# Semiparametrisches Kredit Scoring

Marlene Müller


Fraunhofer Institut für Techno- und Wirtschaftsmathematik (ITWM)

Kaiserslautern

Bernd Rönz, Wolfgang Härdle

Center for Applied Statistics and Economics (CASE)

Humboldt-Universität zu Berlin

## Übersicht

- Ausfallwahrscheinlichkeiten
- Problem und Datenbeschreibung
- Logistisches Kredit Scoring
- Semiparametrisches Kredit Scoring
- Test des semiparametrischen Modells
- Missklassifikation und Performance-Kurven

# Kredit Rating/Scoring

- neue Bedeutung durch Basel II:
  Eigenkapitalunterlegung von Kreditrisiken angepaßt an individuelles Kreditrisiko
- Möglichkeiten für Banken:
  - ⋆ Ratings und Ausfallwahrscheinlichkeiten externer Ratingagenturen
  - ⋆ interne Ratings (IRB-Ansatz)
    $\rightarrow$ bessere Anpassung an eigenes Kreditportfolio
- Kernproblem:
  Schätzung von Ausfallwahrscheinlichkeiten (PDs)

Referenz: The New Basel Capital Accord ("Basel II")

## Aus Ausfallwahrscheinlichkeiten ....

werden bestimmt

- Ratingklassen

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

  entsprechen z.B. PDs

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

- erwarteter Verlust

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

## Methoden zur Bestimmung von PDs

- Diskriminanzanalyse/Klassifikationsmethoden
  $\rightarrow$ Bestimmung von Scores
- kategorielle Regressionsmodelle (Logit/Probit, Panel, geordnete Kategorien)
  $\rightarrow$ Bestimmung von Scores + Ausfallwahrscheinlichkeiten
- Merton-Ansatz (Aktie des Unternehmens wird als Option auf Firmenwert betrachtet)
  $\rightarrow$ Ausfallwahrscheinlichkeit durch "distance to default"
- Jarrow, Lando, Turnbull (Übergangswahrscheinlichkeiten)

## Kategorielle Regressionsmethoden

Modell und Prognose für

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

- Logit-Modelle (Logistische Diskriminanzanalyse)
- Modifikationen
  - Probit-Modell (andere Linkfunktion)
  - Panelmodelle
  - Stichprobenselektionskorrektur (Heckman-Schätzer)

# Datenbeispiele

## Stichprobe für Autokredite

| | | Ja | Nein | (in %) | |
|---|---|---|---|---|---|
| $Y$ | **Kreditausfall** | 26.4 | 73.6 | | |
| | **Bisherige Kredite OK** | 66.2 | 33.8 | | |
| | **Beschäftigt** | 73.2 | 26.8 | | |
| | | Min | Max | Mittel | St.Abw. |
| | **Kreditdauer (Monate)** | 4 | 54 | 21.8 | 10.6 |
| | **Kreditbetrag (DM)** | 428 | 14179 | 3902.3 | 2621.9 |
| | **Alter (in Jahren)** | 19 | 75 | 34.2 | 10.8 |

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

## Französische Bankdaten

- Abhängige Variable Y
  (Kreditstatus, 0=“Nicht-Ausfall”, 1=“Ausfall”)
- Metrische Variablen X2 bis X9.
- Kategorielle Variablen X10 bis X24.

| | Schätz-stichprobe | Validierungs-stichprobe |
|---|---|---|
| 0 (“Nicht-Ausfälle”) | 5808 (94%) | 1891 (94.6%) |
| 1 (“Ausfälle”) | 372 ( 6%) | 107 ( 5.4%) |
| total | 6180 | 1998 |

Tabelle 1. Zusammenfassung.

# Deskriptive Analyse

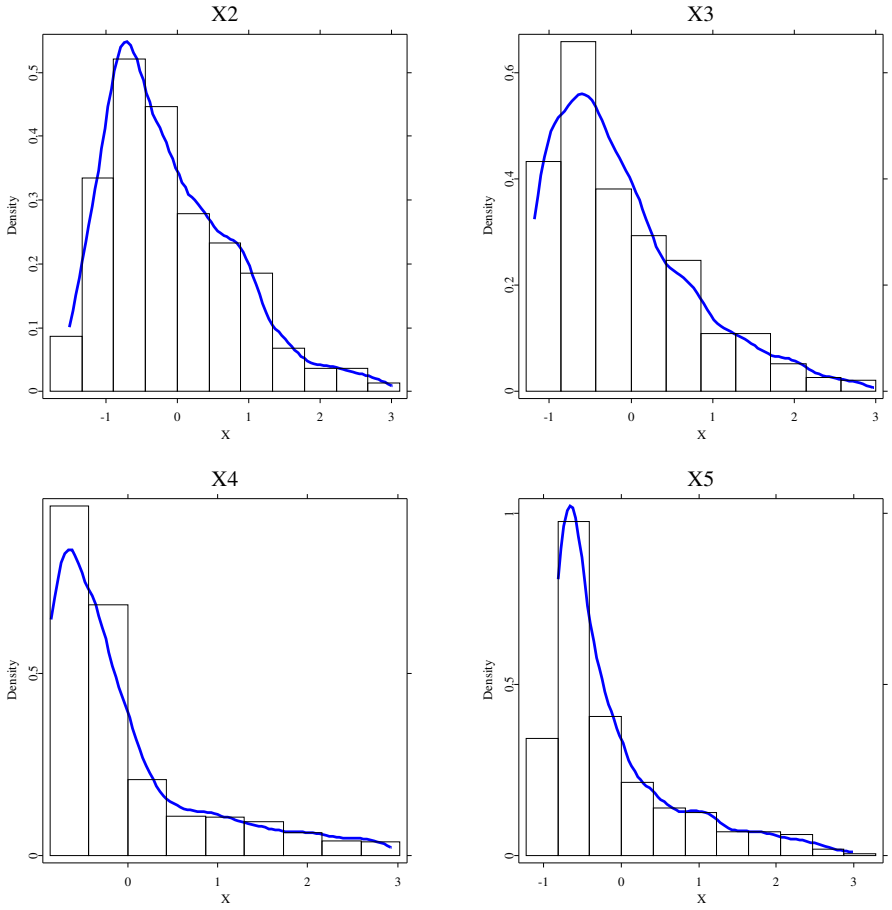



Abbildung 1. Kerndichteschätzungen, Variablen X2 bis X5 .

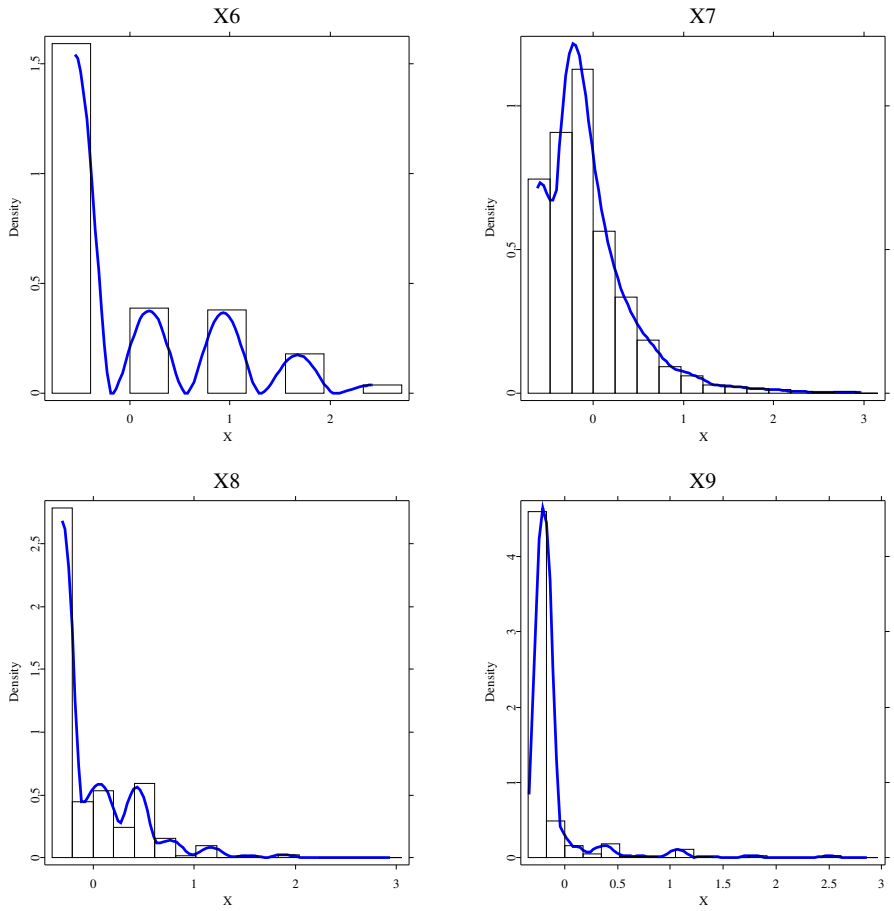



Abbildung 2. Kerndichteschätzungen, Variablen X6 bis X9.

# Scatterplots

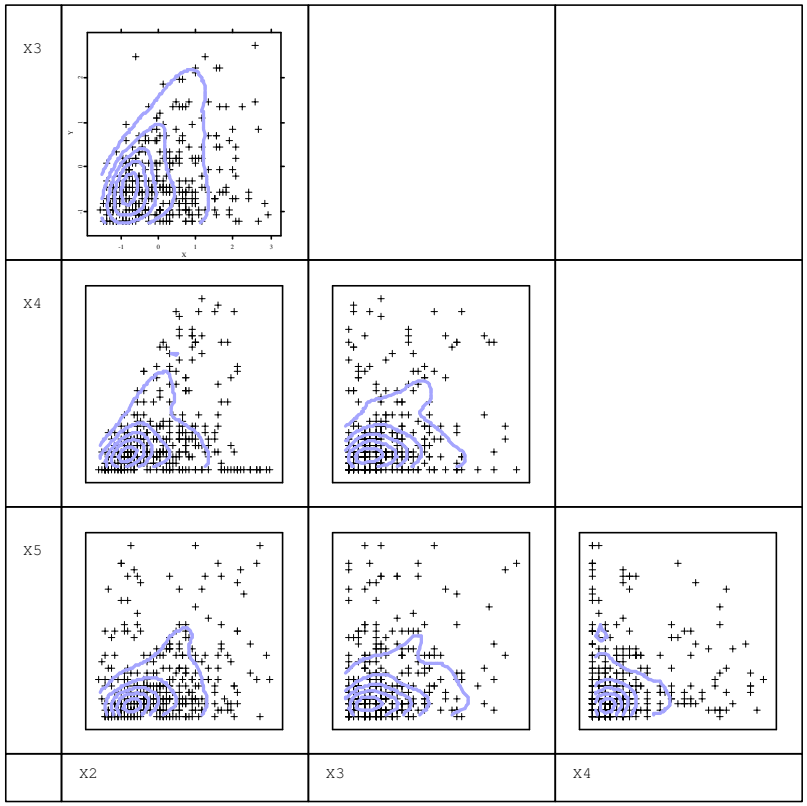



Abbildung 3. Scatter-Kontur-Plots, Variablen X2 bis X5. (Beobachtungen zu Y=1 in schwarz.)

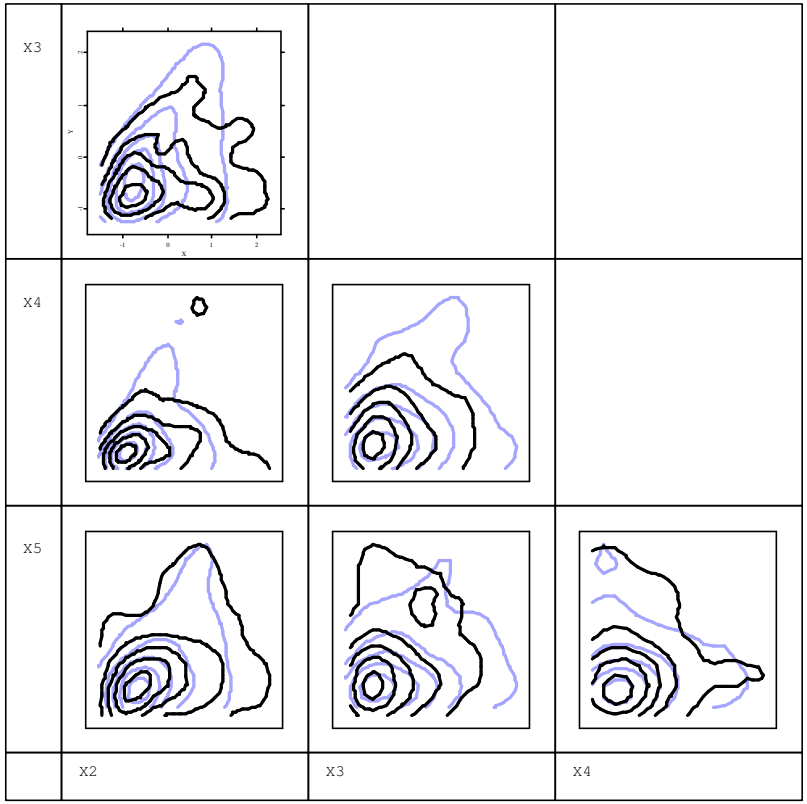



Abbildung 4. Kontur-Kontur-Plots, Variablen X2 bis X5. (Beobachtungen zu Y=1 in schwarz.)

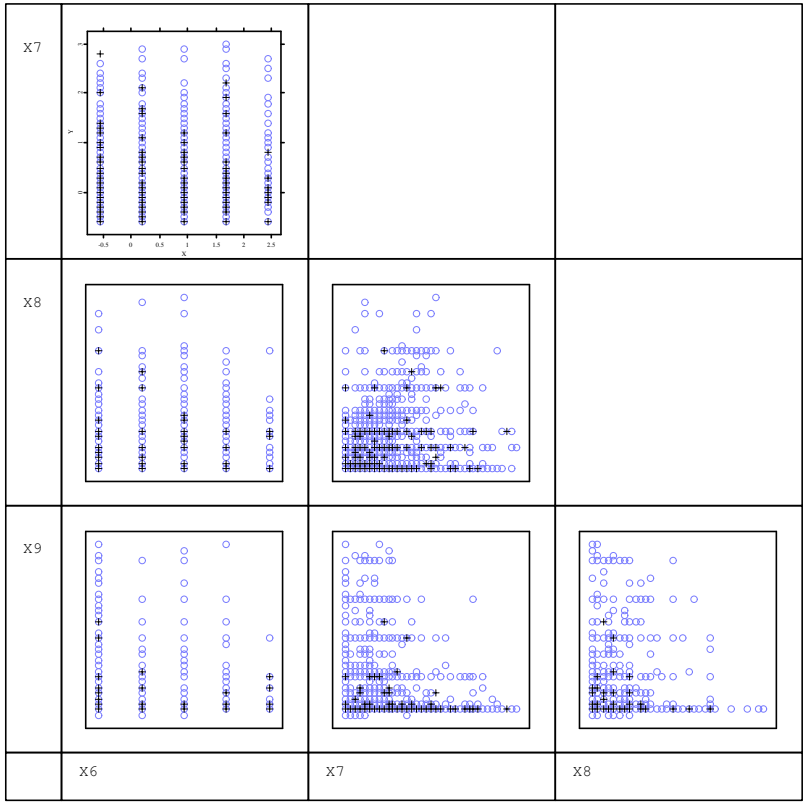



Abbildung 5. Scatter-Plots, Variablen X6 bis X9. (Beobachtungen zu Y=1 in schwarz.)

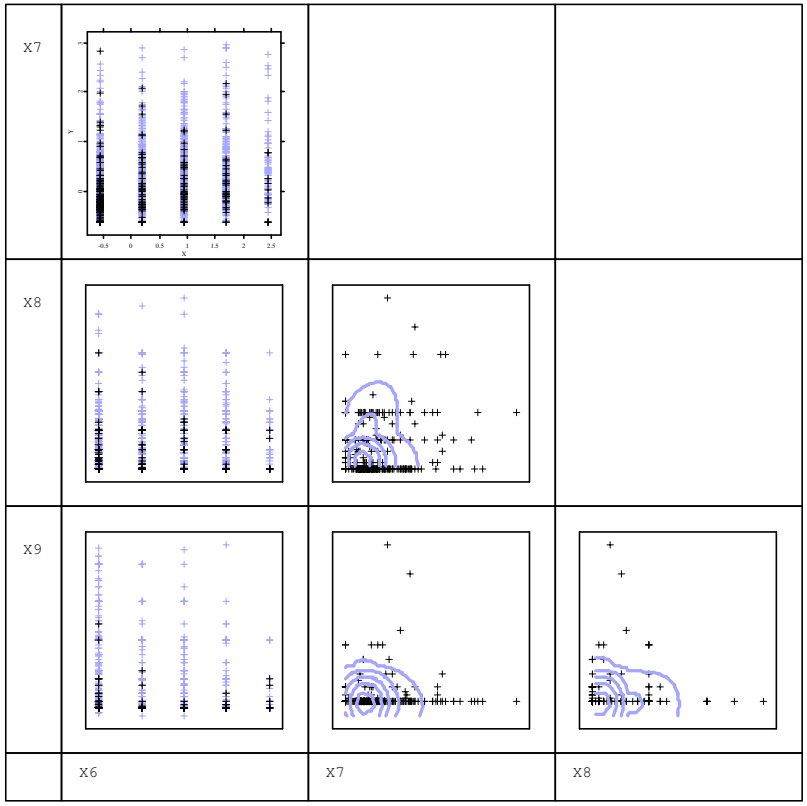



Abbildung 6. Scatter-Kontur-Plots, Variablen X6 bis X9. (Beobachtungen zu Y=1 in schwarz.)

# Logistisches Kredit Scoring

Logit-Modell (logistische Diskriminanzanalyse)

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

$X_j$ bezeichnet

- $j$te Variable, falls X$j$ metrisch ($j \in \{2, \ldots, 9\}$)
- Vektor von Dummies, falls X$j$ kategoriell ($j \in \{10, \ldots, 24\}$)

## Logit-Modell

- binäre abhängige Variable

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

- $Y^* =$ latente Variable, (negativer) Kredit-Score
- $v(\bullet) =$ Indexfunktion, die Beziehung zwischen $X$ und $Y^*$ modelliert, z.B.

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

- $u \sim F$ Fehlerterm.

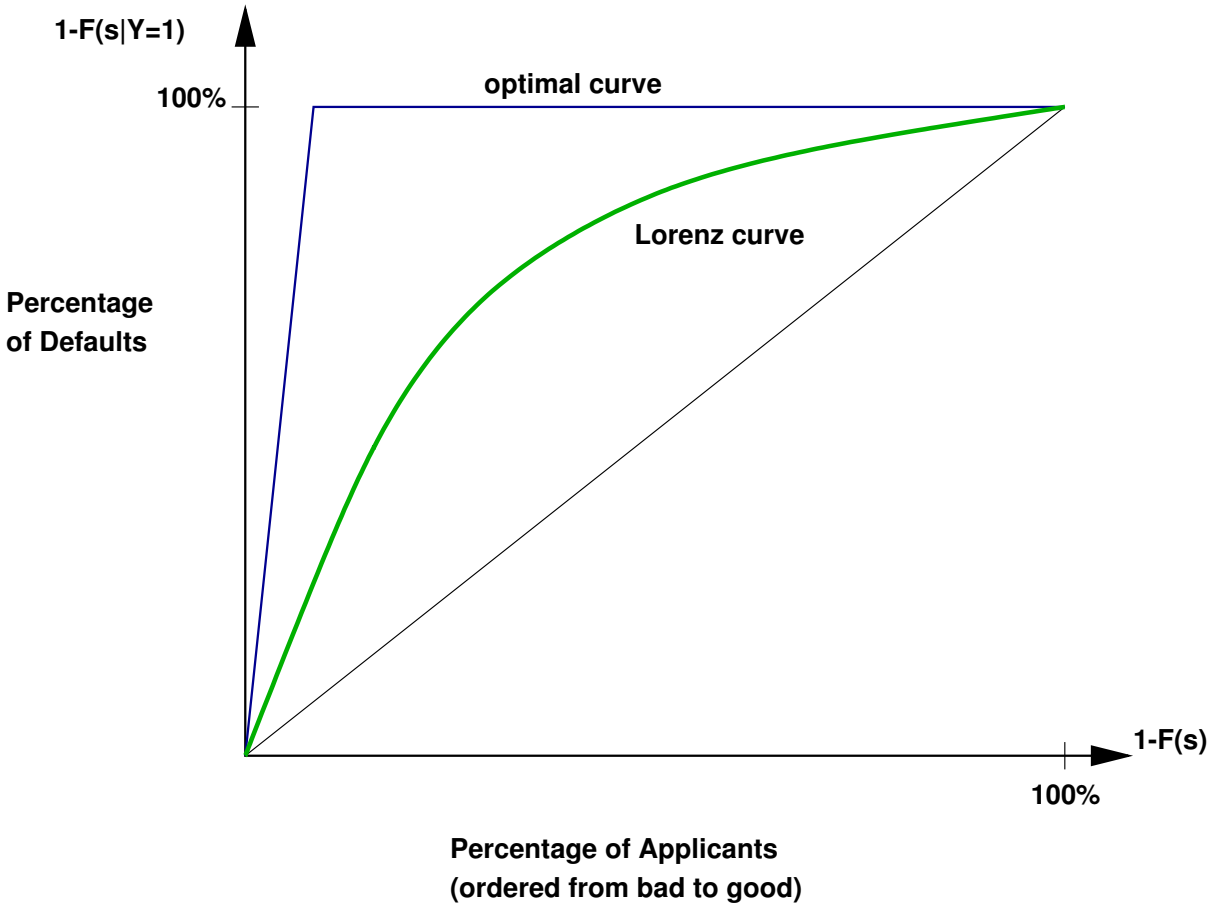
| Variable | Koeffizient | St.Abw. | $t$-Wert | Variable | Koeffizient | St.Abw. | $t$-Wert |
|---|---|---|---|---|---|---|---|
| X0 (const.) | **-2.605280** | 0.5890 | -4.42 | X19#2 | -0.086954 | 0.3082 | -0.28 |
| X2 | **0.246641** | 0.1047 | 2.35 | X19#3 | 0.272517 | 0.2506 | 1.09 |
| X3 | **-0.417068** | 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 | Devianz | | | 2399.2556 |

## Performance (Lorenzkurven)

- berechne Scores, z.B.

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

- plotte
  - $\star$ $1 - F(s) = P(S > s)$ ("Ausfall" klassifiziert) vs.
  - $\star$ $1 - F_1(s) = P(S > s | Y = 1)$
    ("Ausfall" klassifiziert und tatsächlich "Ausfall")


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)

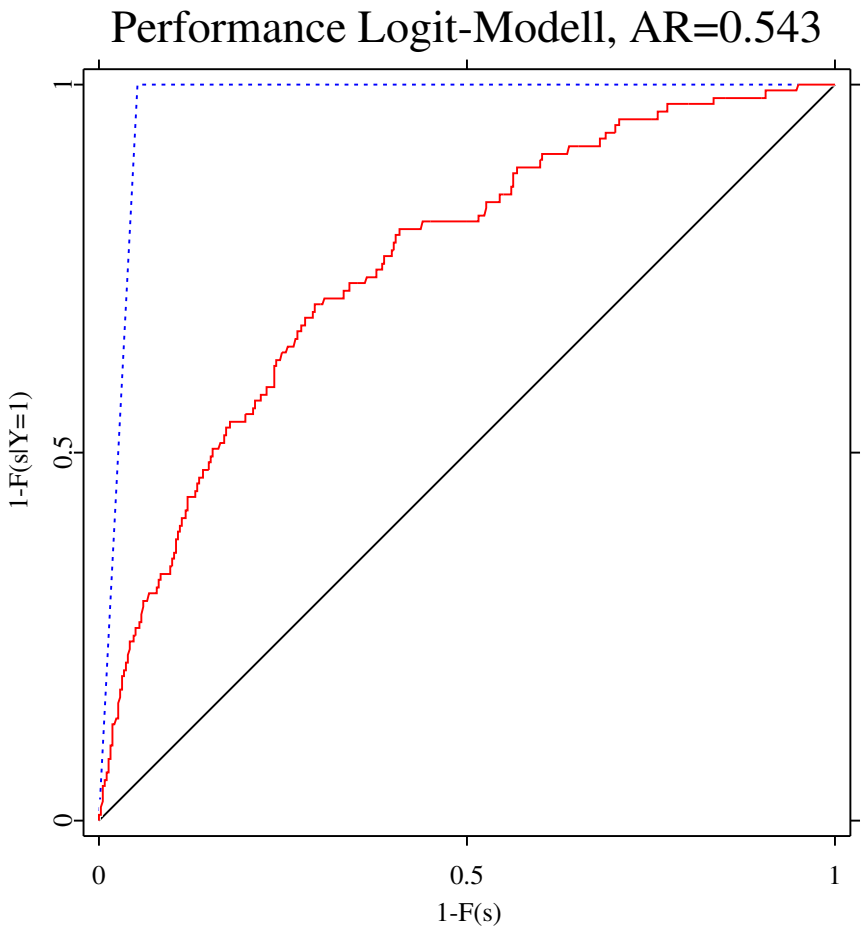



Abbildung 7. Performance-Kurve für Logit-Modell (rot) und optimale Kurve (blau).

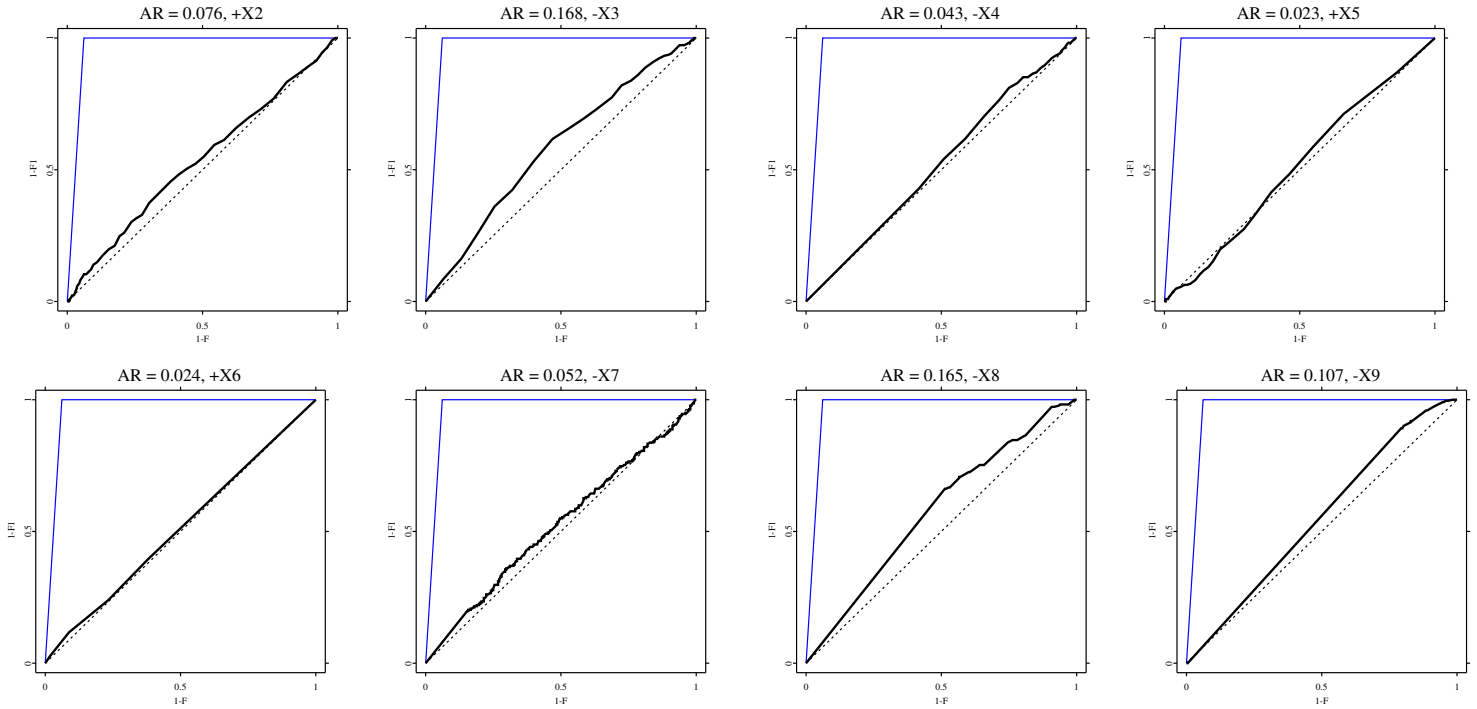



Abbildung 8. Lorenzkurven für alle Variablen

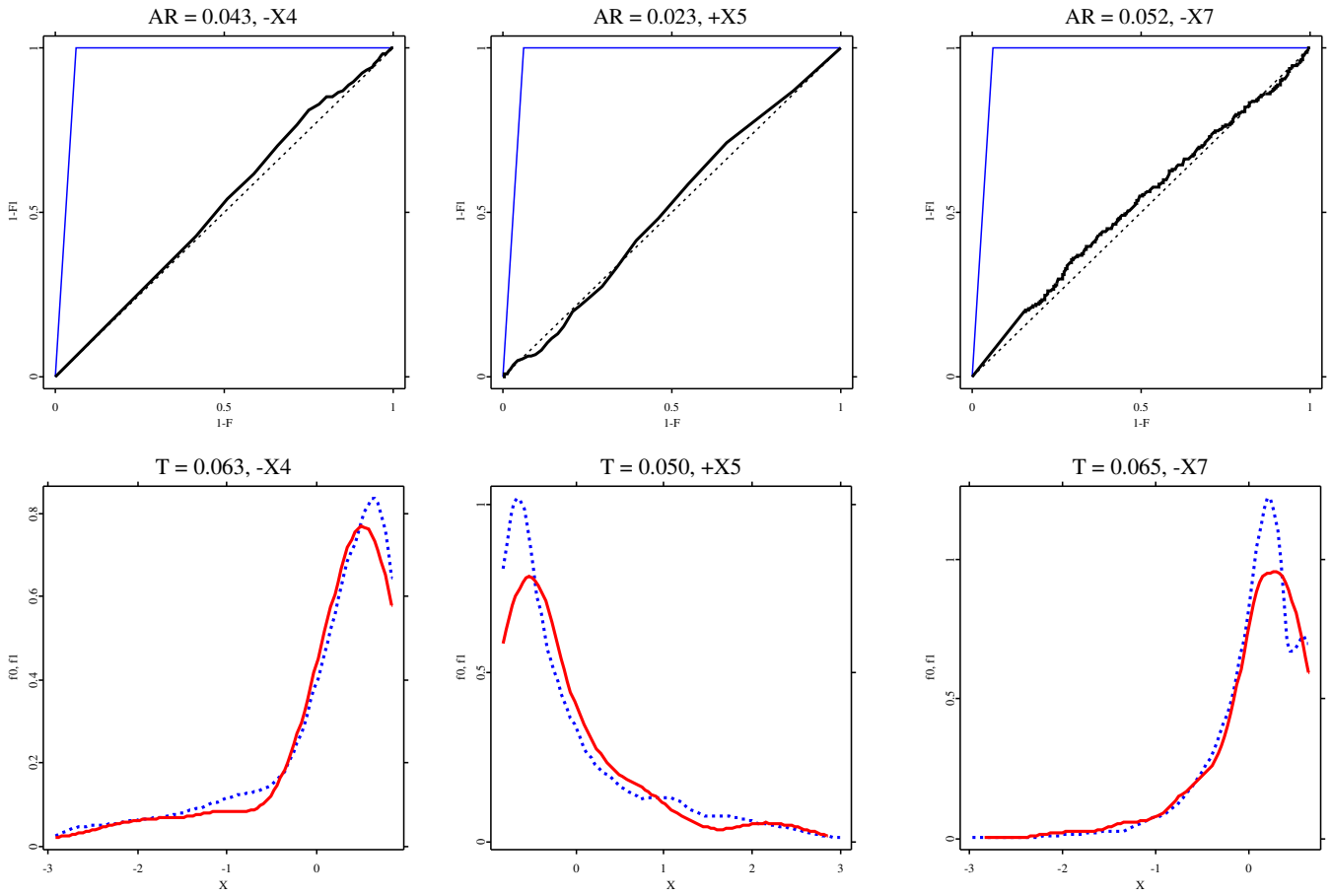



Abbildung 9. Lorenzkurven, Dichten (bed. auf $Y$) für ausgewählte Variablen.

# Wie hängt $Y$ (genauer: $\log(\frac{p}{1-p})$) von einer Variablen ab?

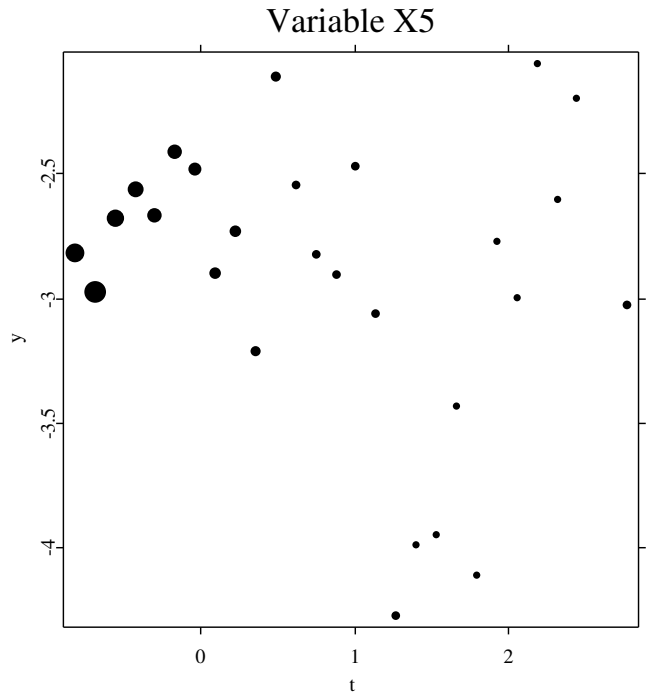



Abbildung 10. Marginale Abhängigkeit, Variable X5. Dickere Punkte entsprechen mehr Beobachtungen.

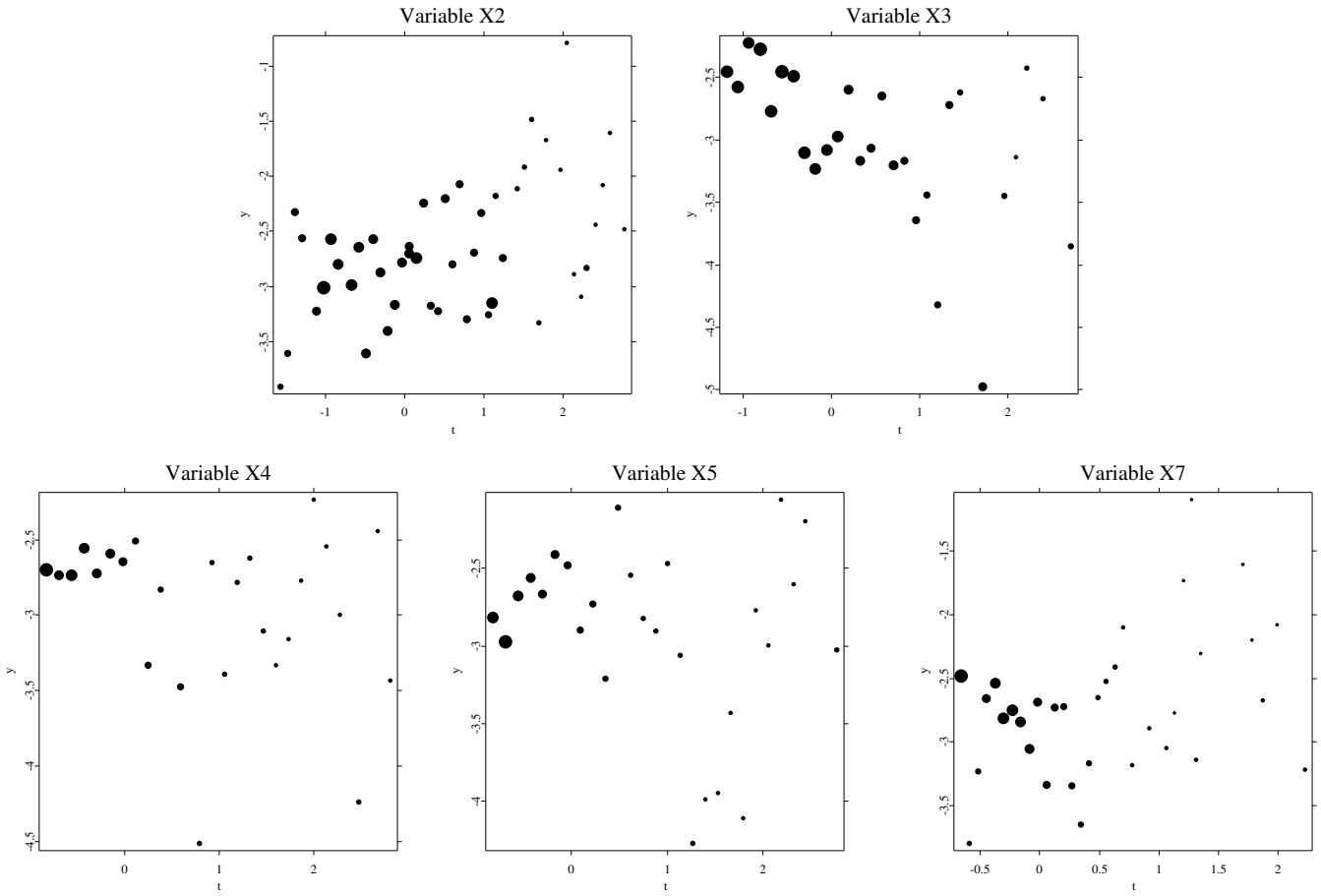



Abbildung 11. Marginale Abhängigkeiten, Variablen X2 bis X5, X7.

# Semiparametrisches Kredit Scoring

Partiell lineares Logit-Modell

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

wobei

- $F(\bullet)$ bekannte (Link-)Funktion,
  hier wie zuvor $F(\bullet) = \frac{1}{1+e^{-\bullet}}$
- $m(\bullet)$ unbekannte glatte Funktion
- $\beta$ unbekannter Parametervektor

# Maximum–Likelihood im GLM

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

- Maximierung der (Log-)Likelihoodfunktion

$$\begin{aligned} \ell(Y,\mu) &= \sum_i \ell_i(Y_i,\mu_i) \\ &= \sum_i Y_i \log(\mu_i) + (1-Y_i)\log(1-\mu_i) \end{aligned}$$

  bzw. Minimierung der Devianz

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

- Algorithmus: "iteratively reweighted least squares"

Referenz: McCullagh & Nelder (1989)

## Semiparametrisches Maximum–Likelihood

GPLM ("generalized partial linear model")

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

- Kombination von klassischem und geglättetem (Log-)Likelihood
- für $\beta$:
  "iteratively reweighted least squares" +modifizierte Designmatrix
- für $m(\bullet)$:
  Nadaraya-Watson (oder anderer) Glätter

# Schätzung im GPLM

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

- $\widehat{\beta}$ kann bei bekanntem $m$ geschätzt werden (parametrische Methode, gewichtete KQS),
- $\widehat{m}$ kann bei bekanntem $\beta$ geschätzt werden (nichtparametrische Methode, Nadaraya-Watson-Typ)

**Schätzverfahren**

- Profile Likelihood, (verallgemeinerter) Speckman-Schätzer
- Backfitting, modified Backfitting

Referenzen: Severini & Staniswalis (1994), Speckman (1988), Hastie & Tibshirani (1990)

**Algorithmus** (verallgemeinerter Speckman-Schätzer)

- *parametrischer Teil*

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

- *nichtparametrischer Teil*

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

mit

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

---

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

## Schätzmatrix

Das Updaten des Index $\mathcal{X}\beta + m$ kann durch eine lineare Schätzmatrix ausgedrückt werden:

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

mit

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

# $\chi^2$ Test

wenn (bei Konvergenz)

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

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

**Approximative Freiheitsgrade**

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

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

Referenz: Hastie & Tibshirani (1990)

## Anwendung

- nichtparametrische Einbeziehung von Variable $T =$ X5

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

- nichtparametrische Einbeziehung von $T =$ (X4,X5)

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

| | | Nichtparametrisch in | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Logit | X2 | X3 | X4 | X5 | X7 | X4,X5 | X2, X4,X5 |
| Konstante | **-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** |

Tabelle 2. Parametrische Koeffizienten für Variablen X2 bis X9. (Fettgedruckte Werte sind zu 5% signifikant.)

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

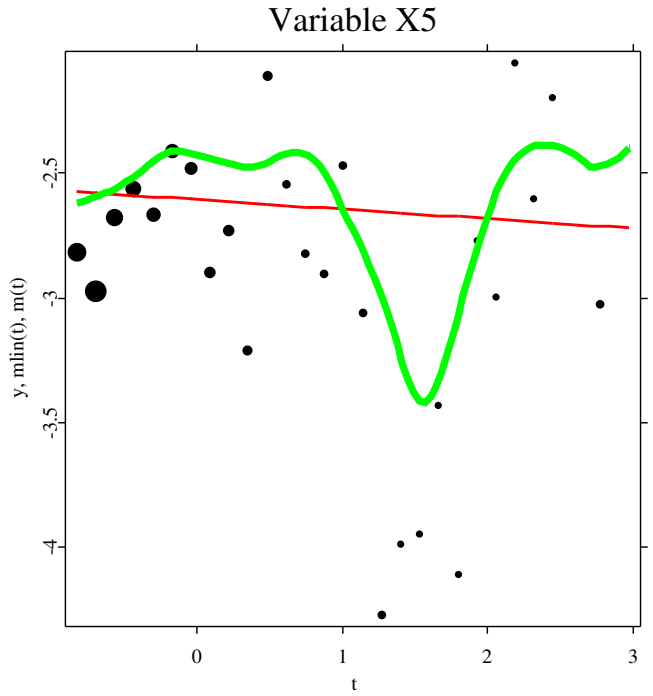



Abbildung 12. Marginale Abhängigkeit, Variable X5. Dickere Punkte entsprechen mehr Beobachtungen. Parametrisches (rot) und GPLM-Logit-Modell (grün).

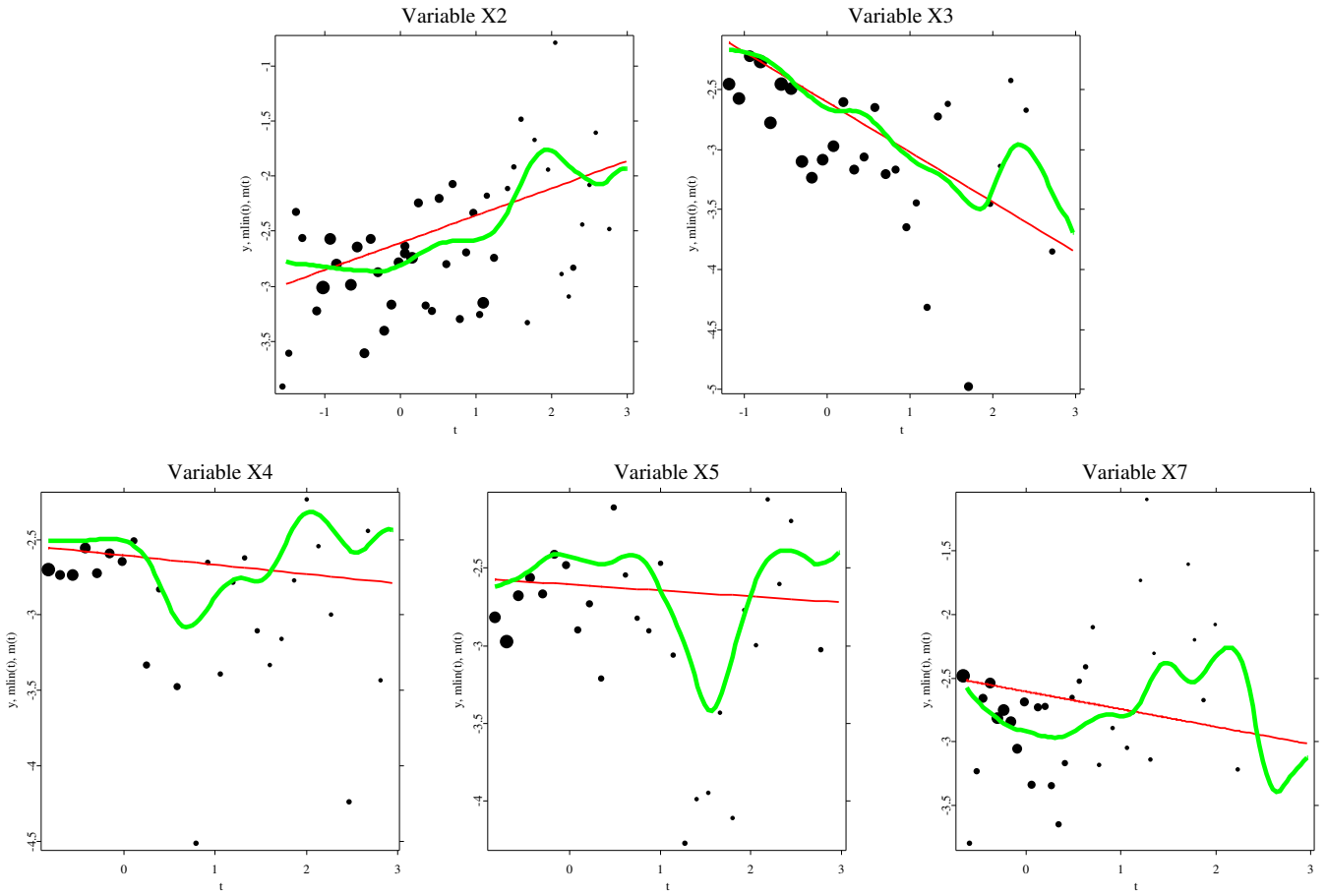



Abbildung 13. Marginale Abhängigkeiten, Variablen X2 bis X5, X7. Parametrisches (rot) und GPLM-Logit-Modell (grün).

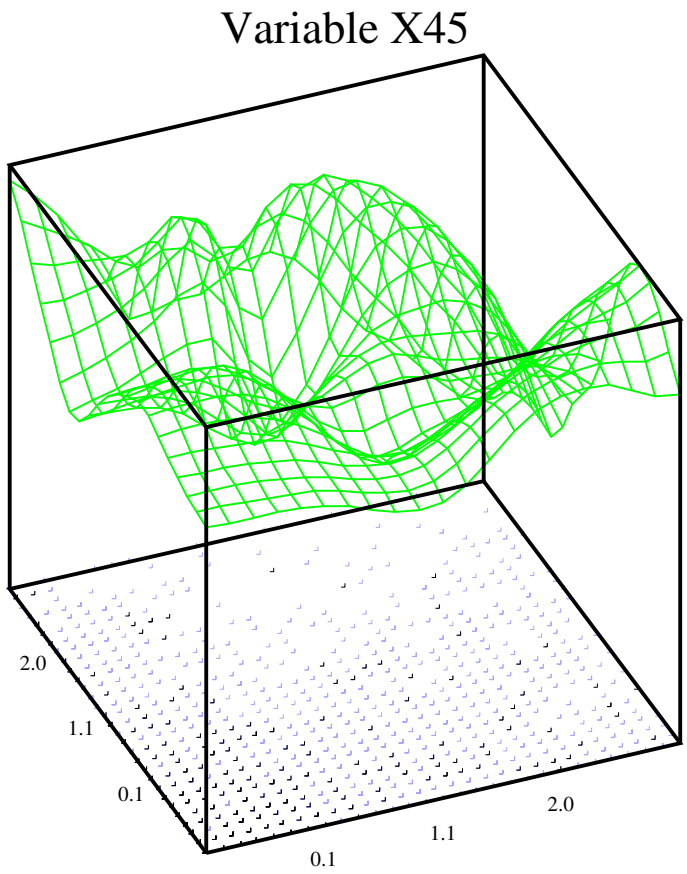



Abbildung 14. Bivariate nichtparametrische Schätzung für Variablen X4, X5.

# Test des Semiparametrischen Modells

| | | Nichtparametrisch in | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Logit | X2 | X3 | X4 | X5 | X7 | X4,X5 | X2, X4,X5 |
| Devianz | 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 |
| $\alpha$ | – | 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% |

Tabelle 3. Statistische Charakteristika der parametrischen und semiparametrischen Logit-Fits.

# Missklassifikation und Performance-Kurven

| Schranke $s$ | Logit | Nichtparametrisch in X2 | X3 | X4 | X5 | X7 | X4,X5 | X2, X4,X5 |
|---|---|---|---|---|---|---|---|---|
| 0.25 | 129 | 133 | 129 | 136 | 130 | 128 | 132 | 130 |
| "Nicht-Ausfall" | 41 | 44 | 40 | 49 | 40 | 40 | 46 | 40 |
| "Ausfall" | 88 | 89 | 89 | 87 | 90 | 80 | 86 | 90 |
| 0.5 | 111 | 110 | 111 | 111 | 110 | 108 | 111 | 110 |
| "Nicht-Ausfall" | 5 | 5 | 5 | 5 | 5 | 2 | 5 | 4 |
| "Ausfall" | 106 | 105 | 106 | 106 | 105 | 106 | 106 | 106 |
| 0.75 | 107 | 107 | 107 | 107 | 107 | 107 | 107 | 107 |
| "Nicht-Ausfall" | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "Ausfall" | 107 | 107 | 107 | 107 | 107 | 107 | 107 | 107 |

Tabelle 4. Missklassifikationen: $Y =$"Ausfall" obwohl $F(\widehat{S}) \leq s$ bzw. $Y =$"Nicht-Ausfall" obwohl $F(\widehat{S}) > s$. Validierungsstichprobe.

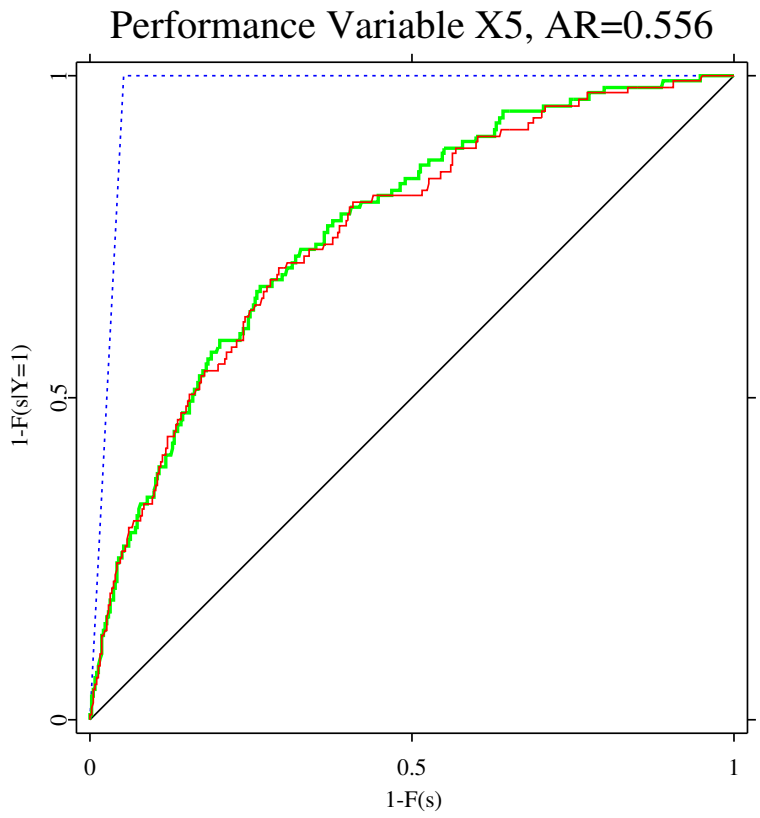



Abbildung 15. Performance-Kurven, parametrisches Logit- (rot) und semiparametrisches Logit-Modell (grün) mit nichtparametrischer Variable X5.

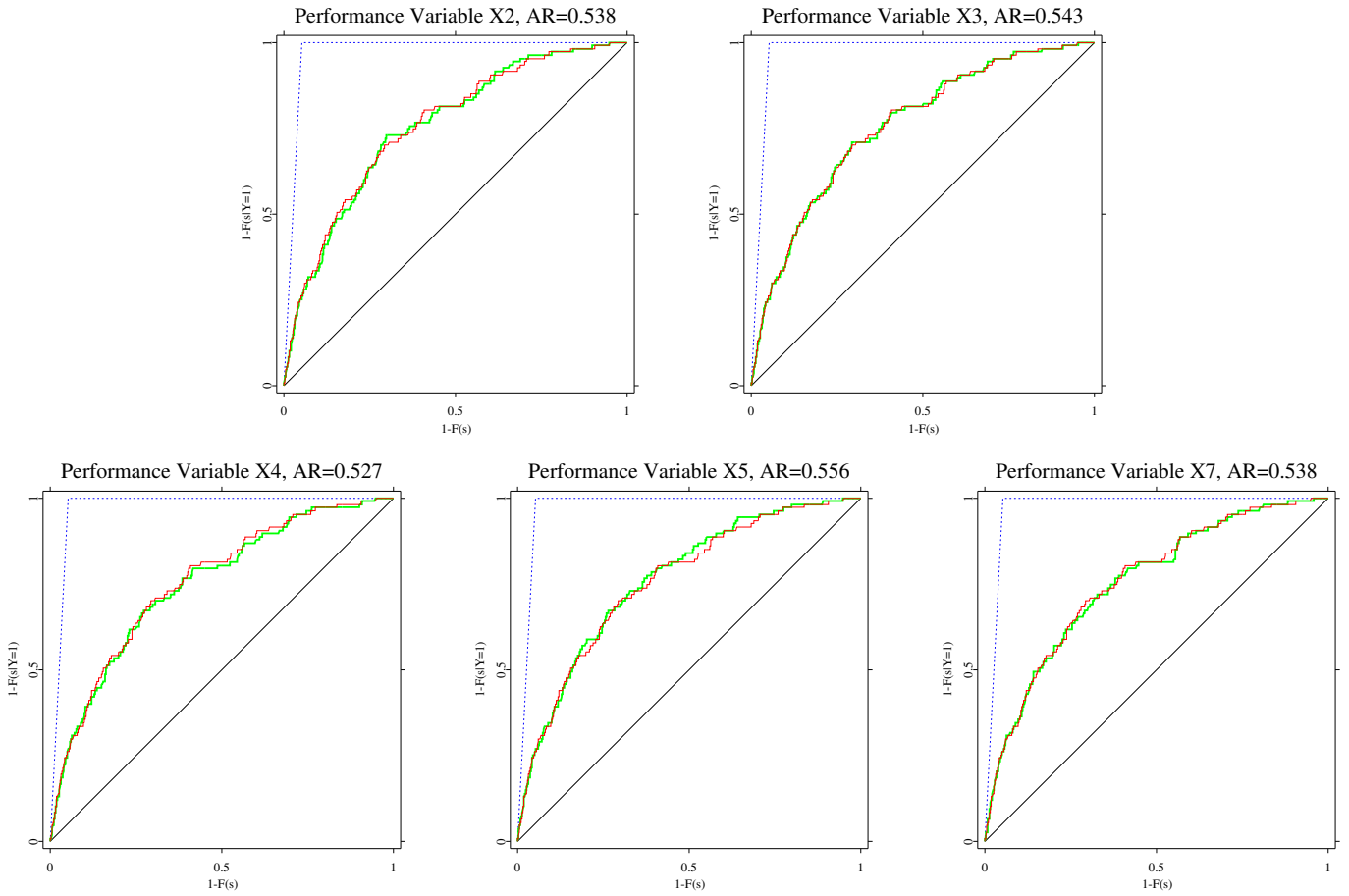



Abbildung 16. Performance-Kurven für nichtparametrische Variablen X2 bis X5, X7 (separat).

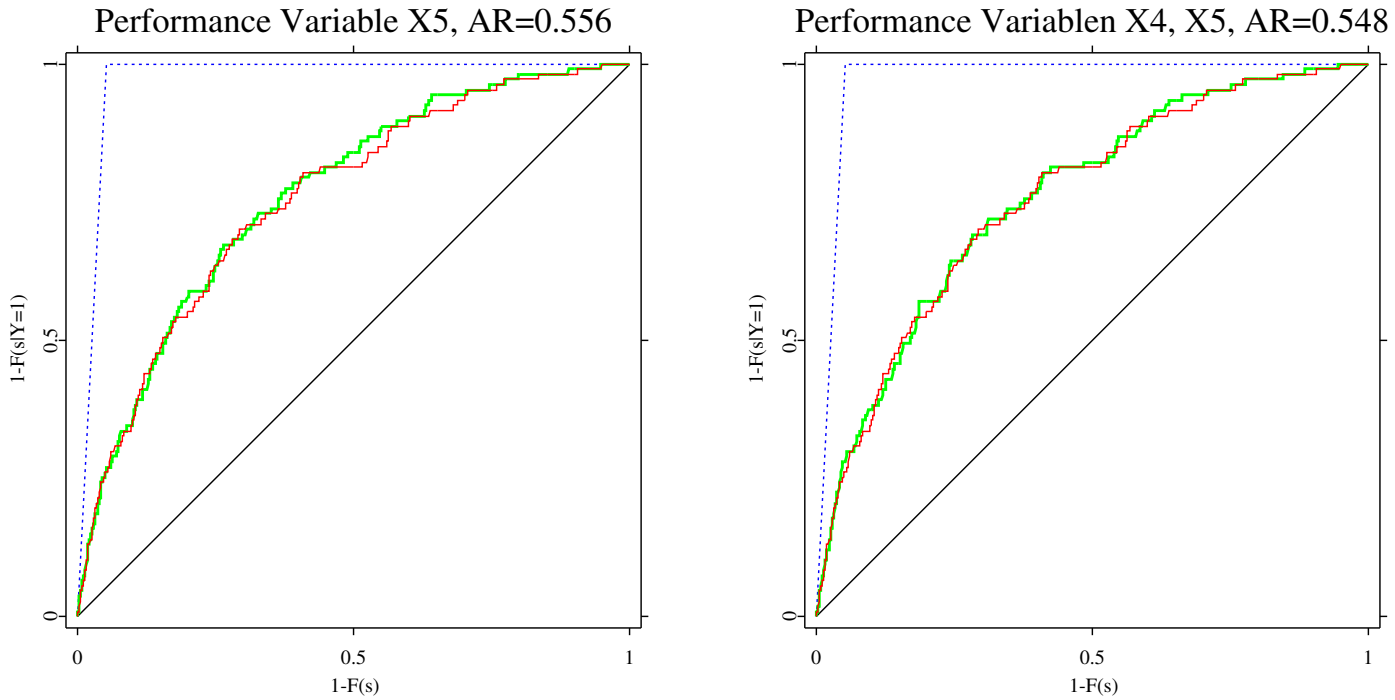



Abbildung 17. Performance-Kurven für nichtparametrische Variablen X5 (links) und X4, X5 (rechts).

# Zusammenfassung

nichtparametrische Komponenten im Modell

- bieten Vorteile nichtparametrischer Diskriminanzanalyse in Kombination mit leichter Intepretierbarkeit, Schätzung von Ausfallwahrscheinlichkeiten
- können explorativ zu Bestimmung von Transformationen von Variablen eingesetzt werden
- können explorativ zu Bestimmung von Interaktionen zwischen Variablen eingesetzt werden
- erlauben Sepzifikationstests auf Gültigkeit der parametrischen Definition von Transformationen

## Ausblick

- Kombination von semiparametrischen Modellen mit Panelansätzen/GEE, geordnete Kategorien
- optimale Bestimmung der Glättungsparameter
- Kombination mit existierenden Verfahren für additive Modelle (gemeinsame Implementierung)