Die Prognose von Spielausgängen in der Fußball-Bundesliga

S. Niermann*

Diskussionspapier 247
ISNN 0949-9962

Zusammenfassung: In dieser Arbeit werden parametrische Modelle zur Prognose von Spielen der Fußball-Bundesliga geschätzt. Dabei werden in einem ersten Schritt die Spielstärken der beteiligten Mannschaften geschätzt, aus denen zu erwartende Tordifferenzen abgeleitet werden. Hierbei werden verschiedenen Modelle und Schätzverfahren verglichen, wobei ein Hauptaugenmerk auf der robusten Schätzung der Modelle liegt. In einem zweiten Schritt werden mit einem Bayesianischen Ansatz die Wahrscheinlichkeiten für die möglichen Spielausgänge (Heimsieg, Unentschieden und Auswärtssieg) geschätzt.

Abstract: In this paper models for the prediction of matches in the German Soccer Bundesliga are estimated. In a first step the expected difference of the number of goals is estimated on the basis of estimated abilities of the teams. Here, a focus lies on the necessity of robust parameter estimation. In a second step the probabilities for home team wins/losses and ties are estimated using a Bayesian updating procedure.

*Adresse des Autors:
Institut für Quantitative Wirtschaftsforschung
Königsworther Platz 1
30167 Hannover
Fax-Nr. 0049 511 762-3923
niermann@mbox.iqw.uni-hannover.de

1 Einleitung

Die wirtschaftliche Bedeutung des Fußballs in Deutschland nimmt stetig zu. Dies äußert sich auf der einen Seite in explodierenden Spielergehältern und Ablösesummen, andererseits gewinnt auch das *Drumherum*, wie Merchandising, Marketing und nicht zuletzt der Markt für Sportwetten an Bedeutung.

Bei der Abgabe von Wetten ist es hilfreich und bei der Erstellung von Quoten ist es notwendig, Aussagen über die Wahrscheinlichkeiten für die verschiedenen Spielausgänge – bei Fußballwetten sind das ein Heimsieg (1), ein Unentschieden (0) und ein Auswärtssieg (2) – treffen zu können. In diesem Papier werden solche Prognosemodelle entwickelt und angewandt auf die Bundesligasaison 1998/99.

Ein rational handelndes, risikoneutrales Individuum wird eine Wette auf einen Spielausgang A abgeben, falls die Quote $q(A)$ größer ist als der Kehrwert der Wahrscheinlichkeit $P(A)$. Bei einer Quote $q(A)$ erhält man pro eingesetzter Geldeinheit $q(A)$ Geldeinheiten zurück, sofern der Spielausgang A eintritt.

Eine ausführliche Darstellung der wirtschaftlichen Bedeutung und Unvollkommenheiten auf dem Wettmarkt findet sich bei Pope und Peel (1989).

Bei der Konstruktion eines Modells, anhand dessen die Ergebnisse in der Fußball-Bundesliga geschätzt werden sollen, werden zwei parametrische Ansätze und verschiedene Schätzansätze verglichen.

Die Prognose des Spielausgangs vollzieht sich in zwei Schritten. Zunächst wird ein Modell spezifiziert, das Aussagen über die zu erwartende Tordifferenz in einer Begegnung liefert. In einem weiteren Schritt werden mit einem bayesianischen Ansatz hieraus die Wahrscheinlichkeiten für einen Heimsieg, ein Unentschieden und einen Auswärtserfolg geschätzt.

2 Modelle für die Erklärung der erwarteten Tordifferenz

Von Interesse ist ein Modell, welches bei einem Spiel der Heimmannschaft i gegen die Auswärtsmannschaft j eine Schätzung $\hat{d}_{ij}$ für die Tordifferenz d_{ij} liefert. Bei einem 3:1 ist diese Tordifferenz 2, bei einem 0:4 ist sie -4, usw.

Diese Tordifferenz hängt natürlich von vielen Faktoren ab: Von den Spielstärken der beteiligten Mannschaften, der Zahl der verletzten Spieler, der Regenerationsphase seit dem letzten Spiel, der Unterstützung durch die Zuschauer, psychologische Effekte etwa nach einer Sieg- oder Niederlagenserie, und – das ist schließlich das Salz in der Suppe - vom Glück.

Die in diesem Papier verwendeten Modelle berücksichtigen zunächst lediglich die Spielstärken der beteiligten Mannschaften, abgebildet durch zu schätzende Parameter oder im einfachsten Fall durch den Tabellenplatz, den Heimvorteil und spezifische Heim- oder Auswärtsstärken der einzelnen Mannschaften. Wünschenswert ist die Integration weiterer Variablen in das Modell, die zu einem späteren Zeitpunkt vorgenommen werden soll.

Bei diesen Modellen wird in einem ersten Schritt überprüft, wie gut die Anpassungsgüte des Modells ist, falls das Modell für alle 306 Spiele einer Bundesligasaison geschätzt wird.
Da die verwendeten Modelle aber zum Teil eine recht hohe Zahl an Parametern aufweisen,

ist es sinnvoll die geschätzten Modelle zu validieren. Dies geschieht, indem die Parameter des Modells für die Hinserie geschätzt werden und dann zu Prognosezwecken auf die Rückserie angewandt wird.

- Modell 0:

  Bei diesem Modell werden die erwarteten Tordifferenzen ausschließlich auf der Grundlage der Tabellenplätze der beteiligten Mannschaften bestimmt:

$$d_{ij} = h + \beta \cdot (Tabellenplatz_{heim} - Tabellenplatz_{ausw}) + u_{ij} \tag{1}$$

  Hierbei ergibt sich die handliche Regel[1]: Die erwartete Tordifferenz ist gleich 0.5 minus einem Zehntel der Differenz der Tabellenplätze der Heim- und der Auswärtsmannschaft. Wenn also beispielsweise der Tabellendritte gegen den Tabellensiebten spielt, ist mit diesem Modell eine Tordifferenz von 0.9 zu erwarten, wenn der Tabellenachtzehnte gegen den Tabellensechsten spielt, ist eine Tordifferenz von -0.6 zu erwarten – im Mittel wird die Auswärtsmannschaft 0.6 Tore mehr schießen.

  Dieses Modell liefert eine einfache Basis für Wahrscheinlichkeitsaussagen. Allerdings kann man sich des Eindrucks nicht erwehren, dass es noch etwas besseres geben muss.

- **Modell 1:**

  Jeder der 18 Mannschaften wird ein Parameter β_i zugeordnet, der die Spielstärke der Mannschaft i abbildet. Zusätzlich wird ein Parameter für den Heimvorteil eingeführt.

$$d_{ij} = h + \beta_i - \beta_j + u_{ij} \tag{2}$$

  Hierbei bezeichnet u_{ij} wie üblich eine Störgröße. Eine gesamte Saison ergibt sich, wenn i und j die Werte 1 bis 18 durchlaufen, jedoch unterschiedlich sein müssen. Das sind dann 306 Spiele pro Saison.

  Die Parameter dieses Modells werden so gewählt, daß die geschätzten Tordifferenzen von den tatsächlichen Tordifferenzen möglichst wenig – nach einem weiter hinten spezifizierten Kriterium – abweichen. Falls als Schätzmethode die Methode der kleinsten Quadrate gewählt wird, entspricht das Modell dem von Clarke und Norman (1995), in dem ebenfalls die Tordifferenz die zu erklärende Variable ist und in dem ein Kleinst-Quadrate-Ansatz verwendet wird. Allerdings zielt die Arbeit von Clarke und Norman (1995) nicht auf die Bestimmung von Wahrscheinlichkeiten für Heimsiege, Unentschieden und Auswärtsiege ab.

- **Modell 2:**

  Bei Modell 2 wird berücksichtigt, daß manche Mannschaften – abgesehen von dem durch den Parameter h abgebildeten Heimvorteil – besondere Heim- oder Auswärtsstärken aufweisen. Hier ergibt sich ein Unterschied zu der ebenfalls bei der Prognose von Fußballergebnissen zum Einsatz kommenden Poisson-Regression.[2] Bei diesen Modellen wird davon ausgegangen, dass die Zahl der Tore der Heimmannschaft und die

[1] Bei den durchgeführten Schätzungen ergab sich diese Regel sowohl bei der LS- als auch bei der L1-Methode.

[2] Vgl. etwa Lee (1997) oder Dixon und Coles (1997).

Zahl der Tore der Auswärtsmannschaft jeweils unabhängig Poissonverteilt mit den Parametern λ bzw. μ sind. Der Parameter λ der Heimmannschaft wird dann üblicherweise erklärt durch die Offensivstärke der Heimmannschaft und die Defensivstärke der Auswärtsmannschaft, der Parameter μ der Auswärtsmannschaft durch die Offensivstärke der Auswärtsmannschaft und die Defensivstärke der Heimmannschaft. Die Parameter lassen sich dann Maximum-Likelihood-Methode schätzen, womit dann für jedes Spiel das λ und das μ geschätzt werden können, woraus sich dann wiederum die Wahrscheinlichkeiten für einen Heimsieg, ein Unentschieden und einen Auswärtssieg berechnen lassen.

Bei diesen Modellen wird also für jede Mannschaft nach Offensivstärke und Defensivstärke differenziert. Dies erscheint mir aber nur zielführend zu sein, wenn der Fokus des Modells auf der Prognose der Ergebnisse und nicht auf der Prognose der Tendenz liegt. Interessieren wir uns also dafür, ob Mannschaft i gegen Mannschaft j eher 3:2 oder 1:0 gewinnen wird, ist die Modellierung der Offensiv- und der Defensivstärke hilfreich. Interessieren wir uns jedoch lediglich für die erwartete Tordifferenz, scheint die Modellierung von Heim- und Auswärtsstärke zielführender zu sein.

In diesem Modell werden jeder Mannschaft 2 Parameter zugeordnet: Ein Parameter h_i für die individuelle Heimstärke und ein Parameter a_i für die individuelle Auswärtsstärke.

$$d_{ij} = h_i - a_j + u_{ij} \tag{3}$$

Ein genereller Parameter für der Heimvorteil wird dann nicht mehr benötigt.

Das Modell 2 enthält 36 Parameter im Vergleich zu den 19 Parametern des ersten Modells. Damit wird die ex post - Anpassung des zweiten Modells immer besser sein als diejenige des ersten Modells. Dennoch können einfache Modelle komplexen Modellen in der Prognosegüte überlegen sein. Dieses Ergebnis hat im Zusammenhang mit der Theorie neuronaler Netze unter dem Begriff *Overfitting* an Aktualität gewonnen. Insofern ist a priori nicht klar, ob Modell 2 tatsächlich eine bessere Prognosequalität aufweisen wird als Modell 1.

Modell 2 wird bei Prognosen lediglich dann bessere Ergebnisse erzielen, wenn einige Mannschaften besonders heim- oder auswärtsstark sind.

3 Die Schätzung der Parameter

Die Grundlage für die Schätzungen bilden die Spiele der Saison 1998/99. Alle Berechnungen und die dafür erforderlichen Funktionen wurden mit dem Programmpaket S-PLUS (Version 3.4) durchgeführt bzw. programmiert.

Die Schätzung der Parameter erfolgt auf drei Arten:

- Nach der Methode der kleinsten Quadrate,
- Minimierung der absoluten Abstände,
- Transformation der abhängigen Variable.

3.1 Modell 1

3.1.1 Schätzung nach der Methode der kleinsten Quadrate.

Hierbei wird die Zielfunktion

$$Z_1 = \sum_{i=1}^{18} \sum_{j \neq i} (y_{ij} - \beta_i + \beta_j - h)^2 \tag{4}$$

bezüglich der Parameter $\beta_1, \ldots \beta_{18}$ und h minimiert.

Bei der Schätzung der Parameter tritt jedoch das Problem auf, daß die Parameter nicht identifizierbar sind. Dies kann man sich leicht klar machen. Wenn zu jedem geschätzen Parameterwert $\hat{\beta}_1, \ldots \hat{\beta}_{18}$ ein konstanter Wert c addiert wird, ändern sich die geschätzten Werte nicht. Insofern ist eine Normierung erforderlich, die bei unseren Schätzungen darin besteht, den Schätzwert $\hat{\beta}_1$ auf Null zu setzen. Die übrigen Parameterschätzer β_i können dann als relative Spielstärke der Mannschaft i zur Mannschaft 1 interpretiert werden.

Mit $\hat{\beta}_1 = 0$ lässt sich das Modell wie folgt matriziell darstellen:

$$\mathbf{d} = \mathbf{X}\beta + \mathbf{u}$$

$$\begin{array}{c} \\ \text{1 gegen 2} \\ \text{1 gegen 3} \\ \vdots \\ \text{1 gegen 18} \\ \text{2 gegen 1} \\ \vdots \\ \text{18 gegen 17} \end{array}
\begin{pmatrix} d_{1,2} \\ d_{1,3} \\ \vdots \\ d_{1,18} \\ d_{2,1} \\ \vdots \\ d_{18,17} \end{pmatrix}
=
\left(\begin{array}{c|cccccc}
h & \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_{18} \\
\hline
1 & 1 & \mathbf{-1} & \mathbf{0} & \cdots & \mathbf{0} \\
1 & 1 & \mathbf{0} & \mathbf{-1} & \cdots & \mathbf{0} \\
 & & \vdots & \vdots & & \\
1 & 1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{-1} \\
1 & -1 & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\
 & & \vdots & \vdots & & \\
1 & 0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}
\end{array}\right)
\begin{pmatrix} h \\ \cancel{\beta_1} \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{18} \end{pmatrix}
+
\begin{pmatrix} U_{12} \\ U_{13} \\ \vdots \\ U_{1,18} \\ U_{21} \\ \vdots \\ U_{18,17} \end{pmatrix}$$

In statistischen und ökonometrischen Programmpaketen wird die *Designmatrix* häufig ohne die erste Spalte für das Absolutglied - in diesem Fall der Heimvorteil - eingegeben. Bei dem S-PLUS-Befehl *lsfit* ist das genauso. Demnach werden dieser fett gesetzte Teil der Matrix $\mathbf{X}$ als Designmatrix und die beobachteten Tordifferenzen y an den Befehl *lsfit(X,y)* übergeben. Die hierfür benötigte Matrix $\mathbf{X}$ wird erzeugt durch die S-PLUS-Funktion *X.design* mit dem Parameter k: Anzahl der Mannschaften. Hierbei ist zu beachten, dass die Spiele in einer bestimmten Reihenfolge aufgeführt sein müssen:

Zuerst alle Heimspiele der Mannschaft 1, dann alle Heimspiele der Mannschaft 2, usw. Diese Blöcke von Spielen müssen wiederum so angeordnet sein, dass die Auswärtsmannschaften ebenfalls geordnet (die mit dem kleinsten Index zuerst) erscheinen. Diese Design-Matrix $\mathbf{X}$ lässt sich mit S-PLUS mit der Funktion *X.design* in Verbindung mit der Funktion *X.soccer* erzeugen. Diese Funktionen sind im Anhang angegeben.

Beispiel: Betrachtet wird eine Liga mit 4 Mannschaften. Dann liefert der Befehl X.design(4) die folgende Matrix:

Spiel	β_2	β_3	β_4
1 gegen 2	*-1*	*0*	*0*
1 gegen 3	*0*	*-1*	*0*
1 gegen 4	*0*	*0*	*-1*
2 gegen 1	*1*	*0*	*0*
2 gegen 3	*1*	*-1*	*0*
2 gegen 4	*1*	*0*	*-1*
3 gegen 1	*0*	*1*	*0*
3 gegen 2	*-1*	*1*	*0*
3 gegen 4	*0*	*1*	*-1*
4 gegen 1	*0*	*0*	*1*
4 gegen 2	*-1*	*0*	*1*
4 gegen 3	*0*	*-1*	*1*

Auf diese Weise erhält man die folgenden Schätzwerte für die Spielstärken der 18 Bundesligavereine der Spielzeit 1998/99. Der Schätzwert $\hat{h}$ ist so zu interpretieren, dass für den Fall, dass gleichstarke Mannschaften aufeinandertreffen, zu erwarten ist, dass die Heimmannschaft im Durchschnitt 0.55 Tore mehr erzielen wird.

Index	Mannschaft	Spielstärke	Index	Mannschaft	Spielstärke
1	1. FC Kaiserslautern	0	10	Hamburger SV	-0.083
2	1. FC Nürnberg	-0.389	11	Hansa Rostock	-0.361
3	1860 München	-0.306	12	Hertha BSC	0.639
4	Bayer Leverkusen	0.750	13	MSV Duisburg	-0.028
5	Bayern München	1.220	14	SC Freiburg	-0.333
6	Bor. Dortmund	0.280	15	VfB Stuttgart	-0.305
7	Bor. M'gladbach	-1.167	16	VfL Bochum	-0.806
8	Eintr. Frankfurt	-0.389	17	VfL Wolfsburg	0.028
9	FC Schalke 04	-0.472	18	Werder Bremen	-0.278
	Heimvorteil	0.549			

Tabelle 1: Geschätzte Spielstärken der Bundesligavereine (Modell 1 und LS)

Spielt also beispielsweise Werder Bremen in Bremen gegen den VfL Wolfsburg, ergibt sich eine prognostizierte Tordifferenz

$$\hat{y}_{ij} = -0.278 - 0.028 + 0.549 = 0.243.$$

Die Prognose des Modells lautet damit, daß Bremen 0.243 Tore mehr schießt als Wolfsburg.

Werden diese prognostizierten Werte den realisierten gegenübergestellt, ergibt sich die folgende Abbildung.

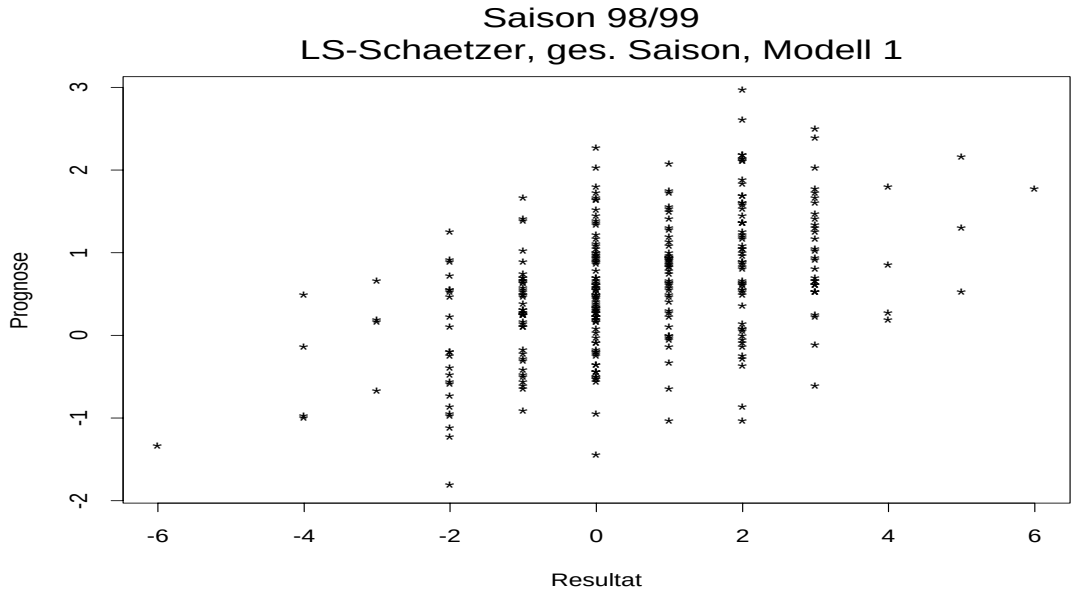

Abbildung 1: Anpassung des Modells (Modell1, LS)

Auf diese Weise lassen sich auch die größten Überraschungen der Saison 98/99 bestimmen. Überraschungen sind diejenigen Spiele, bei denen sich die größten Residuen ergeben haben.

Die größten Überraschungen stellten für das oben spezifizierte Modell die folgenden Spiele dar.

1.FC Nürnberg	-	1860 München	1:5,	
MSV Duisburg	-	VfL Wolfsburg	6:1	und
Bor. M´gladbach	-	Bayer Leverkusen	2:8.	

An dieser Stelle offenbart sich schon ein Problem der Kleinst-Quadrate-Schätzung im Kontext der Prognose von Fußballspielen. Bei Fußballspielen ist häufig die sog. Tendenz (2 – bei Auswärtssiegen, 0 bei Unentschieden und 1 bei Heimsiegen) von primärem Interesse. Bei den drei oben genannten Überraschungen überrascht uns aber in erster Linie die Höhe des Ergebnisses.

Bei der Beurteilung der Anpassungsgüte soll nun der Anteil der korrekt prognostizierten Tendenzen verwendet werden. Hierzu wird ein einfaches Rangverfahren verwendet. Der Anteil der Heimsiege, Unentschieden und der Auswärtssiege läßt sich empirisch bestimmen. In der Saison 98/99 gab es beispielsweise 75 Auswärtssiege (24.5%), 87 Unentschieden (28.4%) und 144 Heimsiege (47.1%). Bei der 24.5% kleinsten Prognosewerten wird damit ein Auswärtssieg prognostiziert, bei der 47.1% größten Prognosewerten ein Heimsieg und bei den übrigen ein Unentschieden.

Damit ergibt sich die folgende Tabelle:

	Prognose			**Treffer** (in Prozent)
	-1	0	1	
bei Auswärtssiegen	31	27	17	41.3
bei Unentschieden	22	32	33	36.8
bei Heimsiegen	22	28	94	65.3

Insgesamt würden mit dem Modell 51.3% der Spiele von der Tendenz her richtig erkannt. Die Prognosegüte des Modells ist mit diesem Wert allerdings noch nicht quantifiziert, da die Parameterschätzung und die Beurteilung der Anpassungsgüte auf derselben Menge von Objekten (hier den 306 Spielen der Saison 98/99) vorgenommen wurden.

Deshalb wird in einem zweiten Schritt die folgende Vorgehensweise gewählt. Die Parameter des Modells werden lediglich auf der Basis der Spiele der Hinrunde geschätzt und dann verwendet, um die Spiele der zweiten Saisonhälfte zu progostizieren. Auf diese Weise erhält man einen Indikator für die Prognosegüte des Modells. Eine solche Prognose wird auch ex ante Prognose bezeichnet.

Index	Mannschaft	Spielstärke	Index	Mannschaft	Spielstärke
1	1. FC Kaiserslautern	0	10	Hamburger SV	-0.167
2	1. FC Nürnberg	-0.608	11	Hansa Rostock	-0.5
3	1860 München	0.448	12	Hertha BSC	0.448
4	Bayer Leverkusen	1.167	13	MSV Duisburg	-0.496
5	Bayern München	1.444	14	SC Freiburg	-0.167
6	Bor. Dortmund	0.333	15	VfB Stuttgart	0.059
7	Bor. M'gladbach	-1.274	16	VfL Bochum	-0.608
8	Eintr. Frankfurt	-0.441	17	VfL Wolfsburg	0.444
9	FC Schalke 04	-0.552	18	Werder Bremen	0
	Heimvorteil	0.532			

Tabelle 2: Geschätzte Spielstärken der Bundesligamannschaften (Modell1 , LS, nur Hinserie)

Bei dieser Vorgehensweise ergibt sich die folgende Tabelle:

	Prognose			**Treffer** (in Prozent)
	-1	0	1	
bei Auswärtssiegen	14	15	9	36.8
bei Unentschieden	14	7	22	16.3
bei Heimsiegen	14	18	40	55.6

Insgesamt wurden damit 39.9% aller Rückrundenspiele auf der Basis der Schätzergebnisse der Hinrunde korrekt prognostiziert.

3.1.2 Die Minimierung der absoluten Abstände

Da *Kantersiege* bei der Kleinst-Quadrate-Regression einen (zu?) großen Einfluß auf die Parameterschätzer aufweisen, wird die Minimierung der absoluten Abstände betrachtet. Diese Vorgehensweise wurde im Zusammenhang mit der Prognose von Fußballspielen von Basset (1997) vorgeschlagen.

Bei der Minimierung der Summe der absoluten Abweichungen – man sagt auch, dass die L1-Norm verwendet wird –, ergibt sich die folgende zu minimierende Zielfunktion:

$$Z_2 = \sum_{i=1}^{18} \sum_{j \neq i} |y_{ij} - \beta_i + \beta_j - h| \tag{5}$$

Hierbei ergeben sich die nachfolgenden Parameterschätzer.

Index	Mannschaft	Spielstärke	Index	Mannschaft	Spielstärke
1	1. FC Kaiserslautern	0	10	Hamburger SV	0
2	1. FC Nürnberg	-0.5	11	Hansa Rostock	-1
3	1860 München	-0.5	12	Hertha BSC	0.5
4	Bayer Leverkusen	0.5	13	MSV Duisburg	0
5	Bayern München	1	14	SC Freiburg	-0.5
6	Bor. Dortmund	0.5	15	VfB Stuttgart	-0.5
7	Bor. M'gladbach	-1.5	16	VfL Bochum	-1
8	Eintr. Frankfurt	-0.5	17	VfL Wolfsburg	0
9	FC Schalke 04	-1	18	Werder Bremen	-0.5
	Heimvorteil	0.5			

Tabelle 3: Geschätzte Spielstärken der Bundesligavereine (Modell 1 und L1)

Bei einem Spiel *Werder Bremen* gegen den *VfL Wolfsburg* lautet der Schätzwert für die Tordifferenz nun

$$\hat{d}_{ij} = \hat{\beta}_{18} - \hat{\beta}_{17} + \hat{h} = -0.5 - 0 + 0.5 = 0.$$

Entsprechend dieser Schätzergebnisse war der Ausgang der Begegnung *MSV Duisburg - VfL Wolfsburg* (6:1) die größte Überraschung der gesamten Saison.

Auch bei diesem Schätzansatz kann die erwartete Tordifferenz der realisierten Tordifferenz gegenübergestellt werden.

Wird nun entsprechend des oben dargestellten Rankingverfahrens die erwartete Tordifferenz in eine Prognose über Heimsieg/Auswärtssieg oder unentschieden transformiert, ergibt sich die folgende Tabelle:

	Prognose			**Treffer** (in Prozent)
	-1	0	1	
bei Auswärtssiegen	24	21	20	45.3
bei Unentschieden	21	34	31	39.5
bei Heimsiegen	20	31	94	64.8

Damit konnten 52.9% der Spielausgänge bezüglich ihrer Tendenzen korrekt prognostiziert werden.

Zur Beurteilung der Prognosequalität des Modells wurde nun wieder überprüft, wie sich das in der Hinserie geschätzte Modell in der Rückserie bewährt.

Wird auch für diesen Fall die Situation einer *echten Prognoseerstellung* betrachtet, ergibt sich, dass 44.4% aller Rückrundenspiele auf der Basis der Schätzergebnisse der Hinrunde korrekt prognostiziert. Die Schätzung, bei der die Summe der absoluten Abweichungen minimiert wird, liefert also bei Modell 1 sowohl bezüglich der Anpassung als auch bezüglich der Prognosequalität bessere Ergebnisse als die LS-Schätzung.

	Prognose			**Treffer** (in Prozent)
	-1	0	1	
bei Auswärtssiegen	15	13	10	39.5
bei Unentschieden	15	8	19	19.0
bei Heimsiegen	14	14	45	61.6

3.1.3 Transformation der abhängigen Variable

In diesem Abschnitt werden 3 weitere Modelle betrachtet. Diese sind zwar prinzipiell von Modell 1 zu unterscheiden, wegen ihrer formalen Ähnlichkeit mit diesem Modell werden sie aber hier als Schätzmethoden für Modell 1 dargestellt. Bei der Schätzung mit der L1-Norm wird der Abstand weit von der Regressionsgerade entfernt liegender Punkte weniger stark – nämlich nicht mehr quadratisch, sondern nur noch absolut – gewichtet. Alternativ dazu ist eine Vorgehensweise denkbar, bei der die zu erklärende Variable vor der Schätzung einer Transformation unterzogen wird. Bei einer solchen Transformation werden *Kantersiege* in nicht ganz so extreme y-Werte transformiert. Natürlich sollte die gewählte Transformation monoton wachsend sein.

$$t(d_{ij}) = \beta_i - \beta_j + h + u_{ij} \tag{6}$$

Hierbei werden 2 Transformationen betrachtet.

- $t_1(d_{ij}) = \Phi(d_{ij})$: Bei dieser Transformation soll ähnlich wie bei der Verwendung der L1-Norm der Einfluss von Spielen mit großen Tordifferenzen reduziert werden.
- $t_2(d_{ij}) = tendenz(d_{ij}) = \begin{cases} 1, & \text{falls Heimsieg,} \\ 0, & \text{falls Unentschieden,} \\ -1, & \text{falls Auswärtssieg} \end{cases}$
- $t_3(d_{ij}) = \begin{cases} 2, & \text{falls } y \geq 2, \\ y, & \text{falls } -2 \leq y \leq 2, \\ -2, & \text{falls } y \leq -2 \end{cases}$

  Die Transformationen t_1 und t_3 stellen einen Kompromiss dar, bei dem mehr Informationen als die bloße Punkteverteilung (Tendenz) eingehen, sehr hohe Ergebnisse aber auf ein *normales Maß zurückgestutzt werden.*

Für die Transformation t_1 ergibt sich, dass über die gesamte Saison gesehen, 51% aller Spiele richtig prognostiziert werden, bei der ex ante Prognose konnten immerhin 45.1% aller Spiele korrekt vorhergesat werden.

Bei der Transformation t_2 ergaben sich die Prozentwerte 52.3% (Anpassung) und 42.5% (ex ante Prognose).

Bei der Transformation t_3 ergaben sich die Prozentwerte 51.6% (Anpassung) und 45.8% (ex ante Prognose).

Festzuhalten bleibt, dass die Kleinst-Quadrate-Methode offensichtlich suboptimale Ergebnisse liefert, wenn primär die *Punkteverteilung* von Interesse ist. Wenn lediglich die Tendenz unter Vernachlässigung der Tordifferenz verwendet wird, erhält man ebenfalls ein Modell mit suboptimaler Prognosegüte.

3.2 Modell 2

In diesem Modell werden jeder Mannschaft 2 Parameter zugeordnet: Ein Parameter h_i für die individuelle Heimstärke und ein Parameter a_i für die individuelle Auswärtsstärke.

$$d_{ij} = h_i - a_j + u_{ij} \tag{7}$$

Ein genereller Parameter für der Heimvorteil wird dann nicht mehr benötigt.

Wird eine zu der bei Modell 1 gewählten Vorgehensweise analoge Vorgehensweise gewählt, lassen sich die Parameter für die Heim- bzw. Auswärtsstärke einer Mannschaft ebenfalls nach der Methode der kleinsten Quadrate, nach der L1-Norm und nach der Transformation der zu erklärenden Variable minimieren.

Die matrizielle Darstellun ist ähnlich der matriziellen Darstellung von Modell 1. Die Design-matrix X lässt sich nun erzeugen durch die Funktion $X2.design(k)$ in Verbindung mit der Funktion $X2.soccer$ und $X1.soccer$ erzeugen. Hierbei steht k wieder für die Zahl der Mannschaften.

Die Spiele müssen in der oben beschriebenen Sortierung aufgelistet sein. In diesem Fall wird ein Modell ohne Absolutglied geschätzt.

3.2.1 Minimierung nach der Methode der kleinsten Quadrate

Bei der Schätzung der Parameter nach der Methode der Kleinsten Quadrate erhält man die folgenden Schätzergebnisse.

	Mannschaft	Heimstärke	Auswärtsstärke
1	1. FC Kaiserslautern	0	-1.01
2	1. FC Nürnberg	-0.89	-0.90
3	1860 München	-0.59	-1.03
4	Bayer Leverkusen	0.09	0.40
5	Bayern München	0.97	0.46
6	Bor. Dortmund	0.59	-1.04
7	Bor. M'gladbach	-1.42	-1.92
8	Eintr. Frankfurt	-0.58	-1.21
9	FC Schalke 04	-1.07	-0.89
10	Hamburger SV	-0.24	-0.93
11	Hansa Rostock	-0.39	-1.34
12	Hertha BSC	0.73	-0.46
13	MSV Duisburg	-0.06	-1.00
14	SC Freiburg	-0.96	-0.72
15	VfB Stuttgart	-0.27	-1.34
16	VfL Bochum	-1.27	-1.34
17	VfL Wolfsburg	0.18	-1.14
18	Werder Bremen	-0.97	-0.60

Tabelle 4: Geschätzte Spielstärken der Bundesligavereine (Modell 2 und LS)

Die Parameter sind hier alle in Relation zu der Heimstärke des 1. FC Kaiserslautern zu interpretiern. Angewendet werden können sie analog zu Modell 1. Wenn der FC Schalke 04 gegen Bayern München spielt, ist zu erwarten, dass die Tordifferenz $-1.07 - 0.46 = -1.53$ beträgt.

Aus dieser Tabelle lässt sich auch ableiten, dass Borussia Dortmund in der Bundesligasaison 1998/99 die - relativ zu den auswärts erbrachten Leistungen von Borussia Dortmund - heimstärkste Mannschaft war. Würde man rein hypothetisch davon ausgehen, dass Borussia Dortmund ein Heimspiel gegen Borussia Dortmund absolvieren würde, wäre zu erwarten, dass die Heimmannschaft 1.63 Tore mehr erzielen würde als die Auswärtsmannschaft. Im Gegensatz dazu war Werder Bremen so heimschwach, dass bei einem - wiederum natürlich rein hypothetischen - Spiel, bei dem Werder Bremen als Heimmannschaft gegen Werder Bremen als Auswärtsmannschaft antritt, zu erwarten wäre, dass die Auswärtsmannschaft 0.37 Tore mehr erzielen wird.

Werden nun also für jede Mannschaft zwei Parameter verwendet, wobei das Modell unter Verwendung der Methode der Kleinsten Quadrate geschätzt wird, ergibt sich eine Anpassungsgüte von 56.2% und eine Prognosegüte von 41.2%. Wie erwartet, ergibt sich bei diesem Modell eine bessere Anpassung. Allerdings wurde diese Prognosegüte bereits durch Schätzungen des Modells 1 übertroffen, sofern nicht nach der Methode der Kleinsten Quadrate geschätzt wird.

In der folgenden Tabelle sind die Anpassungs- und Prognosegüten der verscheidenen Modelle und Schätzverfahren zusammengestellt.

	Modell 1		Modell 2	
	Anpassung	Prognose	Anpassung	Prognose
LS-Schätzung	51.3%	39.9%	56.2%	41.2%
L1-Schätzung	53.6%	43.8%	57.8%	41.2%
$t_1(y) = \Phi(y)$	51.0%	45.1%	56.9%	41.2%
$t_2(y) = tendenz(y)$	52.3%	42.5%	56.2%	43.1%
$t_3(y)$ mit $(\lvert t_3(y)\rvert \leq 2)$	51.6%	45.8%	56.2%	39.2%

Tabelle 5: Modell- und Verfahrensvergleich

4 Die Bestimmung der Wahrscheinlichkeiten

In einem ersten Schritt wurde die bei einem Spiel zu erwartende Tordifferenz geschätzt. In der Regel ist man aber daran interessiert, ob die Heim- oder die Auswärtsmannschaft gewinnen wird oder ob es ein Unentschieden geben wird.

Das oben verwendete Rangverfahren (bei dem für die 75 größten $\hat{d}_{ij}$ ein Heimsieg prognostiziert wird) hilft hier auch nicht weiter, da Wahrscheinlichkeiten für jede mögliche Tendenz bestimmt werden soll. Gesucht ist demnach

$$P(1|\hat{d}_{ij}), \quad P(0|\hat{d}_{ij}), \quad \text{und} \quad P(2|\hat{d}_{ij}).$$

Diese Wahrscheinlichkeiten lassen sich mit dem Satz von Bayes berechnen. Allerdings handelt es sich bei $\hat{d}_{ij}$ in diesem Fall um eine stetige Zufallsvariable. Damit folgt:

$$P(1|\hat{d}_{ij}) = \frac{f_1(\hat{d}_{ij}) \cdot p(1)}{f(\hat{d}_{ij})} = \frac{f_1(\hat{d}_{ij}) \cdot p(1)}{f_1(\hat{d}_{ij}) \cdot p(1) + f_2(\hat{d}_{ij}) \cdot p(2) + f_0(\hat{d}_{ij}) \cdot p(0)} \tag{8}$$

Hierbei sind $p(0), p(1)$ und $p(2)$ die a-priori-Wahrscheinlichkeiten für die drei möglichen Spielausgänge. Diese Wahrscheinlichkeiten werden durch die Anteile der verschiedenen Spielausgänge geschätzt. Diese unterlagen in den vergangenen Jahren übrigens nur sehr kleinen Schwankungen:

Saison	Anteil der Heimsiege	Antei der Unentschieden	Anteil der Auswärtssiege
99/00	0.467	0.284	0.248
98/99	0.471	0.284	0.245
97/98	0.474	0.278	0.248
Durchschnitt	$\hat{p}(1) = 0.471$	$\hat{p}(0) = 0.282$	$\hat{p}(2) = 0.247$

Tabelle 6: Anteilsmäßige Aufteilung der Spielausgänge (1997 - 2000)

Daneben werden die bedingten Dichtefunktionen und die unbedingte Dichtefunktion der Zufallsvariable $\hat{d}_{ij}$ benötigt. Diese lassen sich ebenfalls schätzen. Diese sind jedoch für jedes der oben betrachteten Modelle und jede der Schätzmethoden unterschiedlich. In dieser Arbeit werden zwei Modelle geschätzt. Erstens die sehr einfache Tabellenplatzprognose und zweitens das mit der L1-Methode geschätzte Modell 1, mit welchem sich eine gute Anpassung und eine gute Prognose erzielen ließ.

Nachfolgend ist das Histogramm der prognostizierten Tordifferenzen angegeben. Diese Daten lassen sich durch eine mit den Parametern $\mu = 0.549$ und $\sigma = 0.796$ recht gut approximieren. Die bedingten empirischen Verteilungen der erwartenen Tordifferenzen für die verschiedenen möglichen Spielausgänge sind in der folgenden Abbildung gegeben.

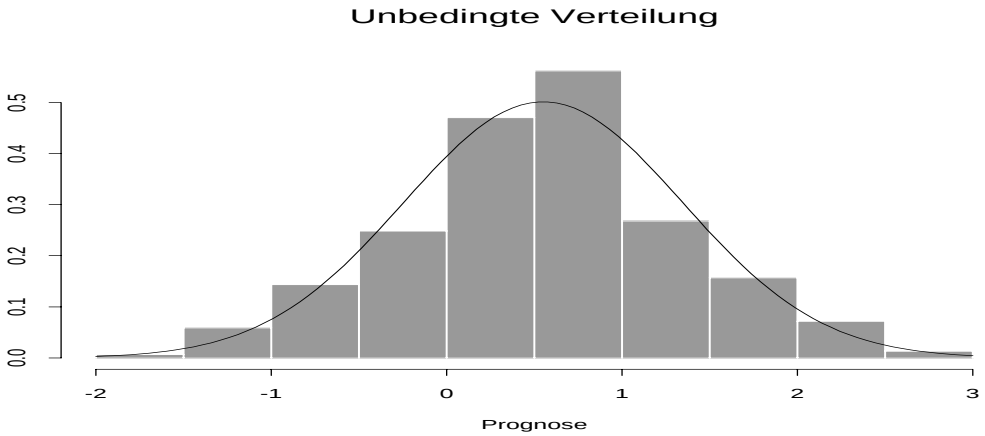

Abbildung 1: Häufigkeitsverteilung der prognostizieten Tordifferenzen (insgesamt)

Unter der Annahme, dass auch diese Verteilungen sich angemessen durch eine normalverteilte Zufallsvariable abbilden lassen, können auch die bedingten Verteilungen geschätzt werden. Damit sind alle Informationen gegeben, um die in Formel 8 angegebenen Wahrscheinlichkeiten zu schätzen.

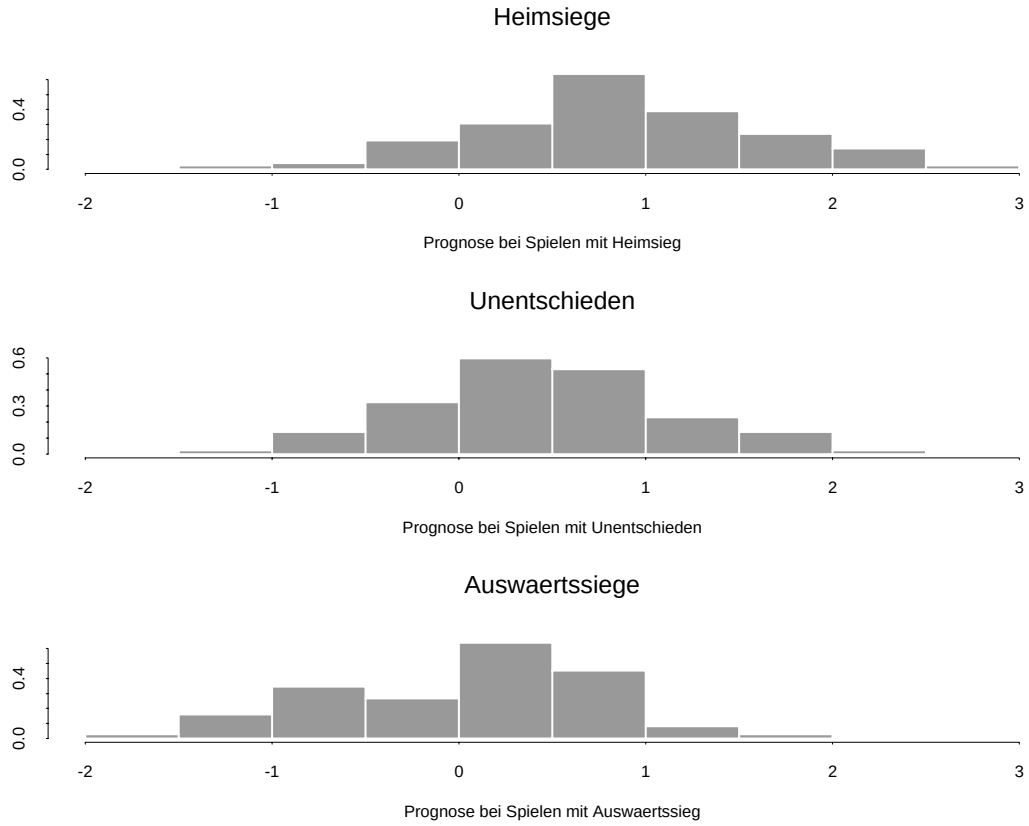

Abbildung 2: Häufigkeitsverteilung der prognostizieten Tordifferenzen (differenziert nach Tendenz)

Diese Wahrscheinlichkeiten werden in S-PLUS mit den im Anhang angegebenen Funktionen bestimmt.

Bei dem oben bereits erwähnten Spiel *Werder Bremen* gegen den *VfL Wolfsburg* hatte die erwartete Tordifferenz den Wert $\hat{d}_{ij}$=0. Dami ergeben sich die folgenden Wahrscheinlichkeiten:

$$P(1|\hat{d}_{ij}=0)=0.322 \quad P(0|\hat{d}_{ij}=0)=0.328 \quad P(1|\hat{d}_{ij}=0)=0.350$$

Falls erwartet wird, dass die Heimmannschaft mit einem Tor gewinnen wird, ergeben sich die folgenden Wahrscheinlichkeiten.

$$P(1|\hat{d}_{ij}=1)=0.584 \quad P(0|\hat{d}_{ij}=1)=0.0.284 \quad P(1|\hat{d}_{ij}=1)=0.131$$

Abschließend sei noch die *handliche Regel* (Modell 0) erwähnt. Bei diesem Modell lässt sich – wie oben bereits erwähnnt – die erwartete Tordifferenz sehr einfach berechnen. Spielt etwa

der 17. gegen den dritten, so lautet die erwartete Tordifferenz $\hat{d}_{ij} = 0.5 + 0.1 \cdot (5 - 17) = -0.7$. Mit dem obigen Bayes-Ansatz lassen sich dann die Wahrscheinlichkeiten für die möglichen Spielausgänge ermitteln. Diese sind in der nachfolgenden Tabelle angegeben. Sie können auch in der Abbildung 4 abgelesen werden.

erw.Tordifferenz	$P(1)$	$P(0)$	$P(2)$
-1.2	0.142	0.290	0.569
-1.1	0.150	0.297	0.552
-1.0	0.160	0.305	0.536
-0.9	0.170	0.312	0.519
-0.8	0.181	0.318	0.501
-0.7	0.193	0.324	0.483
-0.6	0.206	0.330	0.464
-0.5	0.220	0.334	0.446
-0.4	0.236	0.338	0.426
-0.3	0.252	0.341	0.407
-0.2	0.270	0.343	0.387
-0.1	0.290	0.343	0.367
0.0	0.310	0.343	0.347
0.1	0.332	0.341	0.327
0.2	0.356	0.338	0.306
0.3	0.381	0.333	0.286
0.4	0.407	0.328	0.266
0.5	0.434	0.320	0.246
0.6	0.463	0.311	0.226
0.7	0.492	0.301	0.207
0.8	0.522	0.290	0.188
0.9	0.553	0.277	0.170
1.0	0.584	0.263	0.153
1.1	0.615	0.248	0.137
1.2	0.645	0.233	0.122
1.3	0.675	0.217	0.107
1.4	0.705	0.201	0.094
1.5	0.733	0.185	0.082
1.6	0.760	0.169	0.071
1.7	0.786	0.154	0.061
1.8	0.810	0.138	0.052
1.9	0.832	0.124	0.044
2.0	0.853	0.110	0.037
2.1	0.871	0.098	0.031
2.2	0.888	0.086	0.026

Tabelle 8: Tabellenplatzprognose

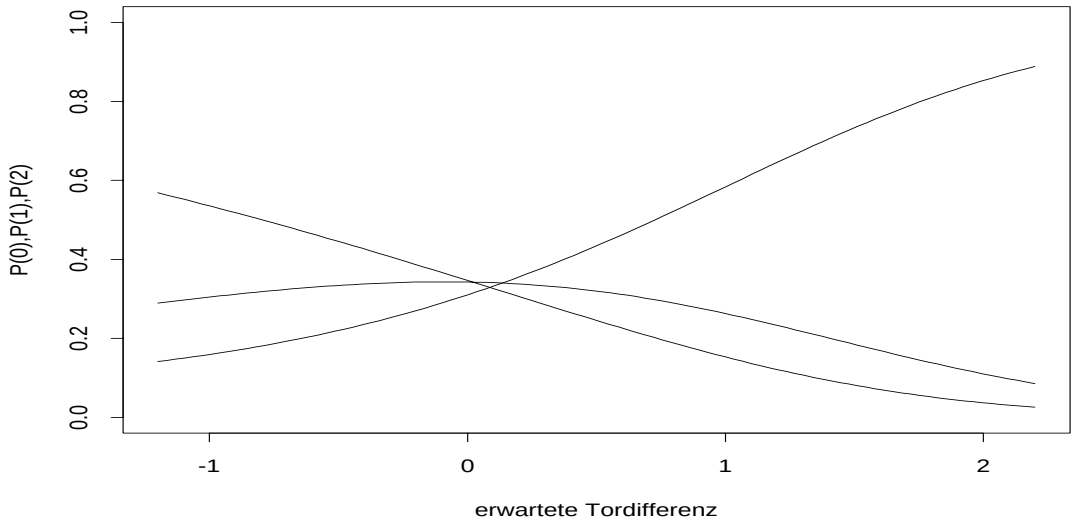

Abbildung 4: Grafische Darstellung der Tabelle 8

Ein solches Modell kann genutzt werden, wenn nicht viele Informationen über die beteiligten Mannschaften bekannt sind. Dies kann der Fall sein bei ausländischen Fußballligen oder bei den ersten Spielen einer Saison, bei denen die Mannschaften lediglich in eine Rangordnung gebracht werden können.

A Literatur

Basset Jr., G.W. (1997): Robust sports ratings based on least absolute errors, in: *American Statistician*, Vol. 51, S.99-105.

Bloomfield, P. and Steiger, W. L. (1983): Least Absolute Deviations: Theory, Applications, and Algorithms. Birkhauser, Boston, Mass.

Dixon, M.J. und S.G. Coles (1997): Modelling association football scores and inefficiencies in the football betting market, in: *Applied Statistics*, Vol. 46, S. 246-280.

Lee, A.J. (1997): Modelling scores in the Premier Leage: is Manchester United really the best?, in: *Chance*, Vol. 10 (1), S. 15-19.

Pope, P.F. und D. Peel (1989): Information, Prices and efficiency in a fixed-odds betting market, in: *Economica*, vol. 56, S. 323-341.

B S-PLUS-Funktionen und -Code

In diesem Abschnitt sind die verwendeten S-PLUS-Funktionen und -Befehle zur Erzeugung der oben dargestellten Schätzergebnisse angegeben.

Der Funktionsaufruf *X.design*(18) erzeugt die für Modell 1 benötigte Deignmatrix.

```
X.design<-function(K)
{
  X.d <- X.soccer(K, 1)
  for(i in 2:K)
  {
     X.d <- rbind(X.d,X.soccer(K,i))
  }
  return(X.d)
}



X.soccer<-function(K, k)
{
# K: Anzahl der Mannschaften
# k: k'te Blockmatrix
#
# Die Designmatrix ergibt sich, falls
# die Blockmatrizen X1,...Xk,...
# untereinandergeschrieben werden

  Xk <- rep(0, (K - 1) * (K -  1))
  Xk <- matrix(Xk, nrow = (K -  1))
  for(i in 2:(K - 1))
     { if(i < k)
         { index <- (i - 1)}
       if(!(i < k))
         { index <- i }
       Xk[i, index] <- (-1)
      }

  if((k - 1) > 0)
     { Xk[, (k - 1)] <- 1 }
  if(k == 1)
     { Xk <- diag(K - 1)
       Xk <-  - Xk
     }
        return(Xk)
}

 X2.design<-function(K)
   {
   X1 <- X1.soccer(K, 1)
```

```
   for(i in 2:K) {X1 <- rbind(X1, X1.soccer(K, i))}
   X2 <- X2.soccer(K, 1)
   for(i in 2:K) { X2 <- rbind(X2, X2.soccer(K, i))}
   X <- cbind(X1, X2)
   X[, K:(2 * K - 1)] <-  - X[, K:(2 * K - 1)]
   return(X)
   }
X2.soccer<-function(K, k)
   {
   X2 <- diag(K)
   X2 <- X2[ - k,  ]
   return(X2)
   }

X1.soccer
function(K, k)
   {
        x <- rep(0, K)
        x[k] <- 1
        x <- rep(x, (K - 1))
        X1 <- matrix(x, byrow = T, nrow = (K - 1))
        X1 <- X1[, -1]
        return(X1)
   }
```

Mit diesen Funktionen können die L1- und LS-Schätzer des Modells problemlos bestimmt werden. Da in der Datenmatrix *buli*9899 ferner Informationen über den Spieltag enthalten sind, können die Schätzungen für die Hinserie und die daraus resultierenden Prognosen ebenfalls leicht erstellt werden.

```
Pvon1.11
function(x)
{
(0.471 * dnorm(x, 0.868, sd = 0.751))/
(0.471 * dnorm(x, 0.868, sd = 0.751) + 0.282 * dnorm(x, 0.459, sd = 0.692) +
 0.247 * dnorm(x, .041, sd = 0.705))
}

Pvon0.11
function(x)
{
(0.282 * dnorm(x, 0.459, sd = 0.692))/
(0.471 * dnorm(x, 0.868, sd = 0.751) + 0.282 * dnorm(x, 0.459, sd = 0.692) +
 0.247 * dnorm(x, .041, sd = 0.705))
}

Pvon2.11
function(x)
```

```
{
(0.247 * dnorm(x, 0.041, sd = 0.705))/
(0.471 * dnorm(x, 0.868, sd = 0.751) + 0.282 * dnorm(x, 0.459, sd = 0.692) +
 0.247 * dnorm(x, .041, sd = 0.705))
}
```