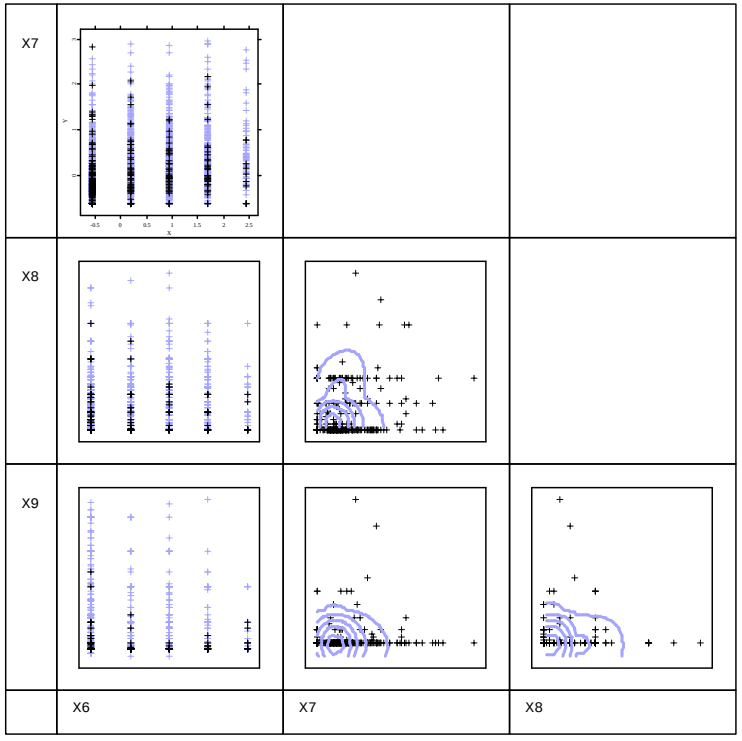


Figure 6: Scatter-contour-plots, variables X6 to X9. Observations corresponding to Y=1 are emphasized in black.

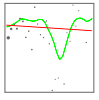
Logistic Credit Scoring

logit model (logistic discriminant analysis)

$$P(Y=1|X) = F\left(\sum_{j=2}^{24} \beta_j^\top X_j + \beta_0\right), \qquad F(\bullet) = \frac{1}{1+e^{-\bullet}} \text{ logistic cdf}$$

X_j denotes here

- j-th variable if Xj is metric ($j \in \{2, \ldots, 9\}$)
- vector of dummies if Xj is categorical ($j \in \{10, \ldots, 24\}$)



Logit Model

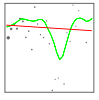
- binary response

$$Y = \begin{cases} 1 & \text{if } Y^* = v(X) - u > 0 \\ 0 & \text{otherwise} \end{cases}$$

- $Y^* =$ latent variable, (negative) credit score
- $v(\bullet) =$ index function that relates X to Y^*, e.g.

$$EY^* = v(X) = \sum_{j=2}^{24} \beta_j^\top X_j + \beta_0$$

- $u \sim F$ unobserved error term.



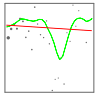
Variable	Coefficient	S.E.	t-value	Variable	Coefficient	S.E.	t-value
X0 (const.)	**-2.605280**	0.5890	-4.42	X19#2	-0.086954	0.3082	-0.28
X2	**0.246641**	0.1047	2.35	X19#3	0.272517	0.2506	1.09
X3	**-0.417068**	0.0817	-5.10	X19#4	-0.253440	0.4244	-0.60
X4	-0.062019	0.0849	-0.73	X19#5	0.178965	0.3461	0.52
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

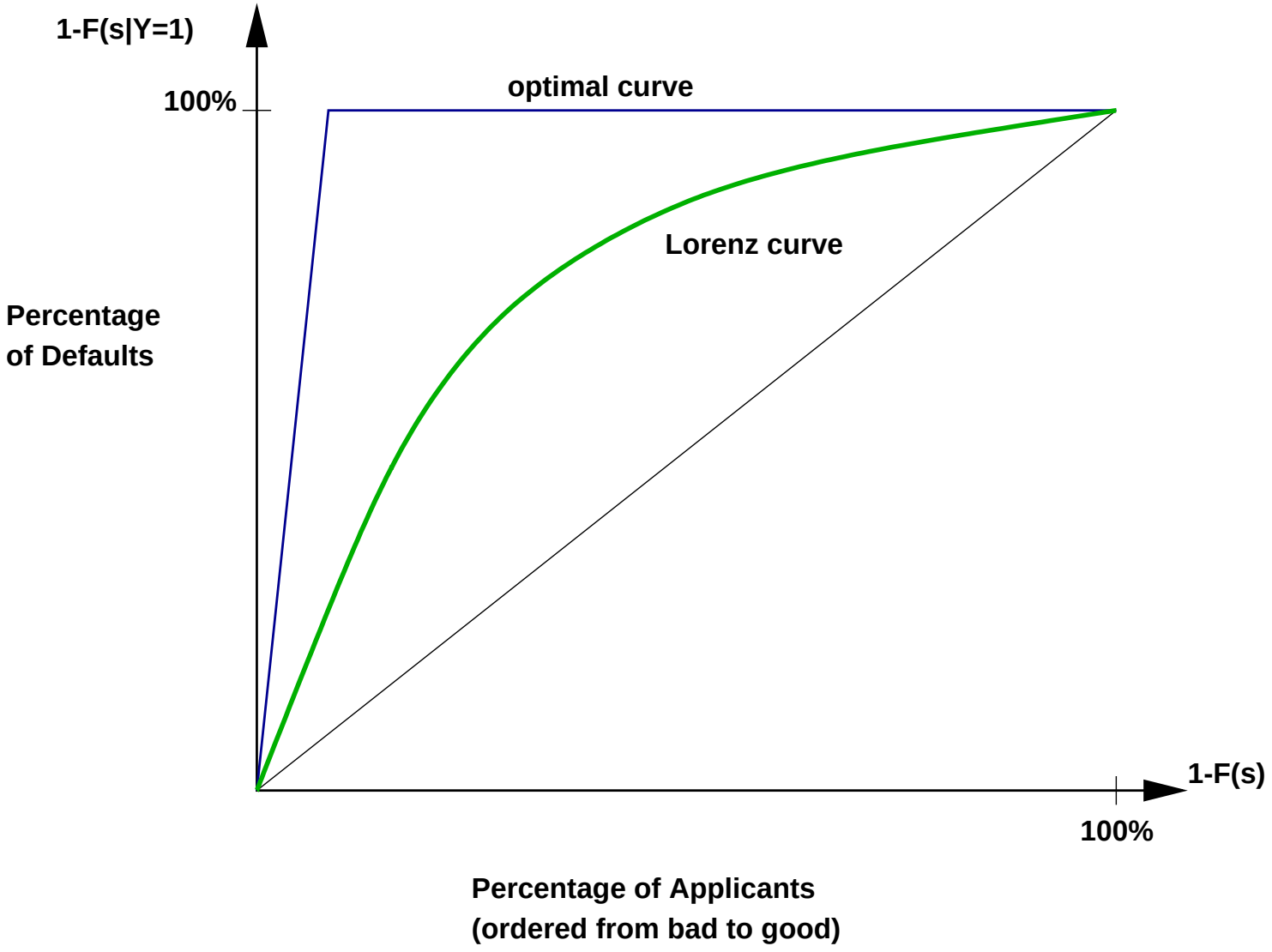
Performance (Lorenz curves)

- calculate scores, e.g.

$$S = X_5 \quad \text{oder} \quad S = \sum_{j=2}^{24} \beta_j^\top X_j + \beta_0$$

- plot
  - $\star$ $1 - F(s) = P(S > s)$ (classified as "Default") vs.
  - $\star$ $1 - F_1(s) = P(S > s|Y = 1)$
    (classified as "Default" and true observation is indeed "Default")




1-F(s|Y=1)
100%
optimal curve
Lorenz curve
Percentage
of Defaults
1-F(s)
100%
Percentage of Applicants
(ordered from bad to good)

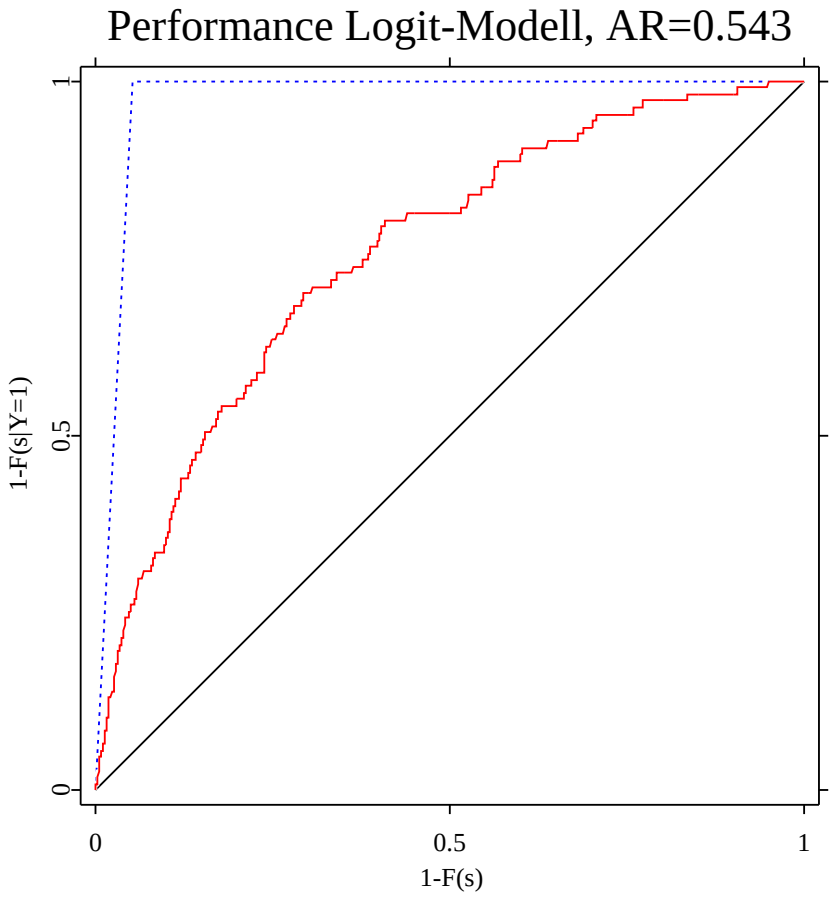
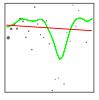


Figure 7: Performance curve for logit model (red) and optimal curve (blue).



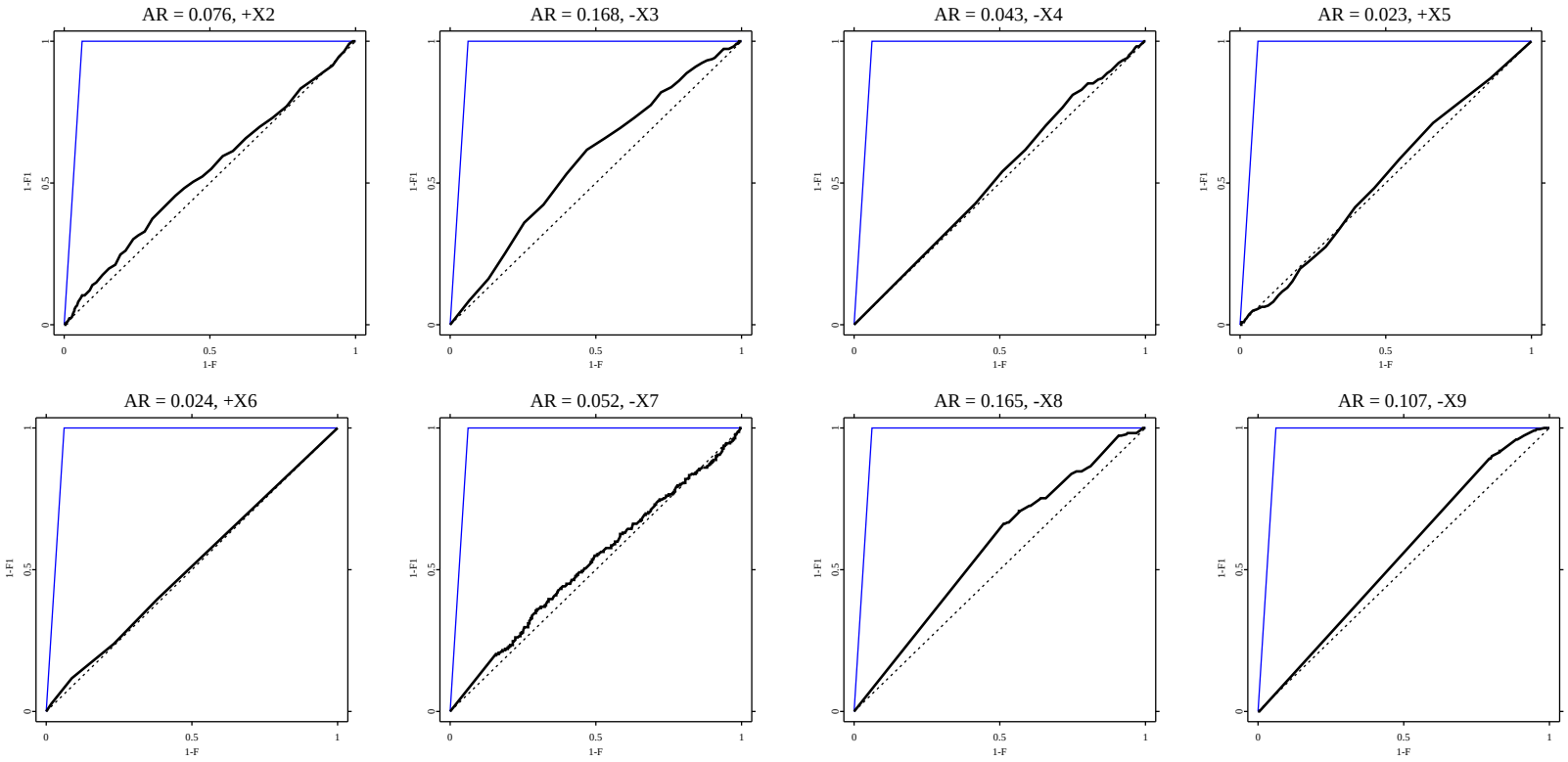
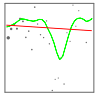


Figure 8: Lorenz curves for variables X2 to X9.



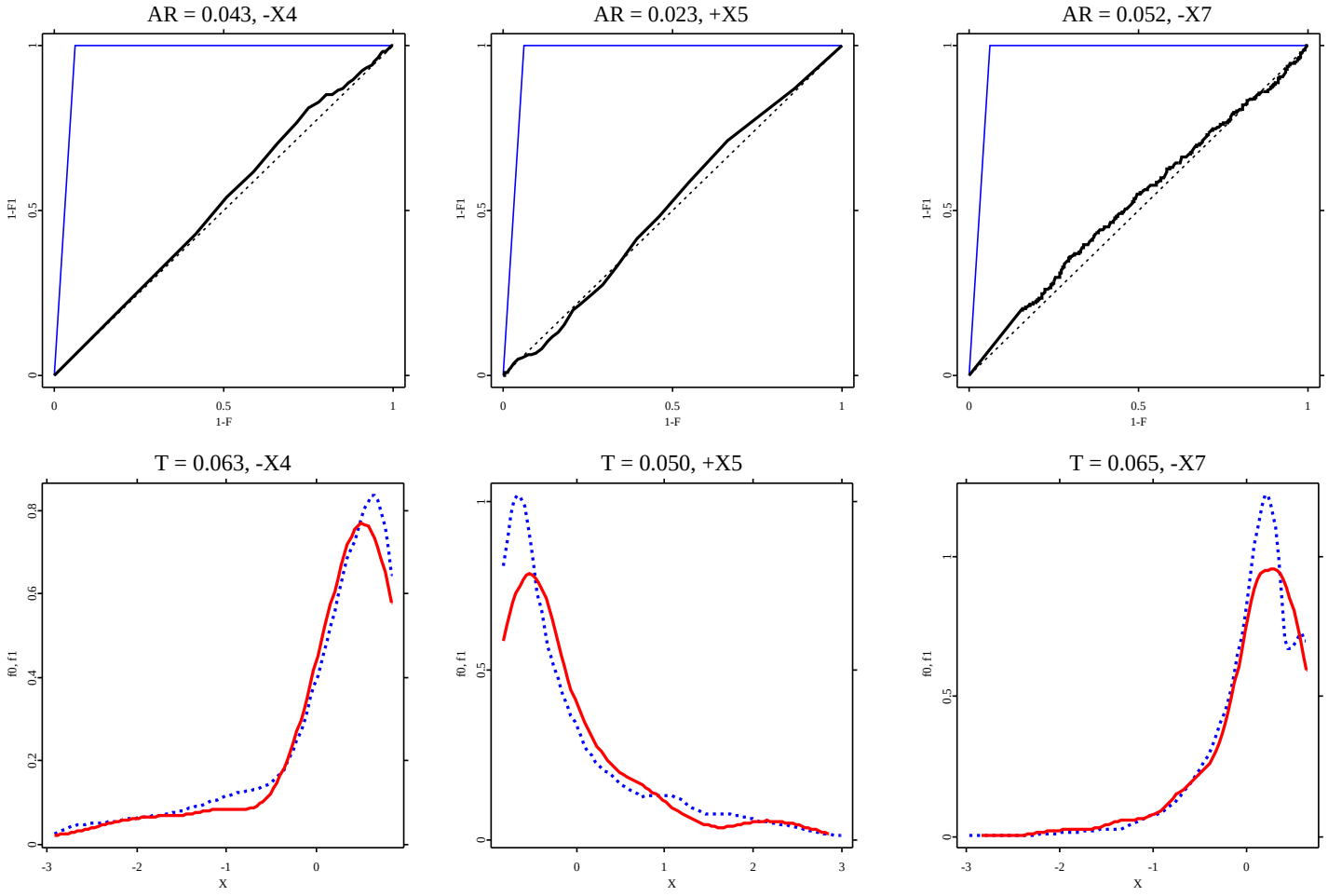
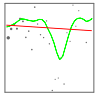


Figure 9: Lorenz curves, density estimates (cond. on Y) for X4, X5, X7.



How does Y (more exactly: $\log(\frac{p}{1-p})$) depend on the the variables?

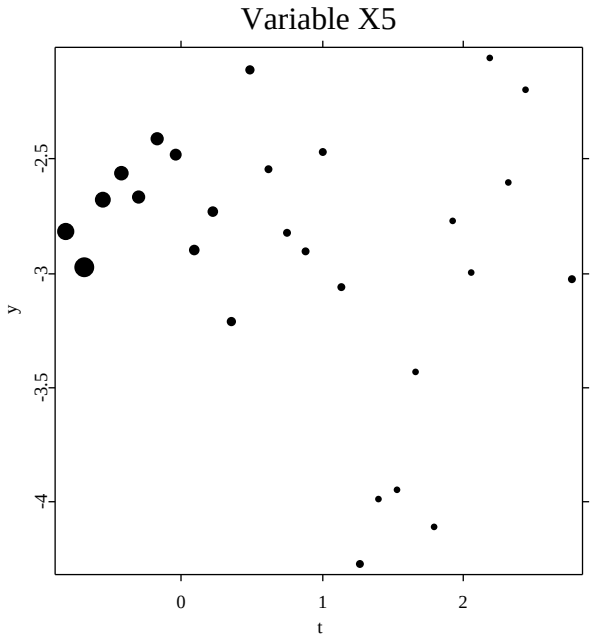
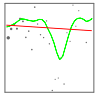


Figure 10: Marginal dependence, variable X5. Thicker bullets correspond to more observations.



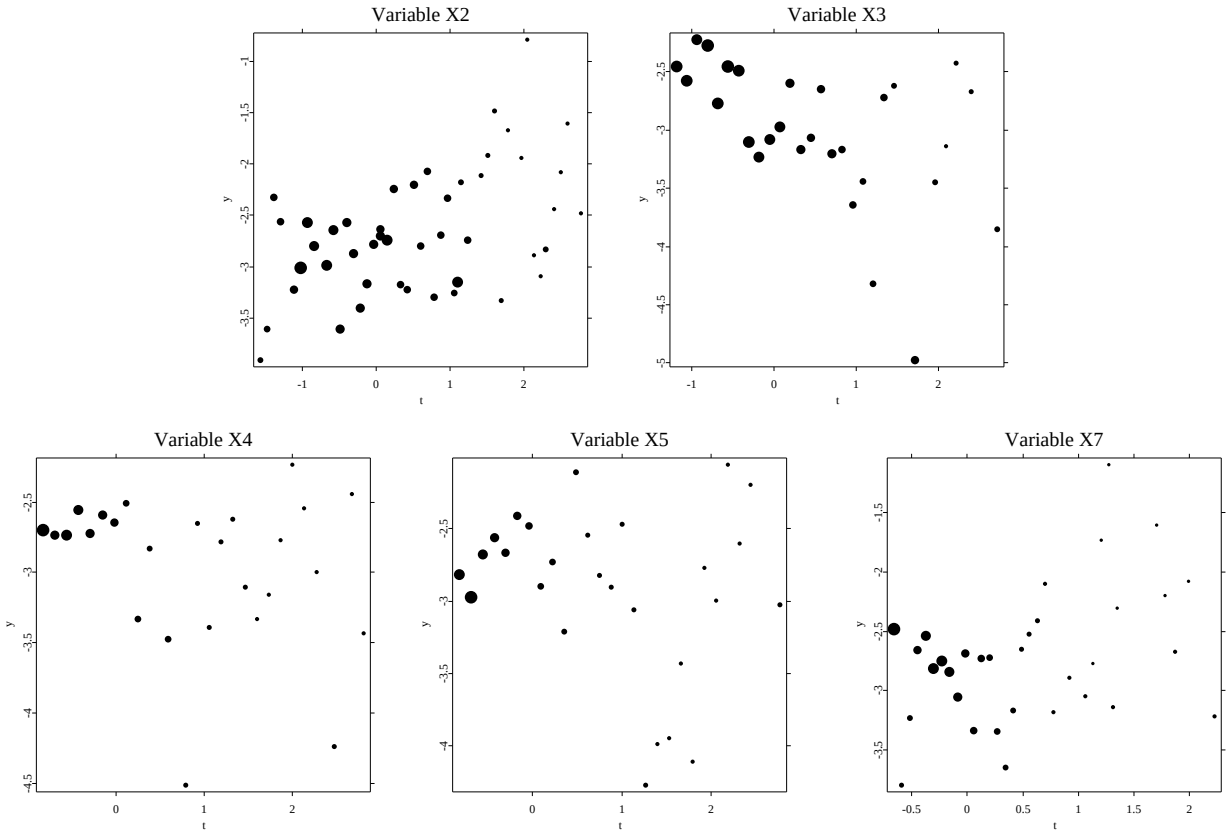
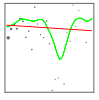


Figure 11: Marginal dependencies, variables X2 to X5, X7.



Semiparametric Credit Scoring

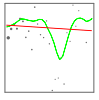
Modelling Strategy

piecewise linear logit model (GPLM)

$$P(Y = 1|X, T) = F\{\beta^\top X + pl(T)\}$$

where

- $F(\bullet)$ known (link) function, here: logistic cdf
- $pl(\bullet)$ piecewise linear function
- β unknown parameter vector



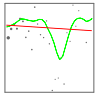
Strategy in this Study

partial linear logit model (GPLM)

$$P(Y = 1|X, T) = F\{\beta^\top X + m(T)\}$$

where

- $F(\bullet)$ known (link) function, here: logistic cdf
- $m(\bullet)$ unknown smooth function
- β unknown parameter vector



Maximum–Likelihood for the GLM

$$E(Y|X) = \mu = G\{X^\top \beta\},\ Var(Y|X) = \sigma^2 V(\mu)$$

- maximization of the (log-)likelihood

$$\ell(Y,\mu) = \sum_i \ell_i(Y_i,\mu_i)$$

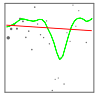
$$=_{\text{here}} \sum_i Y_i \log(\mu_i) + (1-Y_i)\log(1-\mu_i)$$

  or minimzation of the deviance

$$\text{Dev}(Y,\mu) = 2\max_{\widetilde{\mu}} \ell(Y,\widetilde{\mu}) - 2\ell(Y,\mu)$$

- algorithm: "iteratively reweighted least squares"

Reference: McCullagh & Nelder (1989)

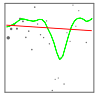


Semiparametric Maximum–Likelihood

GPLM ("generalized partial linear model")

$$E(Y|X,T) = \mu = G\{X^\top \beta + m(T)\},\ Var(Y|X,T) = \sigma^2 V(\mu)$$

- combining parmetric and smoothed (log-)likelihood
- for β:
  "iteratively reweighted least squares" + modified design matrix
- for $m(\bullet)$:
  Nadaraya–Watson (or other) smoother



Estimation of the GPLM

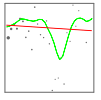
$$E(Y|X,T) = G\left(X^\top \beta + m(T)\right)$$

- $\widehat{\beta}$ can be estimated if m known
  (parametric method, weighted LSE),
- $\widehat{m}$ can be estimated if β known
  (nonparametric method, e.g. Nadaraya–Watson type)

estimation method

- profile likelihood
  generalized Speckman estimator
- backfitting, modified Backfitting

References:
Severini & Staniswalis (1994), Speckman (1988), Hastie & Tibshirani (1990), Müller (2001)



Algorithm (generalized Speckman estimator)

- *parametric part*

$$\beta^{new} = (\widetilde{\mathcal{X}}^\top \mathcal{W} \widetilde{\mathcal{X}})^{-1} \widetilde{\mathcal{X}}^\top \mathcal{W} \widetilde{z},$$

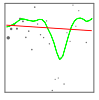
- *nonparametric part*

$$m^{new} = \mathcal{S}(z - \mathcal{X}\beta)$$

where

$$\begin{aligned} \mathcal{S} &= \textit{smoother matrix}, \\ \widetilde{\mathcal{X}} &= (\mathcal{I} - \mathcal{S})\mathcal{X}, \\ \widetilde{z} &= (\mathcal{I} - \mathcal{S})z = \widetilde{\mathcal{X}}\beta - \mathcal{W}^{-1}v. \end{aligned}$$

$\mathcal{X}$ design, $\mathcal{I}$ identity, $v = (\ell_i')$, $\mathcal{W} = \text{diag}(\ell_i'')$



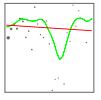
Estimation Matrix

the update of $\mathcal{X}\beta + m$ can be expressed by a linear estimation matrix:

$$\mathcal{X}\beta^{new} + m^{new} = \mathcal{R}z$$

where

$$\mathcal{R} = \widetilde{\mathcal{X}}\{\widetilde{\mathcal{X}}^\top \mathcal{W}\widetilde{\mathcal{X}}\}^{-1}\widetilde{\mathcal{X}}^\top \mathcal{W}(\mathcal{I} - \mathcal{S}) + \mathcal{S}.$$



LR Test

if (at convergence)

$$\widehat{\eta} = \mathcal{R}z = \mathcal{R}(\widehat{\eta} - \mathcal{W}^{-1}v), \quad \eta = \mathcal{X}\beta + m$$

$$\Rightarrow \quad \text{Dev}\,(y, \widehat{\mu}) \approx (z - \widehat{\eta})^{\top}\mathcal{W}^{-1}(z - \widehat{\eta})$$

approximative degrees of freedom

$$df^{err}(\widehat{\mu}) = n - \text{tr}\left(2\mathcal{R} - \mathcal{R}^{\top}\mathcal{W}\mathcal{R}\mathcal{W}^{-1}\right)$$

or $\quad df^{err}(\widehat{\mu}) = n - \text{tr}\,(\mathcal{R})$

Reference: Hastie & Tibshirani (1990)

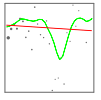
Application

- to include variable X5 in a nonlinear way:

$$P(Y = 1|X_{-5}, X_5) = F\left(\sum_{j\neq 5} \beta_j^\top X_j + m_5(X_5)\right)$$

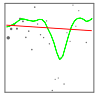
- to include variable X5 in a nonlinear way:

$$P(Y = 1|X_{-4,-5}, (X_4, X_5)) = F\left(\sum_{j\neq 4,5} \beta_j^\top X_j + m_{45}(X_4, X_5)\right)$$



		nonparametric in						
	Logit	X2	X3	X4	X5	X7	X4,X5	X2, X4,X5
constant	**-2.605**	–	–	–	–	–	–	–
X2	**0.247**	–	**0.243**	**0.241**	**0.243**	**0.233**	**0.228**	–
X3	**-0.417**	**-0.414**	–	**-0.414**	**-0.416**	**-0.417**	**-0.408**	**-0.399**
X4	-0.062	-0.052	-0.063	–	-0.065	-0.054	–	–
X5	-0.038	-0.051	-0.045	-0.034	–	-0.042	–	–
X6	**0.188**	**0.223**	**0.193**	**0.190**	**0.177**	**0.187**	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**	**-0.793**	**-0.792**	**-0.796**
X9	**-1.215**	**-1.228**	**-1.213**	**-1.222**	**-1.216**	**-1.227**	**-1.214**	**-1.215**

Table 2: Parametric coefficients for variables X2 to X9. Bold values are significant at 5%.



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

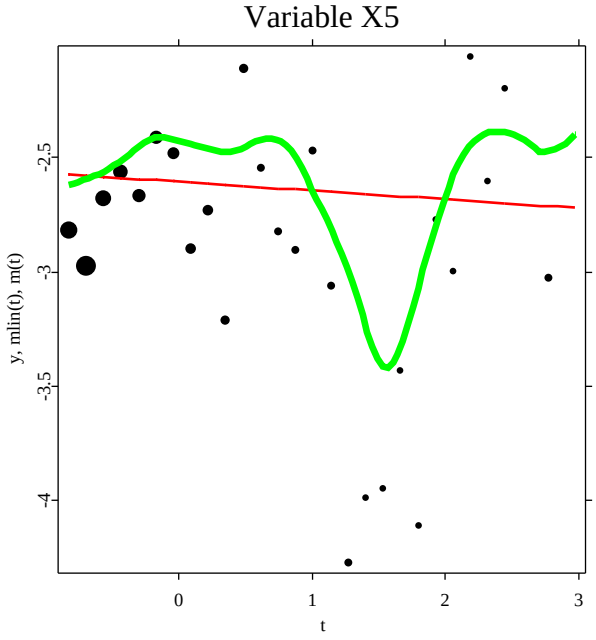
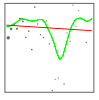


Figure 12: Marginal dependence, variable X5. Thicker bullets correspond to more observations. Parametric (red) and GPLM logit fit (green).



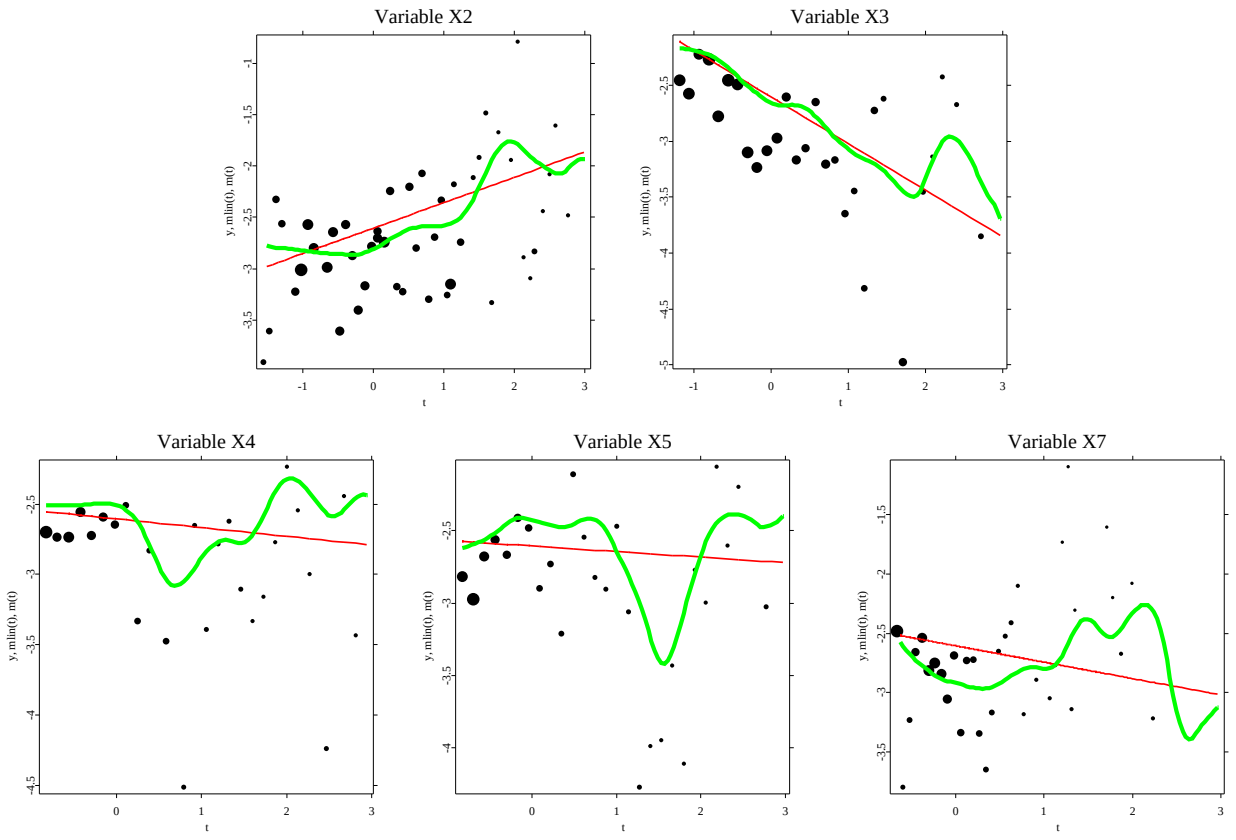
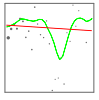


Figure 13: Marginal dependencies, variables X2 to X5. Parametric logit fits (red) and GPLM logit fits (green).



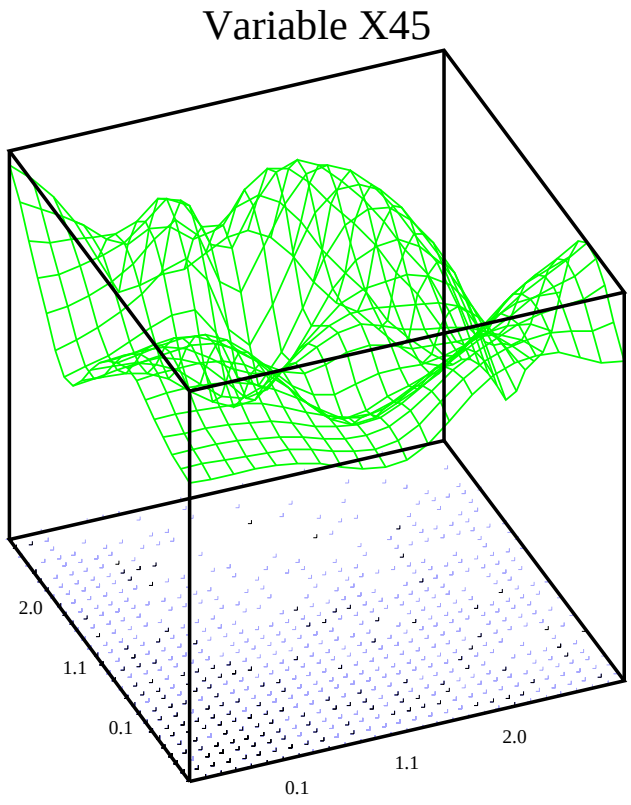
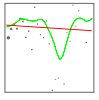


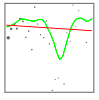
Figure 14: Bivariate nonparametric surface for variables X4, X5.



Testing the Semiparametric Model

		nonparametric in						
	Logit	X2	X3	X4	X5	X7	X4,X5	X2, X4,X5
deviance	2399.26	2393.03	2395.19	2391.29	**2387.17**	**2388.63**	**2372.63**	**2372.43**
df	6118.00	6113.72	6113.57	6113.34	6113.41	6113.61	6103.82	6100.23
α	–	0.210	0.458	0.133	**0.026**	**0.041**	**0.023**	**0.077**
AIC	2523.3	2525.6	2528.0	2524.6	2520.4	2521.4	2525.0	2533.0
pseudo R^2	14.7%	14.9%	14.8%	15.0%	15.1%	15.1%	15.6%	15.6%

Table 3: Statistical characteristics in parametric and semiparametric logit fits. Bold values are significant at 10%.



Misclassification and Performance Curves

threshold s		Logit	nonparametric in X2	X3	X4	X5	X7	X4,X5	X2, X4,X5
0.25		129	133	129	136	130	128	132	130
	"ND"	41	44	40	49	40	40	46	40
	"D"	88	89	89	87	90	80	86	90
0.5		111	110	111	111	110	108	111	110
	"ND"	5	5	5	5	5	2	5	4
	"D"	106	105	106	106	105	106	106	106
0.75		107	107	107	107	107	107	107	107
	"ND"	0	0	0	0	0	0	0	0
	"D"	107	107	107	107	107	107	107	107

Table 4: Misclassifications for $\widehat{Y}$ ="Default" if $F(S) \leq s$ and $\widehat{Y}$ ="Non-Default" if $F(S) > s$. Validation data set.

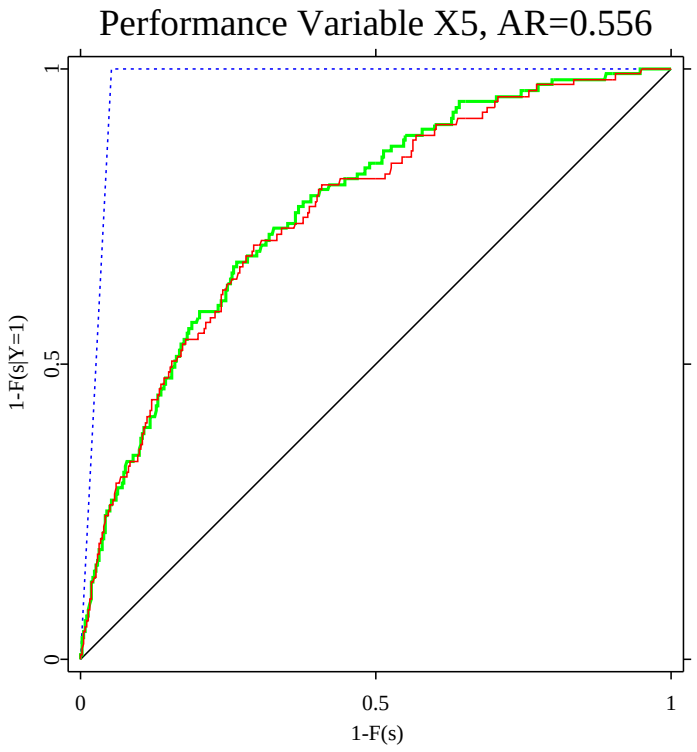
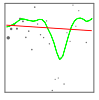


Figure 15: Performance curves, parametric logit (red) and semiparametric logit (green) with variable X5 included nonparametrically. Validation data set.



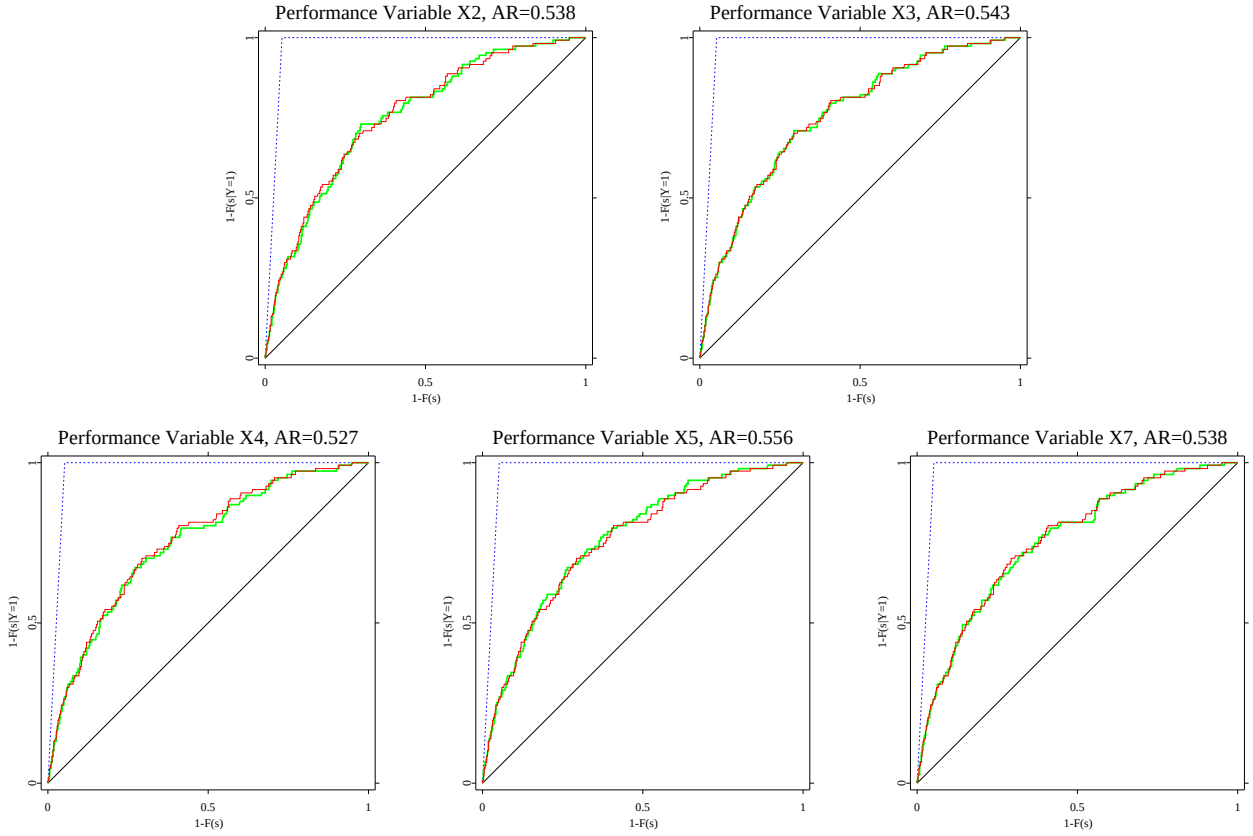
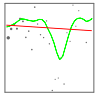


Figure 16: Performance curves with variables X2 to X5 (separately) included nonparametrically. Validation data set.



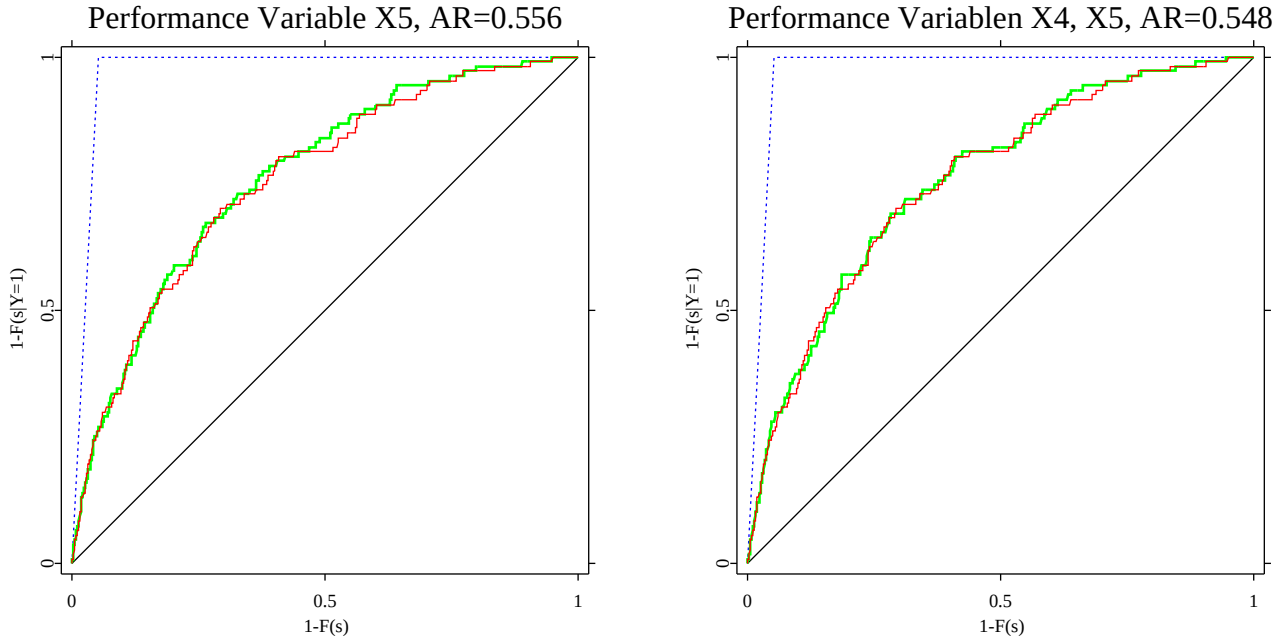
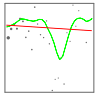


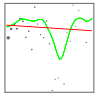
Figure 17: Performance curves with variables X5 (left) and with variables X4, X5 (right) jointly included nonparametrically. Validation data set.



Summary

nonparametric components

- combine advantages of nonparametric discriminance analysis with easy interpretation, estimation od PDs
- can be used as an exploratory tool to determine nonlinear transformations of variables
- can be used as an exploratory tool to determine interaction between variables
- allow for specification tests



Outlook

- combination of semiparametric aproaches with panel estimators/ GEE/ ordered responses
- optimal smoothing parameters
- combination with existing methods for additive models (joint implementation