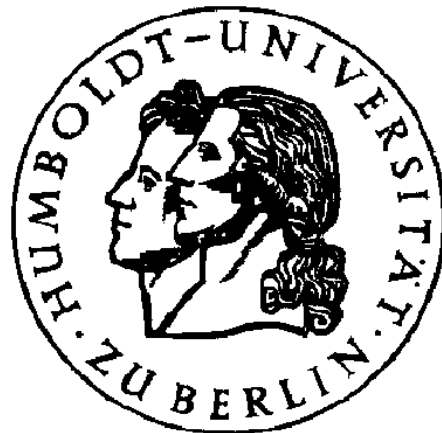


Credit Scoring using Semiparametric Methods

Marlene Müller

Humboldt University Berlin (Germany)



Plan

- Problem and Data Description
- Logistic Credit Scoring
- Semiparametric Credit Scoring
- Testing the Semiparametric Model
- Miss-classification and Performance Curves



Problem and Data Description

- Response variable Y
(credit worthiness, 0= "good" , 1= "bad")
- Metric variables X2 to X9.
- Categorical variables X10 to X24.

	Estimation data set	Validation data set
0 ("good")	5808 (94%)	2045 (94.8%)
1 ("bad")	372 (6%)	113 (5.2%)
total	6180	2158

Table 1: Responses.



Aim

model and prediction for

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

methods

- linear / quadratic discriminant analysis
- logistic discriminant analysis
- alternatives:
 - nonparametric (nearest neighbor)
 - neural networks
 - classification trees



Density Plots

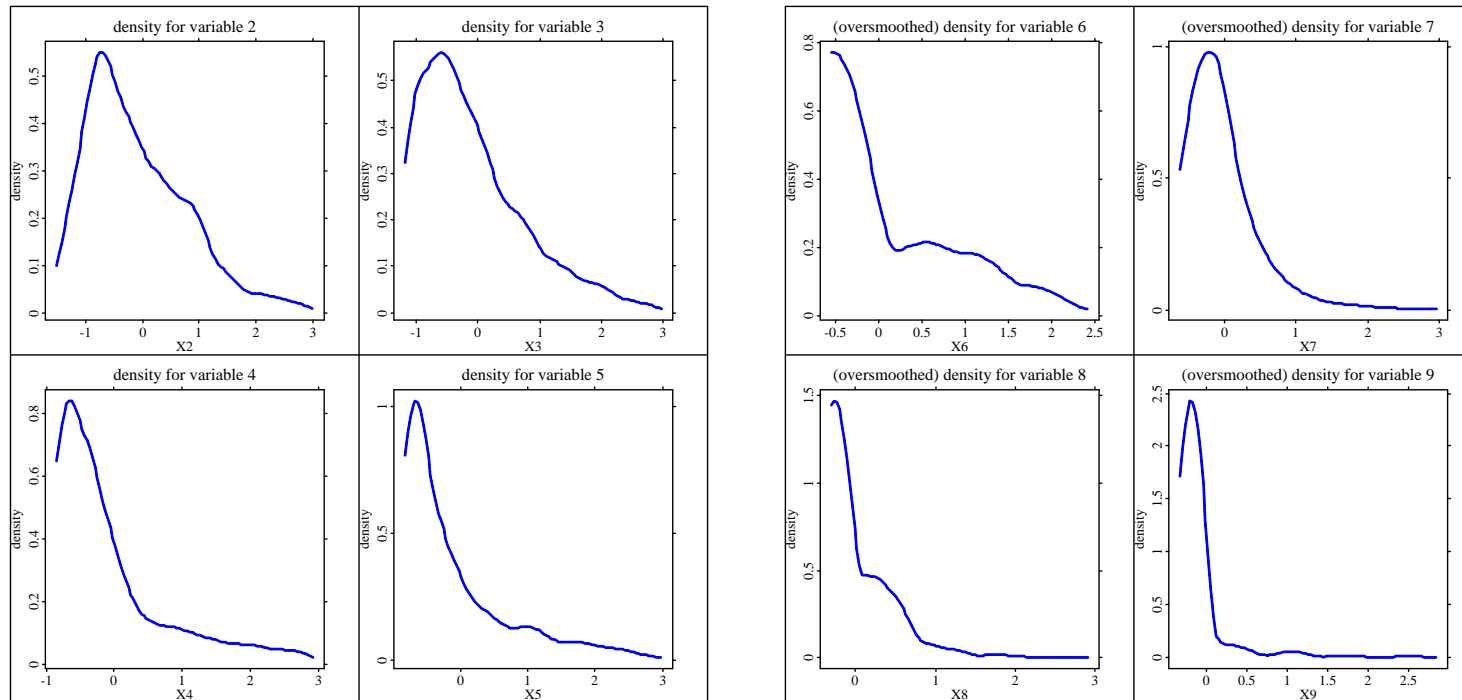


Figure 1: Kernel density estimates, variables X2 to X5 (left) and X6 to X9 (right).



Scatter Plots

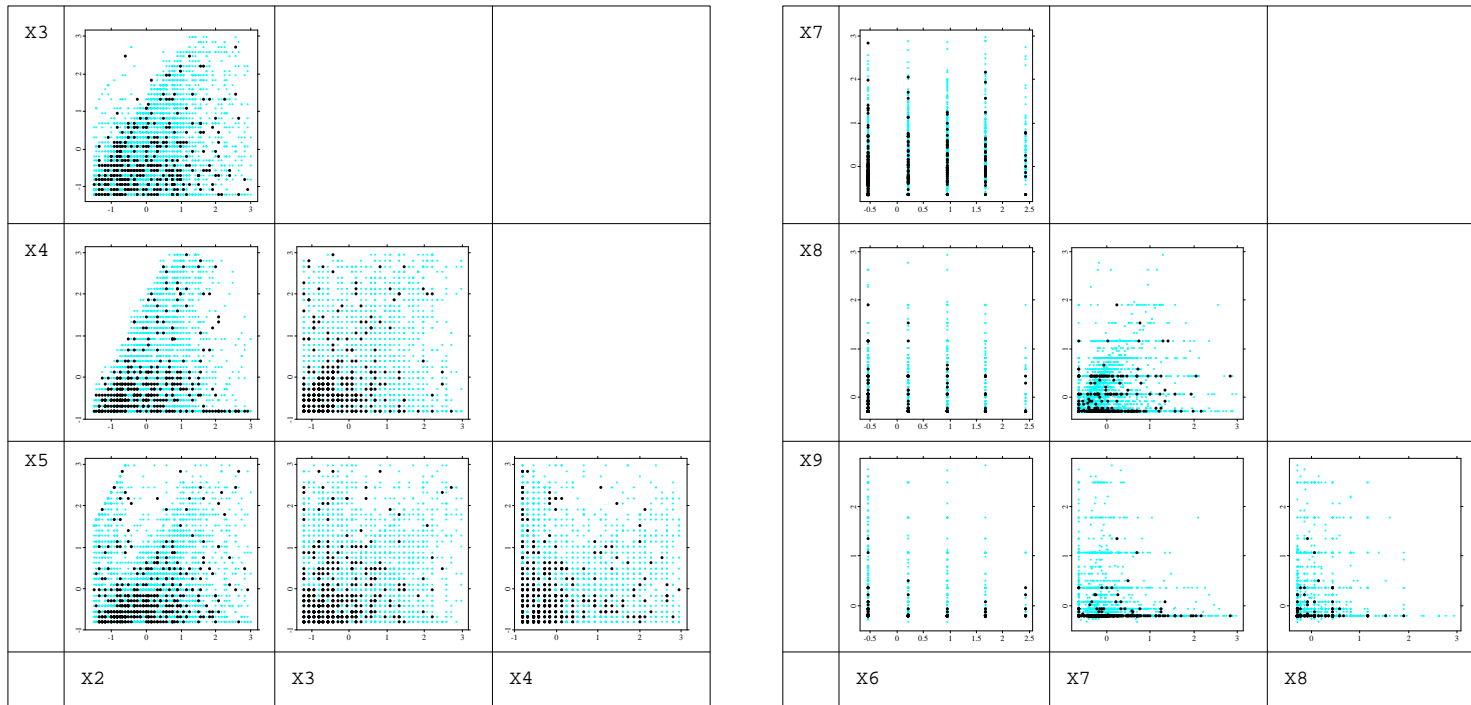


Figure 2: Scatterplots, variables X2 to X5 (left) and X6 to X9 (right). Observations corresponding to $Y=1$ are emphasized in black.



Logistic Credit Scoring

logit model (logistic discriminant analysis)

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

X_j denotes here

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



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				
X18#6	0.463575	0.2399	1.93				
X18#7	0.568302	0.2579	2.20				
				df			6118
				Log-Lik.			-1199.6278
				Deviance			2399.2556



Semiparametric Credit Scoring

generalized partial linear model (GPLM)

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

where

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

Reference for estimation: Severini & Staniswalis (1994)



E.g.: to include variable X_5 in a nonlinear way:

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

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



	Logit	Nonparametric in					
		X2	X3	X4	X5	X4,X5	X2, X4,X5
const.	-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 2: Parametric coefficients for variables X2 to X9. Bold values are significant at 5%.



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

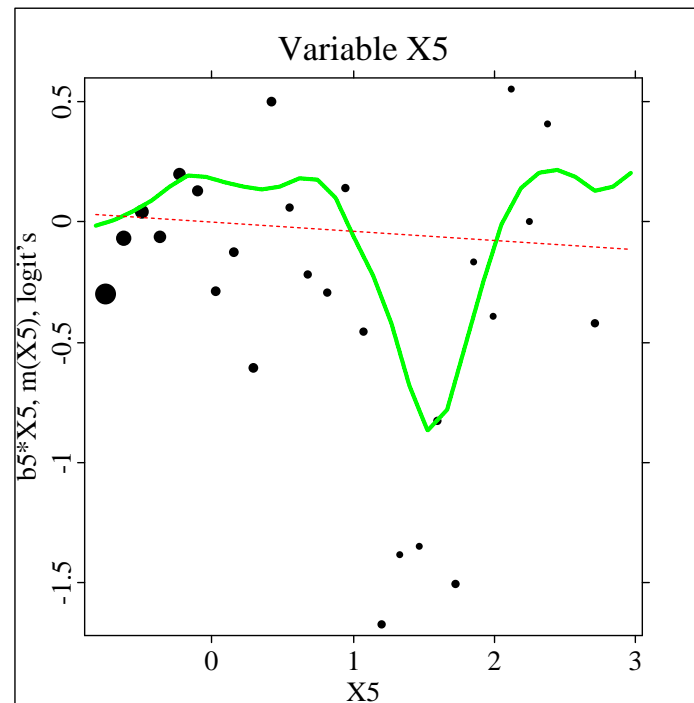


Figure 3: Marginal dependence, variable X5. Thicker bullets correspond to more observations. Parametric (red) and GPLM logit fit (green).



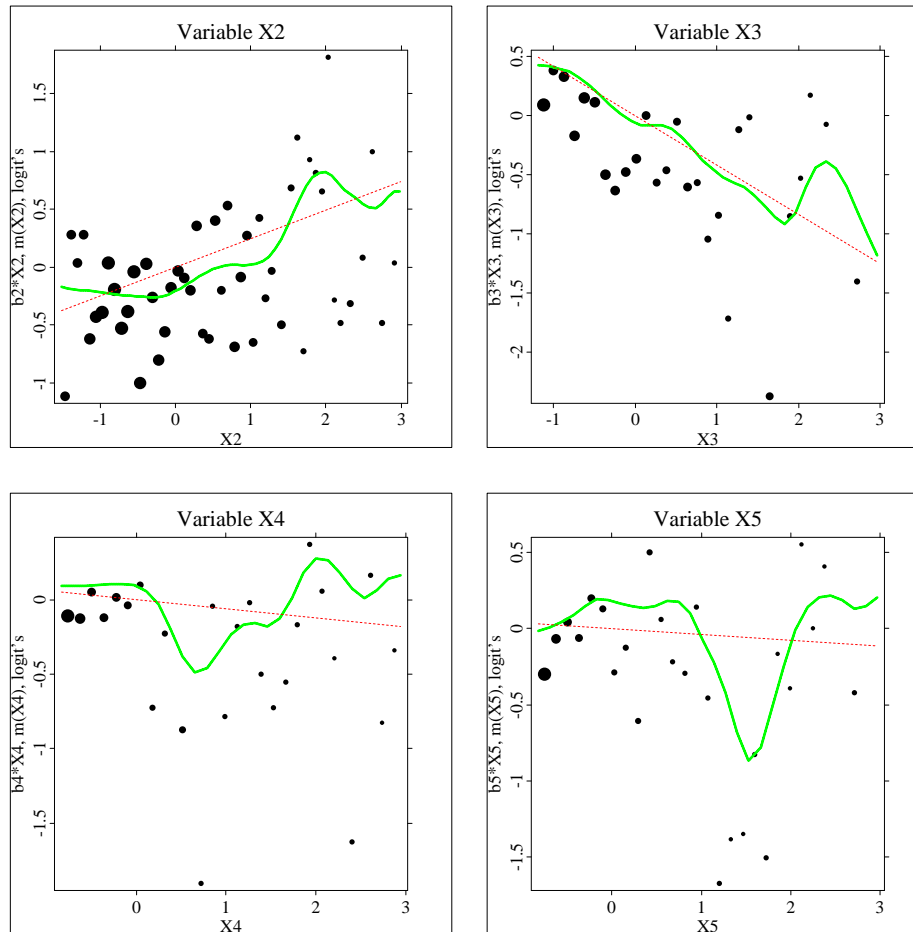


Figure 4: Marginal dependencies, variables X2 to X5. Parametric logit fits (red) and GPLM logit fits (green).



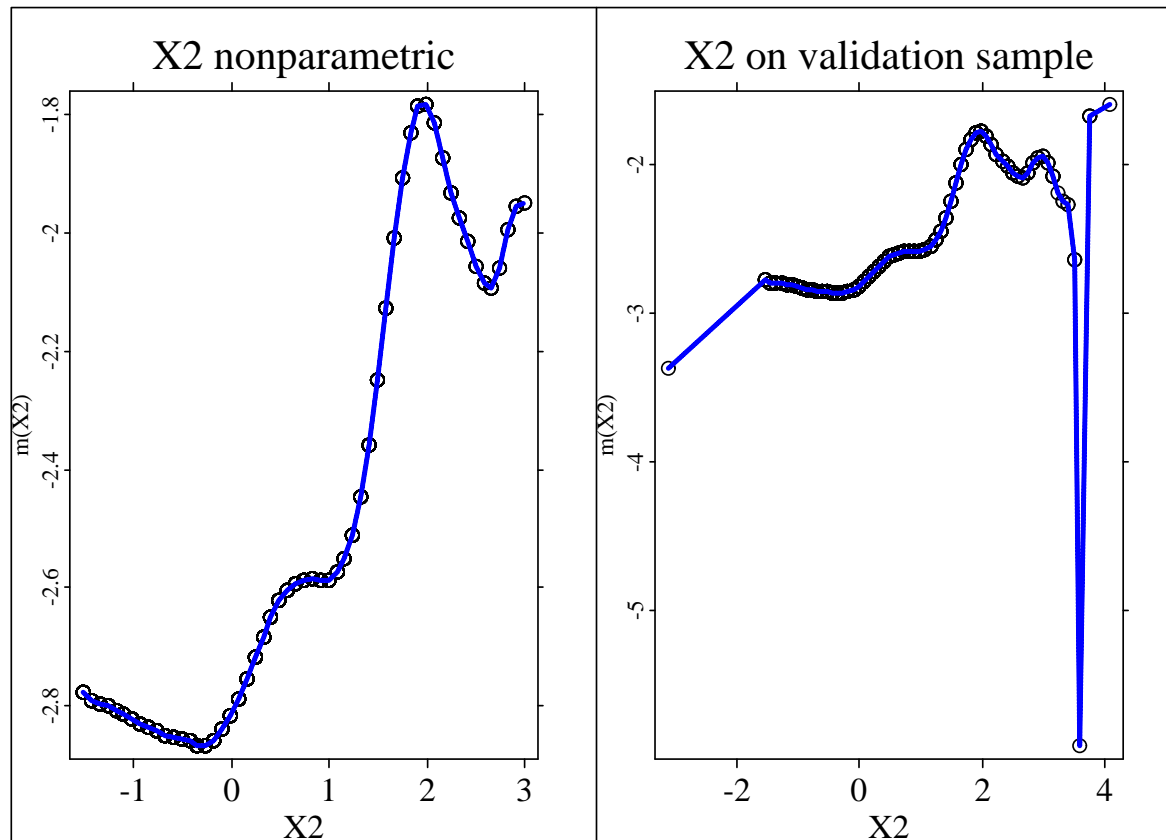


Figure 4.2: Semiparametric logit model, nonparametric curve for variable X_2 . Estimation data set (left panel) and validation data set (right panel), bandwidth $h = 15\%$.



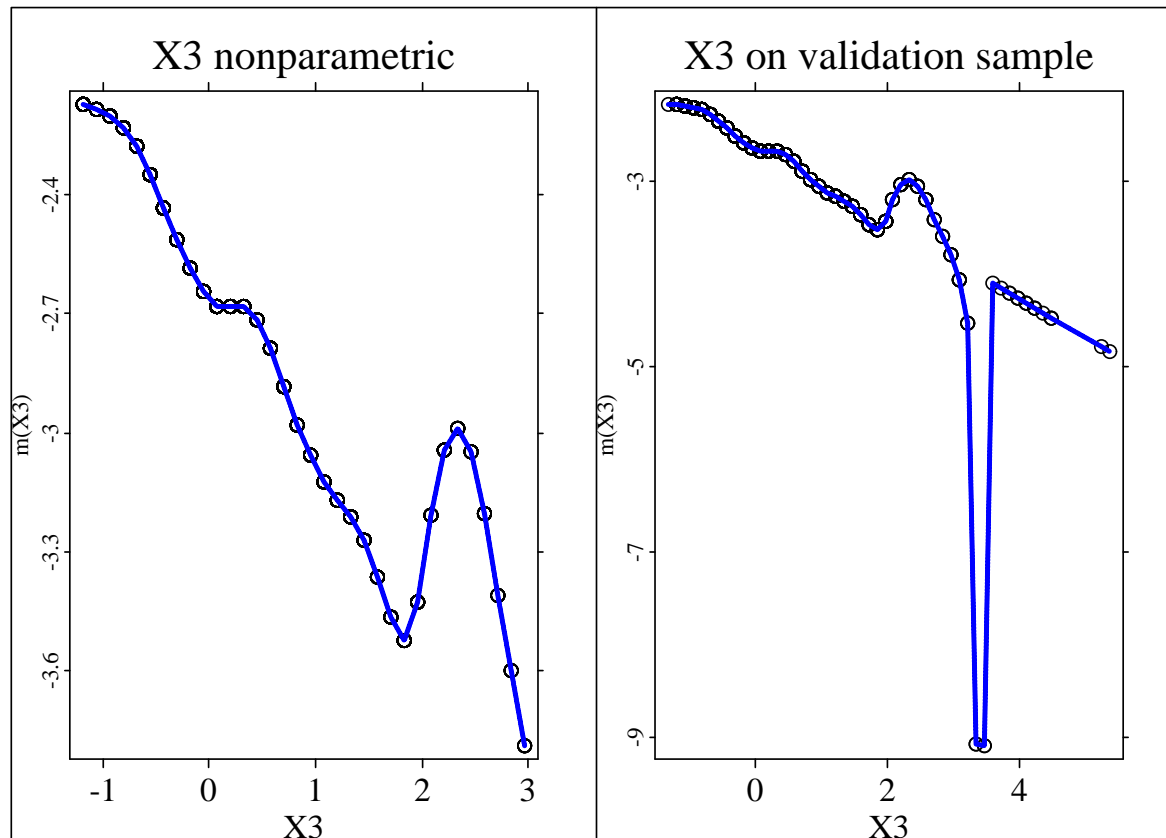


Figure 4.3: Semiparametric logit model, nonparametric curve for variable X_3 . Estimation data set (left panel) and validation data set (right panel), bandwidth $h = 15\%$.



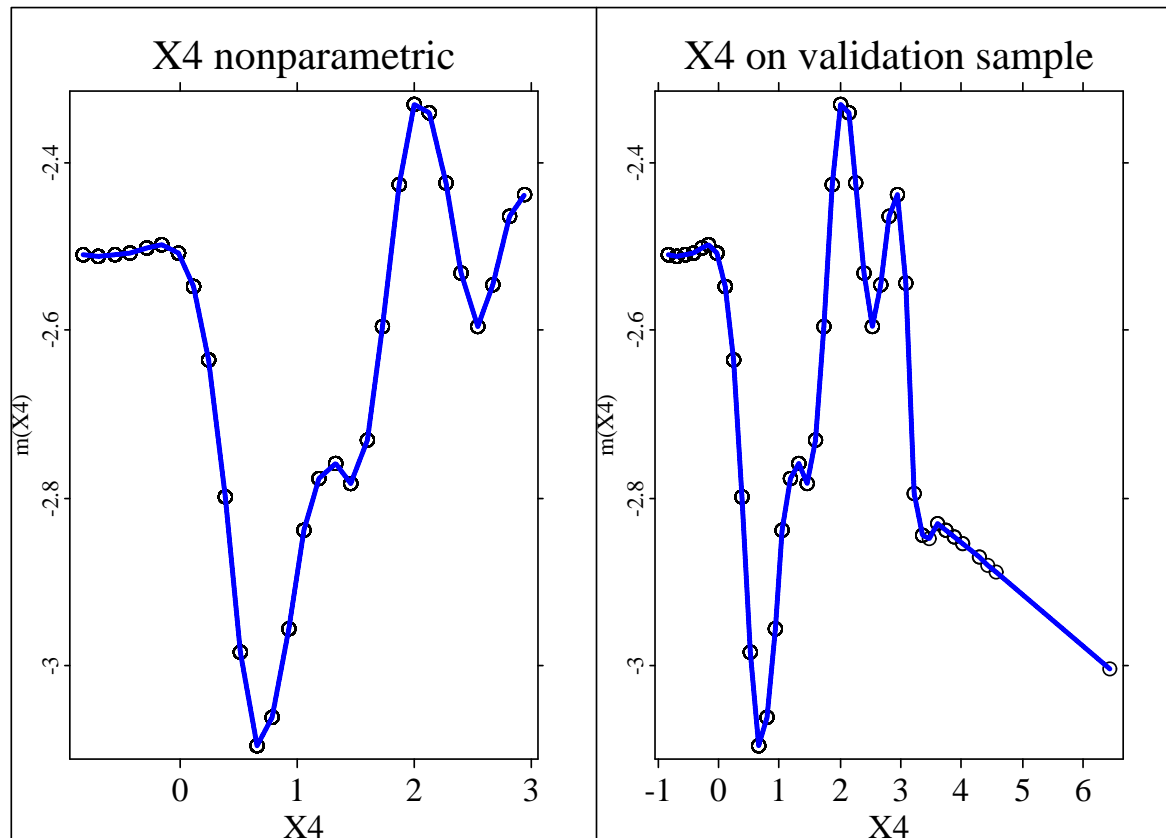


Figure 4.4: Semiparametric logit model, nonparametric curve for variable X_4 . Estimation data set (left panel) and validation data set (right panel), bandwidth $h = 15\%$.



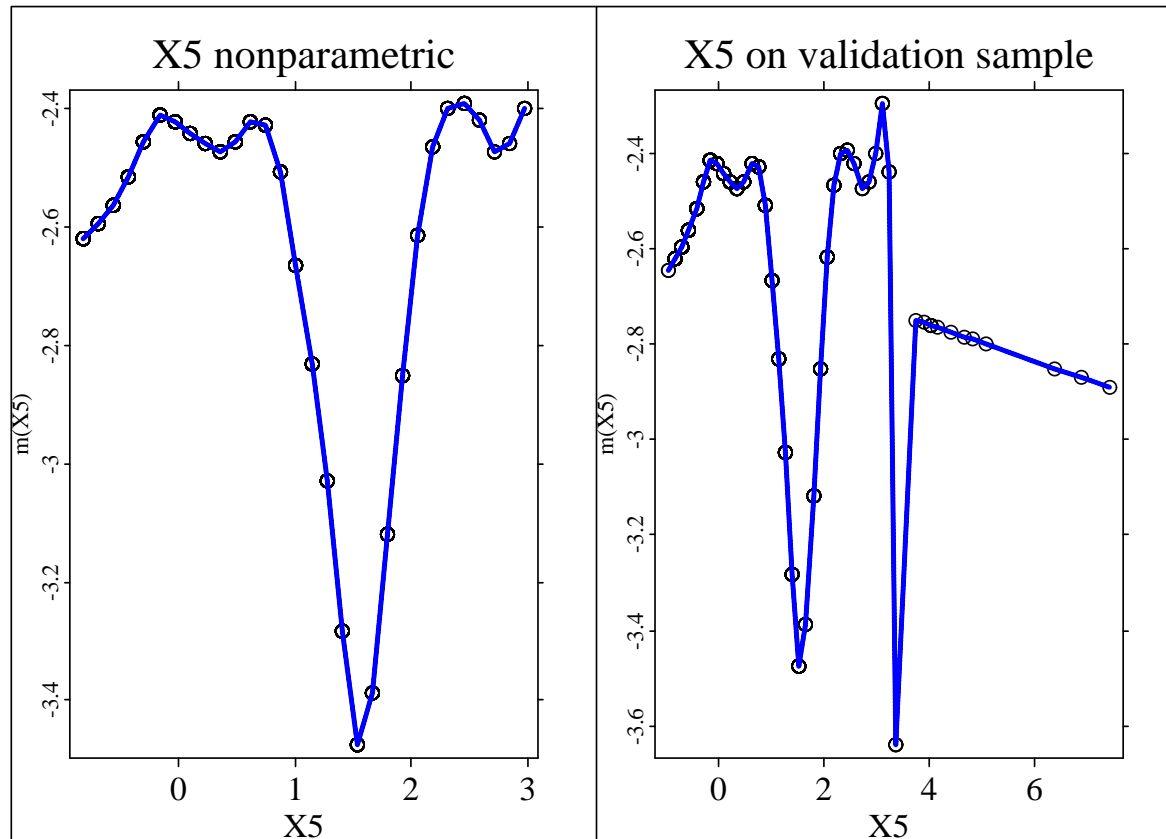


Figure 4.5: Semiparametric logit model, nonparametric curve for variable X_5 . Estimation data set (left panel) and validation data set (right panel), bandwidth $h = 15\%$.



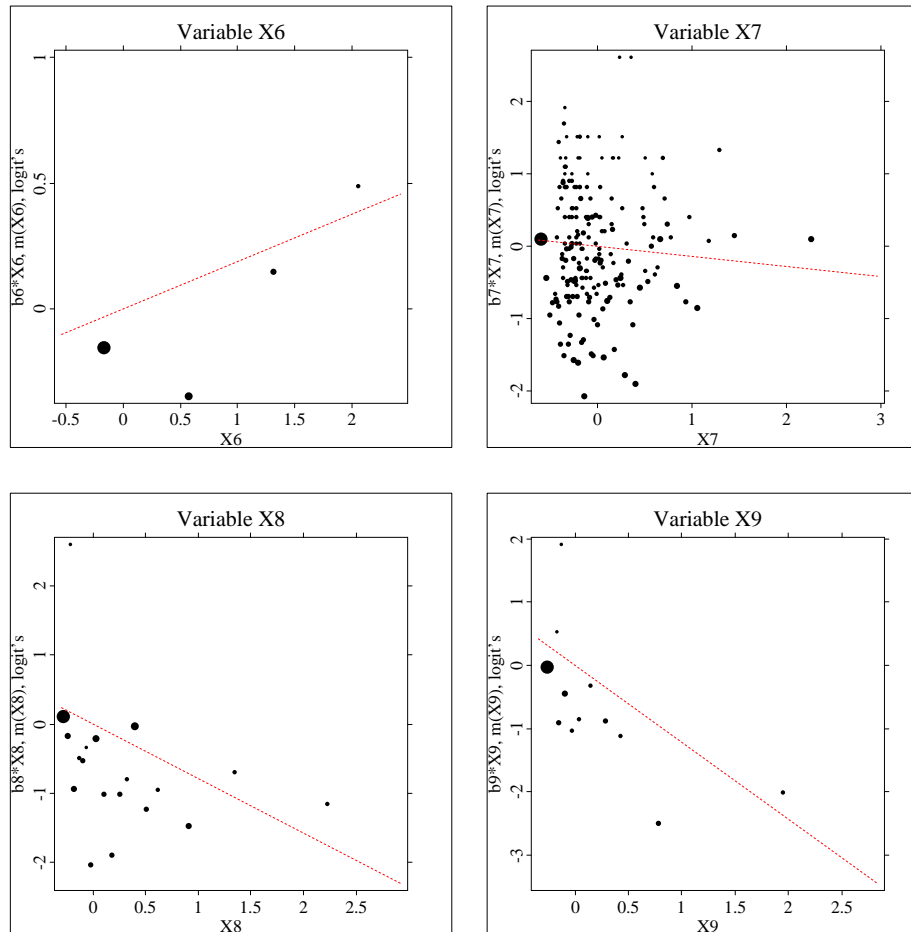


Figure 5: Marginal dependencies, variables X6 to X9. Parametric logit fits (red).



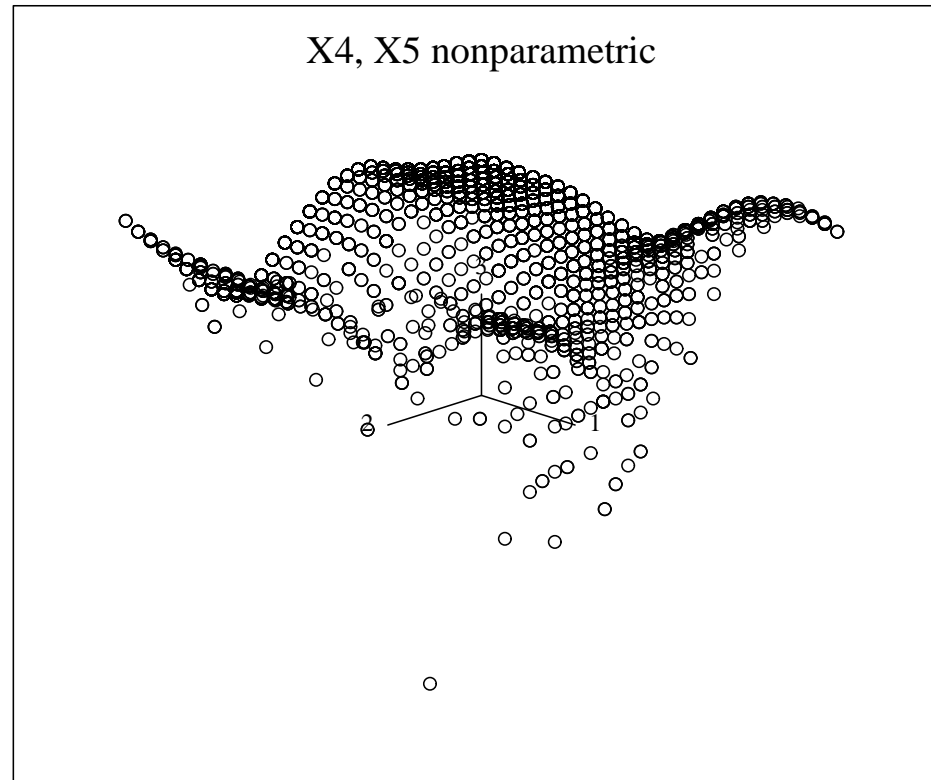


Figure 6: Bivariate nonparametric surface for variables X4, X5.

Testing the Semiparametric Model

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



Miss-classification and Performance Curves

Threshold s	Logit	Nonparametric in					
		X2	X3	X4	X5	X4,X5	X2, X4,X5
0.25	136	139	136	142	136	138	138
"good"	42	44	41	49	40	44	42
"bad"	94	95	95	93	96	94	96
0.5	117	116	117	117	116	117	116
"good"	5	5	5	5	5	5	4
"bad"	112	111	112	112	111	112	112
0.75	113	113	113	113	113	113	113
"good"	0	0	0	0	0	0	0
"bad"	113	113	113	113	113	113	113

Table 4: Miss-classifications for $\hat{Y} = \text{"bad"}$ if $F(S) \geq s$ and $\hat{Y} = \text{"good"}$ if $F(S) < s$. Validation data set.



Performance curves (Lorenz curves)

- calculate scores, e.g.

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

- plot $P(S < s)$ (classified as “good”) versus $P(S < s | Y = 1)$ (classified as “good” although observation is “bad”)



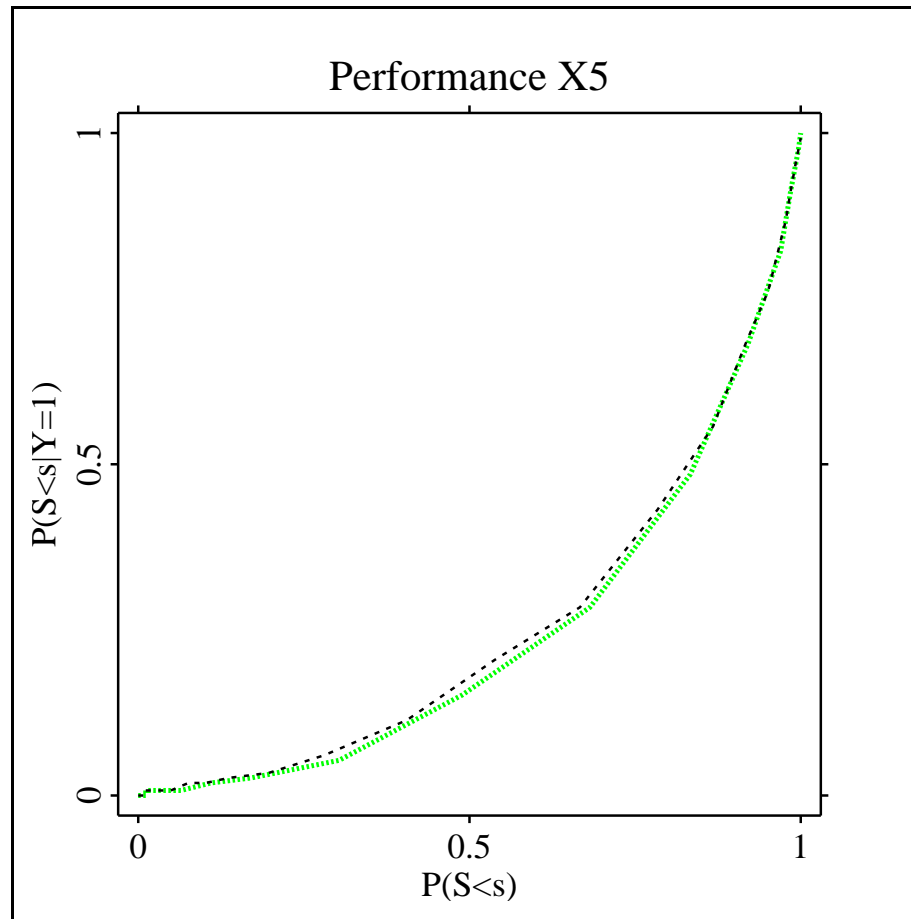


Figure 7: Performance curve, parametric logit (dashed) and semiparametric logit (green) with variable X5 included nonparametrically. Validation data set.



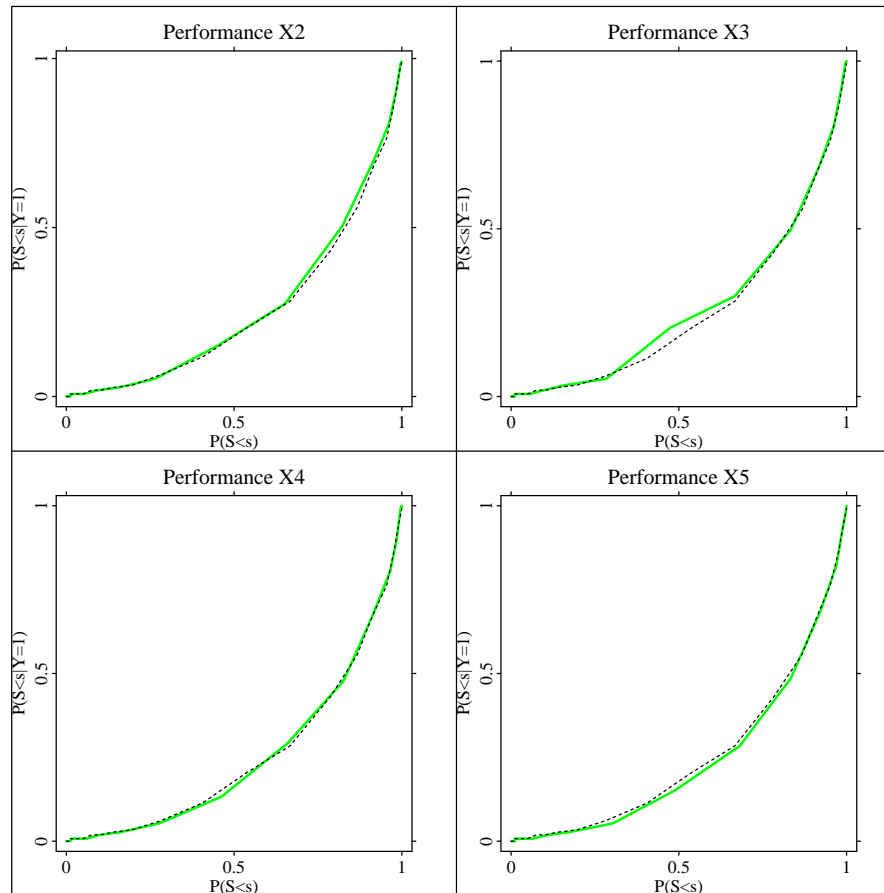


Figure 8: Performance curves with variables X2 to X5 (separately) included nonparametrically. Validation data set.



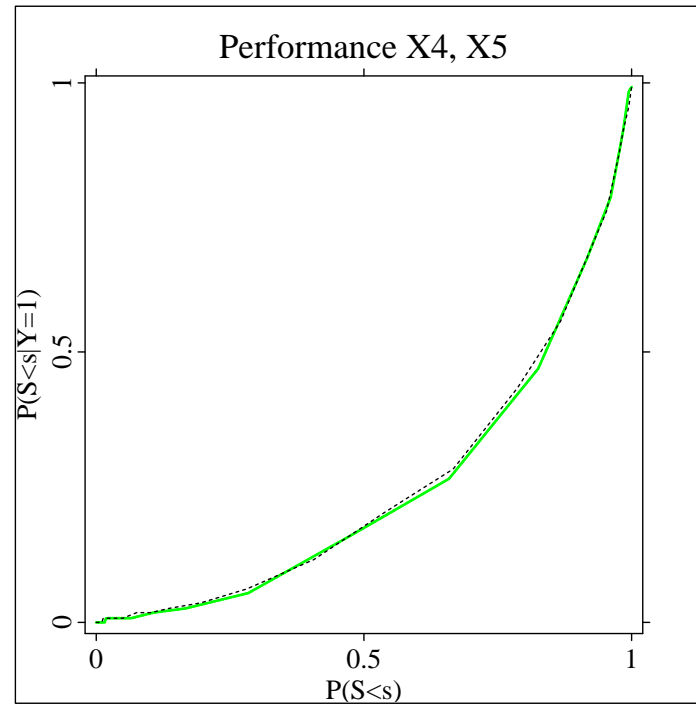
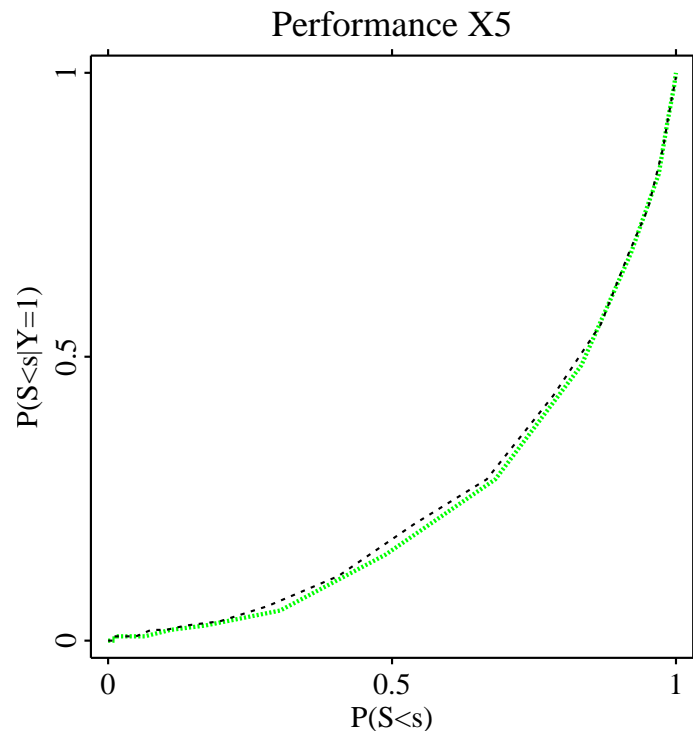


Figure 9: Performance curves with variables X5 (left) and with variables X4, X5 (right) jointly included nonparametrically. Validation data set.

