Arch. Tierz., Dummerstorf **50** (2007) 1, 47-58

Aus dem [1]Institut für Agrar- und Ernährungswissenschaften der Martin-Luther-Universität Halle-Wittenberg und dem [2]Forschungsinstitut für Tierproduktion Prag-Uhříněves

NORBERT MIELENZ[1], HANA KREJČOVÁ[1], JOSEF PŘIBYL[2] und LUTZ SCHÜLER[1]

Anpassung eines Fixed Regression Modells für die tägliche Zunahme von Fleckviehbullen mit Hilfe von Informationskriterien*

Abstract

Title of the paper: **Fitting a fixed regression model for daily gain of bulls using information criterion**

In this study the model choice is demonstrated exemplarily on data of 6405 Czech Simmental bulls using information criterion. Per bull up to 8 observations were available for the trait daily gain. Because the animals showed different age on control day, the expected gain curves were described in the population and within the herd*year*season-classes by second, third or fourth order Legendre polynomials of age. For optimization of the fixed effects and to choice the covariance structure of the repeated records the information criteria of Akaike (AIC), the Bayesian criteria (BIC) and the ICOMP-criteria, developed mainly from Bozdogan, were used. Within and over all covariance structures AIC selected generally the most complex model. On the other hand, BIC and ICOMP favoured a model with second order polynomials of age nested within the head*year*season-classes. All criterion selected models with nested second order polynomials within the herd*year*season-classes in comparison to models with non-nested polynomials of age.

Key Words: model selection, information criterion, repeated measurements, daily gain, Simmental bulls

Zusammenfassung

In dieser Studie wird die Modellwahl mit Hilfe von Informationskriterien exemplarisch an Daten von 6405 Fleckviehbullen demonstriert. Pro Bulle lagen bis zu 8 wiederholte Beobachtungen für die tägliche Zunahme vor. Da sich die Tiere am Kontrolltag im Alter unterscheiden, wurden die zu erwartenden täglichen Zunahmen in der Population bzw. innerhalb der Herde*Jahr*Saison (HYS)-Klassen mit Legendre Polynomen 2. bis maximal 4. Ordnung beschrieben. Für die Optimierung der Erwartungswertstruktur und der Kovarianzstruktur der wiederholten Beobachtungen wurden das Informationskriterium von Akaike (AIC), das Bayessche Informationskriterium (BIC) und das hauptsächlich von Bozdogan entwickelte ICOMP-Kriterium genutzt. Innerhalb und über alle acht untersuchten Kovarianzstrukturen hinweg wählte AIC generell das komplexeste Modell aus. Dagegen bevorzugten BIC und ICOMP innerhalb der Kovarianzstrukturen den HYS-Klassen hierarchisch untergeordnete Polynome 2. Grades. Alle Kriterien favorisierten Modelle mit hierarchisch untergeordneten Polynomen 2. Ordnung im Vergleich zu Modellen ohne Hierarchie der Polynome innerhalb der HYS-Klassen.

Schlüsselwörter: Modellwahl, Informationskriterien, wiederholte Messungen, tägliche Zunahme, Fleckviehbullen

1. Einleitung

Der Auswahl der einzubeziehenden fixen und zufälligen Faktoren und deren zeitabhängige Beschreibung innerhalb eines genetisch-statistischen Modells kommt sowohl in der Tierzüchtung als auch im Tierversuchswesen eine entscheidende Bedeutung zu. Sollen Modellwahl und Parameterschätzung an derselben Stichprobe erfolgen, so bedient man sich solcher analytischer Kriterien, wie dem Informationskriterium von AKAIKE (1973, 1974), dem Bayesschen Informationskriterium (SCHWARZ, 1978) oder dem hauptsächlich von BOZDOGAN (1990, 2000) entwickelten ICOMP-Kriterium. In dieser Studie werden die oben angeführten Kriterien zur Anpassung eines Fixed-Regression Modells an die berechneten täglichen Zunahmen von Fleckvieh-

* unterstützt vom Ministerium für Landwirtschaft der Tschechischen Republik (Projekt Nr. 0002701401)

bullen genutzt, wobei berücksichtigt wird, dass pro Tier wiederholte Leistungen vorliegen (vgl. PTAK und SCHAEFFER, 1993; SWALVE, 1995). Die über einen längeren Zeitraum an mehreren Kontrolltagen erfassten Leistungen unterliegen den Einflüssen von Herde, Jahr und Saison und sind zusätzlich abhängig von tierspezifischen und umweltbedingten zufälligen Effekten. Da sich die an einem Kontrolltag geprüften Tiere im Alter unterscheiden, beschreibt man die mittleren täglichen Zunahmen beispielsweise aller Tiere innerhalb einer Herde oder innerhalb einer Herde*Jahr*Saison (HYS)-Klasse durch altersabhängige Kovariablen mit fixen Regressionskoeffizienten. Zusätzlich muss die Struktur der Varianz-Kovarianz Matrix der wiederholten Leistungen eines Tieres modelliert werden. Somit besteht die Modellwahl in der Bestimmung der einzubeziehenden festen und zufälligen Effekte sowie deren Kovarianzstruktur. Das Körpergewicht als auch die tägliche Zunahme von Tieren lassen sich in Abhängigkeit vom Prüfzeitraum durch Polynome 1. bis 5. Grades gut beschreiben (ALBUQUERQUE und MEYER, 2001; HUISMAN, 2002; MEYER, 2001a; MALOVRH, 2003; NOBRE u.a., 2003). Folglich ist bei der Modellwahl der Grad der Polynome zu optimieren, wobei zu beachten ist, dass die fixen Regressionskoeffizienten innerhalb der Herde*Jahr*Saison-Klassen unterschiedliche Werte besitzen können. Unter Verwendung von AIC, BIC und ICOMP gelingt die Optimierung der Erwartungswertstruktur nur, falls die Varianzkomponenten im Modell mit der Maximum-Likelihood (ML)-Methode geschätzt werden. Die analytischen Kriterien, wie AIC und BIC, bewerten die Güte der Modellanpassung durch die mit (-2) multiplizierte logarithmierte Likelihoodfunktion und berücksichtigen die Modellkomplexität über einen Strafterm, welcher von der Anzahl der zu schätzenden Parameter abhängig ist. Erfolgt die Schätzung der Varianzkomponenten mit der REML-Methode, so sind die fixen Modellparameter nicht mehr Bestandteil der zugehörigen „restricted“ Likelihoodfunktion (GROENEVELD und BRADE, 1996; RÖHE u.a., 2000). Basiert die Berechnung der analytischen Kriterien auf der Likelihood der REML-Methode, so kann nur die beste Kovarianzstruktur konkurrierender Modelle bei gegebener Struktur der Erwartungswerte bestimmt werden. Diese Untersuchung wird mit folgenden Zielen durchgeführt.

a) Überprüfung, inwieweit die hierarchische Unterordnung der fixen Regressionskoeffizienten unter die HYS-Effekte dem Ansatz ohne Unterordnung überlegen ist.
b) Auswahl einer geeigneten Struktur für die Varianz-Kovarianz (VC)-Matrix der wiederholten Leistungen pro Tier bei gleichzeitiger Variation des Ansatzes für die fixen Modellparameter.
c) Ableitung von Empfehlungen für die Verwendung von AIC, BIC und ICOMP in Kontext einer zweistufigen Modellwahl.

Die unter (a) und (b) formulierten Ziele lassen sich nur mit analytischen Kriterien überprüfen, falls zu deren Berechnung die Likelihoodfunktion der ML-Methode verwendet wird. In den bekannten Programmen für tierzüchterische Probleme können die Varianzkomponenten nur mit der REML-Methode geschätzt werden. Deshalb erfolgte in dieser Studie die Auswertung mit dem Statistikpaket SAS ohne Berücksichtigung der verwandtschaftlichen Beziehungen zwischen den Bullen. Im Vordergrund steht der Vergleich verschiedener Erwartungswertstrukturen wobei zusätzlich das Verhalten der drei unterschiedlichen Informationskriterien untersucht werden soll. Die Anpassung von linearen gemischten Modellen setzt die Auswahl sowohl einer geeigneten Erwar-

tungswert- als auch einer Kovarianzstruktur voraus, wobei beide Strukturen wiederum nicht unabhängig voneinander sind. Deshalb muss in Abhängigkeit von den Ergebnissen dieser Studie in einem zweiten (hier nicht angegebenen) Schritt die Struktur der VC-Matrix der wiederholten Leistungen nochmals mit analytischen Kriterien basierend auf der Likelihood der REML-Methode bei Einbeziehung aller Pedigreeinformationen optimiert werden.

2. Material

Für die Untersuchungen lagen Körpergewichte von 6405 Bullen der Rasse Tschechisches Fleckvieh vor. Pro Bulle waren durchschnittlich 12 Messungen für das Körpergewicht im Alter von 12 bis 420 Tagen vorhanden. Die Datenerfassung wurde über einen Zeitraum von 20 Jahren in 7 Zuchtstationen durchgeführt. Die täglichen Zunahmen wurden aus jeweils drei aufeinander folgenden Beobachtungen für das Körpergewicht berechnet. Um die Leistungen sowohl mit einem Mehrmerkmals- als auch mit einem Wiederholbarkeitsmodell auswerten zu können (vgl. KREJČOVÁ, u.a., 2007), wurde der Prüfzeitraum vom 12. bis zum 420. Tag in acht Altersabschnitte Δt_1=[12;62], Δt_2=[63;113], Δt_3=[114;164], Δt_4=[165;215], Δt_5=[216;266], Δt_6=[267;317], Δt_7=[318;368] und Δt_8=[369;420] unterteilt. Es liegen also sieben 50-tägige und ein 51-tägiges Zeitintervall vor. Als tägliche Zunahme eines Tieres innerhalb der Intervalle wurde die Leistung mit dem geringsten zeitlichen Abstand zum Mittel des Zeitintervalls (also $\bar{t} = 37$; 88; 139; 190; 241; 292; 343 und 394) ausgewählt. Nach der Aufbereitung des Datenmaterials verfügte ein Bulle im Prüfzeitraum über 4 bis 8 wiederholte Beobachtungen für das Merkmal tägliche Zunahme. Aufeinander folgende Beobachtungen besitzen den gleichen zeitlichen Abstand. Die beschreibende Statistik für die berechnete tägliche Zunahme in den acht gebildeten Intervallen ist in Tabelle 1 aufgeführt.

Tabelle 1
Statistiken für die tägliche Zunahme innerhalb von acht Zeitintervallen (Statistics of daily gain for eight time intervals)

	Mittelpunkt der Zeitintervalle								
Alter	**37**	**88**	**139**	**190**	**241**	**292**	**343**	**394**	**[22-420]**
N	978	3948	5236	6052	6095	5610	4593	1729	34213
Mittel	743,7	930,1	1144,3	1242,8	1288,3	1249,5	1124,9	955,7	1156,3
Std.	196,2	213,9	249,1	218,5	206,3	214,7	229,0	247,0	261,4

Die Einflüsse von Station, Jahr und Saison wurden über die Bildung von Station*Jahr*Saison-Klassen, wobei für die Saison eine dreimonatige Einteilung beginnend mit Dezember gewählt wurde, berücksichtigt. Da die täglichen Zunahmen aus drei aufeinander folgenden Beobachtungen für das Körpergewicht berechnet wurden, ist für dieses Merkmal nicht der Kontrolltag sondern der Kontrollmonat (oder weiter gefasst die Kontrollsaison) als Umweltfaktor bei der Modellbildung zu berücksichtigen. Bedingt durch eine kontinuierliche Beschickung der Prüfstationen über mehrere Jahre unterscheiden sich die Tiere am Kontrolltag in ihrem Alter. Folglich sind Modelle zu bevorzugen, die neben den Umwelteffekten der Kontrollsaison auch das Wachstumsstadium der Tiere direkt bei der mathematisch-statistischen Beschreibung der Kontrolltagsleistungen berücksichtigen.

3. Methoden:

3.1 Auswertungsmodelle

Zur statistischen Analyse der täglichen Zunahmen wurden verschiedene Fixed-Regression Modelle (FRM) angepasst, die sich in der Anzahl der Kovariablen mit zufälligen Regressionskoeffizienten und in der Struktur der Varianz-Kovarianz Matrix der wiederholten Leistungen unterscheiden. Als Kovariablen wurden, wie bei KIRKPATRICK u.a. (1990) vorgeschlagen, auf dem Intervall (-1;1) orthogonale Polynome verwendet. Sei y_{ijk} die tägliche Zunahme ermittelt in Herde*Jahr*Saison-Klasse i an Bullen j zum Alter t_{ijk} Im ersten Schritt wurde das folgende Auswertungsmodell zugrunde gelegt:

$$\text{(FRMn)} \qquad y_{ijk} = \beta_{i0} \cdot \phi_0 + \sum_{m=1}^{n} \beta_{im} \cdot \phi_m(t_{ijk}) + \varepsilon_{ijk} \quad \text{mit} \quad E(y_{ijk}) = \sum_{m=0}^{n} \beta_{im} \cdot \phi_m(t_{ijk})$$

Hierbei sind: t_{ijk} =auf das Intervall (-1;1) standardisiertes Alter, $\phi_m(\cdot)$ =Legendre Polynom m-ten Grades, β_{im} =fixer Regressionkoeffizient m innerhalb Herde*Jahr*Saison-Klasse i und ε_{ijk} =zufälliger Resteffekt.

Sei $\varepsilon_{ij} = (\varepsilon_{ij1},...,\varepsilon_{ijN})'$ der Vektor der zufälligen Resteffekte von Tier j und $\Sigma = Var(\varepsilon_{ij})$ die zugehörige Kovarianzmatrix mit den Elementen $\sigma_{kk'}(k,k' = 1,...,N)$. Untersucht wurden die folgenden Strukturen (vgl. MEYER, 2001b):

$$\text{(CS)} \qquad \sigma_{kk'} = \sigma^2 \cdot \rho \quad \text{für} \quad k \neq k' \quad \text{und} \quad Var(y_{ijk}) = \sigma^2$$

$$\text{(AR1)} \qquad \sigma_{kk'} = \sigma^2 \cdot \rho^{|k-k'|} \quad \text{mit} \quad k,k' = 1,...,N$$

$$\text{(ARH1)} \qquad \sigma_{kk'} = \sigma_k \cdot \sigma_{k'} \cdot \rho^{|k-k'|} \quad \text{mit} \quad k,k' = 1,...,N$$

(UN) keine Restriktion auf Σ

Die obigen Bezeichnungen entsprechenden den in SAS gewählten Abkürzungen. Im Fall (CS) besteht zwischen allen Beobachtungen an Tier j die gleiche Korrelation ρ. Dagegen verändert sich in den Fällen (AR1) und (ARH1) die Korrelation ρ autoregressiv, wobei im Fall (ARH1) unterschiedliche Varianzen für jede wiederholte Beobachtung zugelassen wird. Für die Kovarianzmatrix aller Resteffekte von Modell (FRMn) wird blockdiagonale Gestalt vorausgesetzt. Somit sind in den Fällen (CS) und (AR1) lediglich zwei und in den Fällen (ARH1) und (UN) insgesamt (N+1) bzw. N(N+1)/2 Varianzkomponenten zu schätzen. Die Schätzung der Modellparameter erfolgte mit der Prozedur MIXED des Statistikprogramms SAS unter Verwendung der REPEATED Anweisung und der ML-Methode.

Durch den Übergang zu Random-Regression Modellen lassen sich weitere Strukturen von Σ erzeugen (SCHAEFFER und DEKKERS, 1994). Dazu werden tierspezifische, zufällige Regressionskoeffizienten α_{jm} eingeführt und die Resteffekte in (FRMn) mit Hilfe von Legendre Polynomen (PLG) wie folgt zerlegt:

$$\varepsilon_{ijk} = \sum_{m=0}^{n} \alpha_{jm} \cdot \phi_m(t_k) + e_{ijk}$$

Setzt man $\alpha_j = (\alpha_{j0},...,\alpha_{jn})'$ mit $Var(\alpha_j) = K_a$ und $\phi_k = (\phi_0(t_k),...,\phi_n(t_k))'$. Dann ergibt sich für die Elemente von Σ die folgende Darstellung:

(PLGn) $$\sigma_{kk'} = \phi_k' K_a \phi_{k'} + \delta_{kk'} \cdot \sigma_e^2 \quad \text{mit} \quad \sigma_e^2 = Var(e_{ijk}) \quad \text{und} \quad k, k' = 1,..., N$$

Hierbei gilt: $\delta_{kk'}=1$ für $k=k'$ und Null sonst. Die Schätzung der Varianzkomponenten in Darstellung (PLGn) erfolgte in SAS mit der Prozedur MIXED bei Verwendung der RANDOM Anweisung, wobei mit Hilfe der PARMS Anweisung Startwerte für die zu schätzenden Varianzkomponenten vorgegeben wurden.

Bemerkung 1: Da in Modell (FRMn) vorausgesetzt wird, dass alle Zufallseffekte den Erwartungswert Null besitzen, sollte der Polynomgrad der Erwartungswerte nicht kleiner als der Grad der Polynome zur Beschreibung der zufälligen Parameter gewählt werden. Deshalb wurden zur Modellierung der Erwartungswerte generell Polynome 4. Grades verwendet. Um die Anzahl der Regressionskoeffizienten nicht unnötig zu vergrößern, wurden für hierarchische Ansätze 2. Grades $\beta_{i3}=\beta_3$ und $\beta_{i4}=\beta_4$ sowie für hierarchische Ansätze 3. Grades $\beta_{i4}=\beta_4$ gesetzt. Für die fixen Regressionskoeffizienten zugehörig zu Polynomen höheren Grades als im Polynomansatz für die zufälligen Effekte wurde also keine Unterordnung gefordert.

Bemerkung 2: Durch die unter (AR1) und (ARH1) aufgelisteten Kovarianzen wird die Korrelationsstruktur zwischen wiederholten Leistungen pro Tier ausschließlich vom zeitlichen Abstand zweier Beobachtungen pro Tier modelliert. Es gilt also allgemein:

$$\sigma_{kk'} = \sigma_k \cdot \sigma_{k'} \cdot g(|\, t_k - t_{k'}\,|) \quad \text{mit} \quad k, k' = 1,..., N$$

Hierbei ist g(·) eine Abstandsfunktion (mit g(0)=1), die mit wachsendem Abstand monoton gegen Null strebt. Dagegen wird die Kovarianzstruktur innerhalb von (PLGn) ausschließlich über tierspezifische zufällige Regressionskoeffizienten erzeugt. Es sei erwähnt, dass die Strukturen basierend auf subjekt-spezifischen Zufallseffekten um residuale Korrelationsstrukturen basierend auf Abstandsfunktionen erweitert werden können (VERBEKE u.a., 1998). Derartige Erweiterungen sind zu prüfen, falls pro Tier täglich oder wöchentlich erfasste Leistungen über einen längeren Zeitraum vorliegen (MIELENZ u.a., 2006).

3.2. Analytische Kriterien zur Modellwahl

In Modell (FRMn) aus Abschnitt 3.1 ist der Polynomgrad n und die Struktur der Kovarianzmatrix für die wiederholten Leistungen pro Tier zu bestimmen. Deshalb muss zur Schätzung der Varianzkomponenten die ML-Methode verwendet werden. Sei log$L$ die logarithmierte Likelihoodfunktion im Optimum. Dann besitzen AIC und BIC in Programm SAS die folgende Gestalt:

$$AIC = -2 \cdot \log L + 2 \cdot (p + q)$$

$$BIC = -2 \cdot \log L + (p + q) \cdot \log(N_a)$$

Hierbei sind p die Anzahl der fixen Effekte in log$L$, q die Anzahl der Varianzkomponenten in log$L$ und N_a die Anzahl der unabhängigen Objekte, also die Anzahl Tiere in Modell (FRMn). Präziser formuliert entsprechen p dem Rang der Versuchsplanmatrix der festen Effekte und q der Anzahl der Varianzkomponenten bei Ausschluss der mit Null geschätzten Varianzen.

Basierend auf den Arbeiten von BOZDOGAN (1990, 2000) werden zunehmend so genannte „information-theoretic measure of model complexity (ICOMP)“ genutzt. Anstelle der Anzahl der freien Parameter wird die Modellkomplexität bei Verwendung

von ICOMP mit Hilfe der asymptotischen Varianz-Kovarianz-Matrix der geschätzten fixen Effekte und der geschätzten Varianzkomponenten (also mit der inversen Fisher'schen Informationsmatrix F^{-1}, abgekürzt IFIM) bewertet. Dieses Kriterium ist gegeben durch:

$$ICOMP = (-2) \cdot \log L + C_1(F^{-1})$$

$$\text{mit} \qquad C_1(F^{-1}) = q^* \cdot \log[tr(F^{-1})/q^*] - \log|F^{-1}|$$

Hierbei sind: $q^* = (p+q)$, $tr(F^{-1})$ die Spur und $|F^{-1}|$ die Determinante von Matrix F^{-1}.

Sei $\hat{\beta}$ der Vektor der geschätzten fixen Parameter und $\hat{\sigma}$ der Vektor der geschätzten Varianzkomponenten von Modell (FRMn). Dann wurde die Matrix F^{-1} zur Berechnung von ICOMP wie folgt approximiert:

$$\text{(IFIM)} \qquad F^{-1} = \begin{pmatrix} F_\beta^{-1} & 0 \\ 0 & F_\sigma^{-1} \end{pmatrix} \quad \text{mit} \quad Var(\hat{\beta}) = F_\beta^{-1} \quad \text{und} \quad Var(\hat{\sigma}) = F_\sigma^{-1}$$

Der Ausdruck $C_1(F^{-1})$ berücksichtigt also die Genauigkeit und Komplexität der geschätzten Modelleffekte. Da die Berechnung von Determinanten für große Matrizen numerisch instabil ist, wurde ICOMP unter Ausnutzung der Beziehungen $\log\left|F^{-1}\right| = \sum_i \log(\lambda_i)$ und $\mathrm{Sp}\left(F^{-1}\right) = \sum_i (\lambda_i)$ mit Hilfe der Eigenwerte von F_β^{-1} und F_σ^{-1} berechnet. Die entsprechenden SAS-Programme sind bei den Autoren verfügbar.

Bei Nutzung der analytischen Kriterien ist das Modell zu bevorzugen, dass im Vergleich der Modelle den jeweils kleinsten Wert für AIC, BIC bzw. ICOMP aufweist.

4. Ergebnisse

In Tabelle 2 sind die Werte von AIC, BIC und ICOMP für ein Modell mit nicht untergeordneten Polynomen 4. Grades und für ein Modell mit den (HYS)-Klassen untergeordneten Polynomen 2. Grades in Abhängigkeit von den untersuchten Kovarianzstrukturen aufgelistet. Die Datenaufbereitung lieferte 187 HYS-Klassen. Folglich sind für die Struktur mit hierarchisch untergeordneten Polynomen 2. Grades, insgesamt (3·187) Regressionskoeffizienten zu schätzen. Da dieser Ansatz zusätzlich Polynome 3. und 4. Grades enthält, deren Regressionskoeffizienten lediglich populationsspezifisch sind, ergeben sich 563 zu schätzende fixe Parameter. In Tabelle 2 erreichen AIC und ICOMP ihre kleinsten Werte für ein Modell mit Struktur (UN) bei hierarchischer Unterordnung der Regressionskoeffizienten von Polynomen 2. Ordnung. Dieses Modell besitzt 563 fixe Parameter und 36 zu schätzende Varianzkomponenten. Es repräsentiert in Tabelle 2 das komplexeste Modell und zeigt mit 456367,6 erwartungsgemäß den kleinsten (-2)log*L*-Wert. Zur Berechnung von AIC bzw. BIC müssen zu diesem Wert 2(p+q)=1198 bzw. (p+q)ln(N_a)=5250,1 addiert werden. Hierbei entspricht N_a=6405 der Anzahl aller Bullen. Die relativierten Werte von AIC, BIC und ICOMP besitzen innerhalb der untersuchten sieben Kovarianzstrukturen den kleineren Wert für das Modell mit hierarchisch untergeordneten Regressionskoeffizienten. Folglich bevorzugen AIC, BIC und ICOMP innerhalb der untersuchten Strukturen jeweils das hierarchische Modell.

Arch. Tierz. **50** (2007) 1

Tabelle 2
Anzahl Modellparameter, (-2) LogLikelihood, AIC, BIC und ICOMP als Abweichungen vom kleinsten Wert für eine Erwartungswertstruktur (EW) ohne (plg4+hys) und mit (plg2(hys)) Hierarchie geschätzt mit sieben verschiedenen Kovarianzstrukturen (CV) von Σ (Number of model parameters, (-2) LogLikelihood, AIC, BIC and ICOMP for expected structures with and without nested regression coefficients estimated with different covariances)

Struktur		Anz. Parameter		Statistiken			
CV	**EW**	**p**	**q**	**(-2)logL**	**AIC_rel**	**BIC_rel**	**ICOMP_r**
CS	**plg4+hys**	191	2	461754,0	4574	1878	2877
	plg2(hys)	563	2	457321,0	885	706	804
AR(1)	**plg4+hys**	191	2	460651,4	3472	775	1748
	plg2(hys)	563	2	456696,3	261	81	155
PLG2	**plg4+hys**	191	7	460932,3	3763	1100	2347
	plg2(hys)	563	7	456968,2	543	397	437
ARH(1)	**plg4+hys**	191	9	460462,1	3297	648	2303
	plg2(hys)	563	9	456554,1	133	0	82
PLG3	**plg4+hys**	191	11	460647,1	3486	850	2070
	plg2(hys)	563	11	456755,0	337	218	234
PLG4	**plg4+hys**	191	16	460574,5	3423	821	2020
	plg2(hys)	563	16	456670,3	263	178	183
UN	**plg4+hys**	191	36	460177,6	3066	600	2258
	plg2(hys)	563	36	456367,6	0	50	0

Tabelle 3
Anzahl der Modellparameter, (-2) Log-Likelihood, AIC, BIC und ICOMP als Abweichungen vom kleinsten Wert in Abhängigkeit von der Erwartungswert(EW)- und der Kovarianzstruktur(CV) von Σ. (Number of model parameters, (-2) LogLikelihood, AIC, BIC and ICOMP for different expected and covariances structures)

Struktur		Anz. Parameter		Statistiken			
CV	**EW**	**p**	**q**	**(-2)logL**	**AIC_rel**	**BIC_rel**	**ICOMP_r**
PLG2	**plg4+hys**	191	7	460932,3	4308	1100	2347
	plg2(hys)	563	7	456968,2	1088	397	437
	plg3(hys)	749	7	456056,9	549	1116	3668
ARH(1)	**plg4+hys**	191	9	460462,1	3842	648	2303
	plg2(hys)	563	9	456554,1	678	0	82
	plg3(hys)	749	9	455653,6	150	730	3135
PLG3	**plg4+hys**	191	11	460647,1	4031	850	2070
	plg2(hys)	563	11	456755,0	883	218	234
	plg3(hys)	749	11	455830,0	330	924	3474
PLG4	**plg4+hys**	191	16	460574,5	3968	821	2020
	plg2(hys)	563	16	456670,3	806	169	183
	plg3(hys)	749	16	455731,8	240	861	3425
UN	**plg4+hys**	191	36	460177,6	3612	600	2258
	plg2(hys)	563	36	456367,6	546	50	0
	plg3(hys)	749	36	455450,1	0	763	3074

In Tabelle 3 sind die Werte von AIC, BIC und ICOMP für ausgewählte Strukturen bei weiterer Erhöhung des Grades, der den (HYS)-Klassen untergeordneten Polynomen zusammengestellt. Das Modell mit Kovarianzstruktur (UN) und hierarchisch untergeordneten Polynomen 3. Grades stellt in Tabelle 3 das komplexeste Modell dar. Für dieses Modell zeigt sich mit 455450,1 erwartungsgemäß der kleinste Wert für (-2)log*L*. Auch nach Addition des Strafterms lieferte AIC für das Modell mit 749 fixen Parametern und 36 Varianzkomponenten den kleinsten Wert. Innerhalb aller untersuchten Kovarianzstrukturen bevorzugt AIC generell das Modell mit dem höchsten Polynomgrad. Beispielweise wurden innerhalb der Strukturen PLG4 bzw. UN für die Varianten plg4+hys, plg2(hys) und plg3(hys) relativierte AIC-Werte von 3968; 806 und 240 bzw. von 3612; 546 und 0 gefunden. Dagegen wählen BIC und ICOMP in-

nerhalb der Kovarianzstrukturen immer das Modell mit hierarchischen Polynomen 2. Ordnung aus. Klammert man die Erwartungswertstruktur mit hierarchischen Polynomen 2. Grades aus, so bevorzugen BIC und ICOMP die Modelle ohne Hierarchie gegenüber Modellen mit hierarchischen Polynomen 3-ten Grades. Innerhalb der Strukturvariante ARH(1) wurden für die Erwartungswertstrukturen plg4+hys, plg2(hys) und plg3(hys) relativierte AIC-, BIC- bzw. ICOMP-Werte von 3842; 678; 150 von 648; 0; 730 bzw. von 2303; 82 und 3135 geschätzt.

Obwohl BIC und ICOMP innerhalb der CV-Strukturen gleiche Modelle auswählen, zeigen diese Kriterien über alle CV-Strukturen hinweg minimale Werte für zwei verschiedene Modelle. BIC besitzt den kleinsten Wert für Struktur ARH(1) und ICOMP nimmt in Tabelle 3 den kleinsten Wert für Struktur (UN) an.

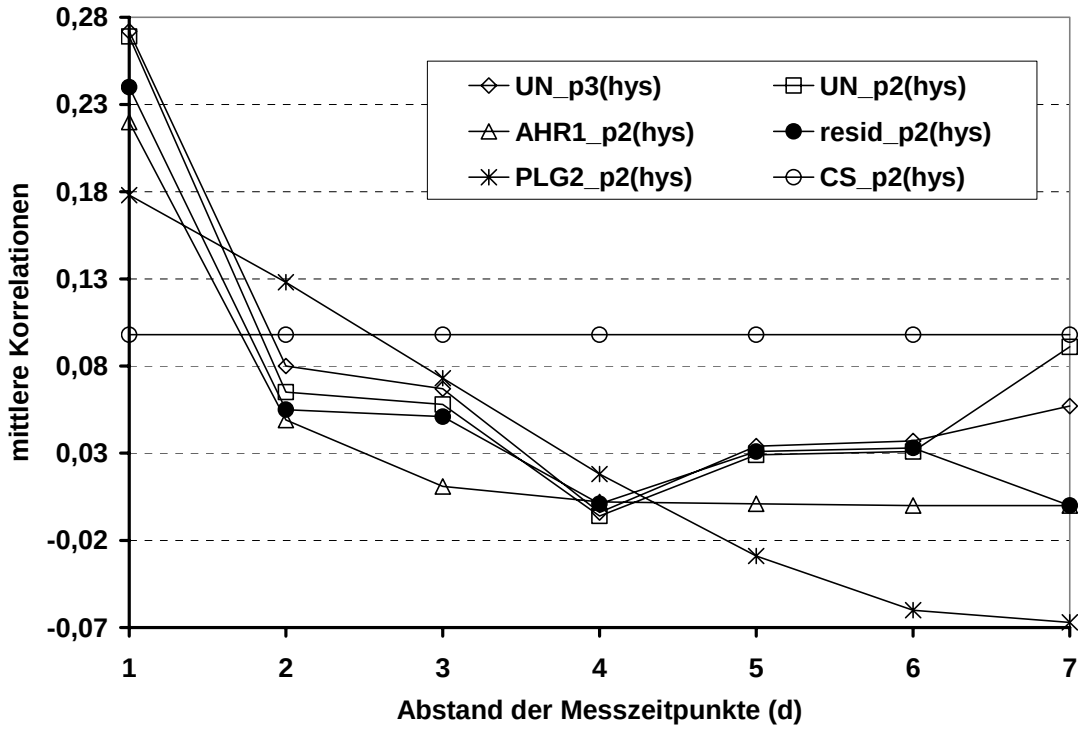

Abb. 1: Mittlere Korrelationen in Abhängigkeit vom Abstand (d) der wiederholten Beobachtungen für unterschiedliche Kovarianzstrukturen (Average correlations as function of lag between the repeated observations for different covariance structures)

Zur Überprüfung der mit AIC, BIC und ICOMP ausgewählten Modelle, wurden sowohl die Korrelationen zwischen den wiederholten Beobachtungen pro Tier (Abb. 1) als auch die Standardabweichungen der Beobachtungen (Abb. 2) geschätzt mit Modell (FRMn) grafisch veranschaulicht. Unter Verwendung der geschätzten Korrelationen zwischen wiederholten Leistungen wurden mittlere Korrelationen als Funktion des Abstandes zwischen den Messzeitpunkten berechnet. Der zeitliche Abstand von 51 Tagen wurde gleich Eins gesetzt. Dadurch ergaben sich sieben mögliche Abstände von 1 bis 7. Die mittleren Korrelationen für ausgewählte Σ-Strukturen als Funktion des Abstandes sind in Abbildung 1 veranschaulicht. Zusätzlich enthält diese Abbildung mittlere Korrelationen geschätzt mit den Residuen resultierend aus einem Regressionsmodell mit hierarchisch untergeordneten Polynomen 2. Ordnung bei Annahme von unkorrelierten Resteffekten. Somit wird zur Schätzung der festen Modelleffekte die „ordinary least square“ (OLS) Methode verwendet. Diese Methode liefert zumindest

erwartungstreue Schätzungen der festen Effekte, deren Standardfehler in Abhängigkeit vom Stichprobenumfang jedoch stark unterschätzt sein kann. Mittlere Korrelationen und Standardabweichungen basierend auf OLS-Residuen von Modell (FRM2) sind in Abbildung 1 und 2 zu Vergleichszwecken mit veranschaulicht.
Durch Verbindung der Punktschätzungen ergeben sich Trendlinien für die Veränderung der mittleren Korrelationen. Die Trendkurven für die Residuen und die gemäß AIC und ICOMP ausgewählten Modelle UN_p3(hys) und UN_p2(hys) zeigen gute Übereinstimmung.

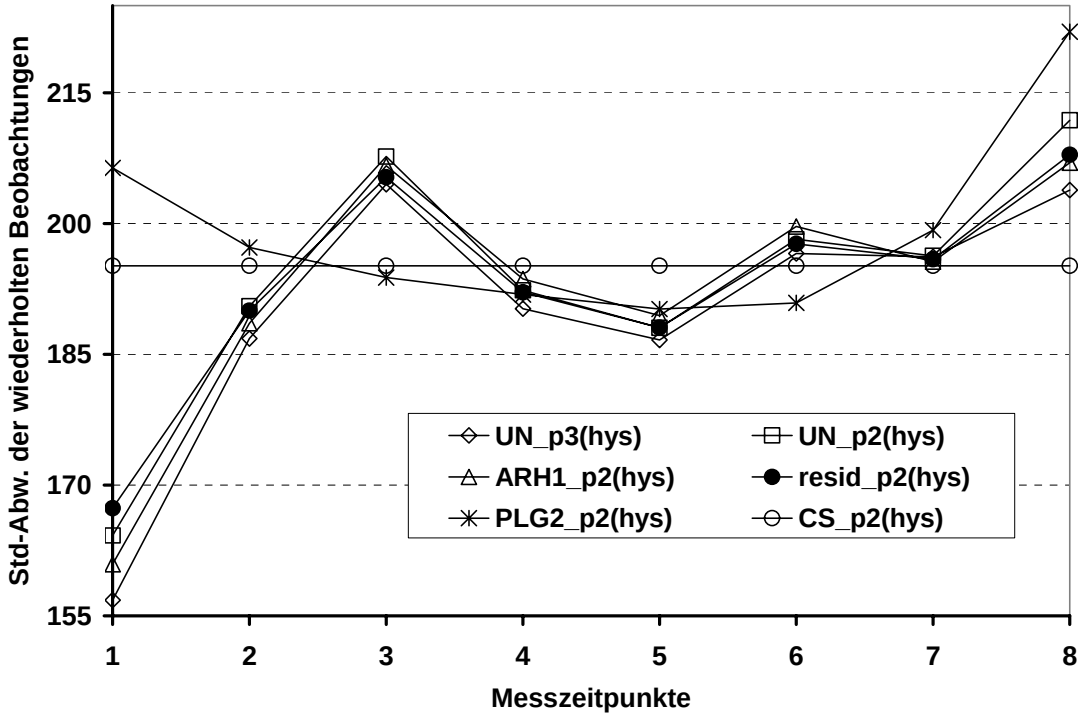

Abb. 2: Standardabweichung der wiederholten Beobachtungen als Funktion der Messzeitpunkte für unterschiedliche Kovarianzstrukturen (Standard deviations of the repeated observations as function at time of measurements for different covariance structures)

Das Modell mit autoregressiver Kovarianzstruktur (ARH1_p2(hys)), ausgewählt mit BIC, glättet die Punktschätzungen der Residuenanalyse. Die mittleren Korrelationen für ARH1_p2(hys) streben monoton fallend beginnend mit einem Wert von 0,22 kontinuierlich gegen Null. Die durch AIC, BIC und ICOMP ausgewählten Modelle zeigen annähernd gleiche Tendenzen und spiegeln den residualen Trend gut wieder. Die durch die drei Kriterien abgelehnten Strukturen CS und PLG2 weisen größere Abweichungen von den ausgewählten Strukturen UN und ARH1 auf. Insbesondere die Variante CS mit einer unveränderlichen Korrelation von 0,098 zeigt ungenügende Anpassung.
In Abbildung 2 sind die geschätzten Standardabweichungen der acht aufeinander folgenden Beobachtungen basierend auf der Residuenanalyse und 4 verschiedenen Kovarianzstrukturen grafisch veranschaulicht. Die durch AIC, BIC und ICOMP ausgewählten Modelle UN_p3(hys), ARH1_p2(hys) und UN_p2(hys) zeigen untereinander ähnliche Tendenzen und stimmen an den acht Messzeitpunkten mit den Schätzwerten der Residuenanalyse gut überein. Die Struktur CS liefert mit einem Wert von 195,1 für alle Messzeitpunkte die gleichen Schätzungen und somit am Anfang und am Ende

größere Abweichungen. Dagegen zeigt PLG2 zumindest ab Messzeitpunkt 4 ähnliche Tendenzen wie die ausgewählten Strukturen.

5. Diskussion

Unabhängig von den untersuchten Strukturen favorisieren AIC, BIC und ICOMP den hierarchischen Ansatz 2. Ordnung gegenüber dem Ansatz mit Polynomen 4. Ordnung ohne hierarchische Unterordnung. Wird der Polynomgrad in den hierarchischen Modellen auf 3 bzw. 4 erhöht, so ergeben sich höhere BIC- und ICOMP-Werte im Vergleich zu den Modellen ohne Hierarchie. Innerhalb der Kovarianzstrukturen zeigten diese beiden Modelle gleiches Verhalten. Sie favorisierten nicht das Modell mit den meisten Parametern sondern das Modell mit Polynomen 2. Grades. Dagegen bevorzugte AIC innerhalb der Kovarianzstrukturen das Modell mit den meisten Parametern und über alle Strukturen hinweg das komplexeste der untersuchten Modelle. Ein Vergleich der drei Kriterien zeigt: AIC bevorzugt tendenziell komplexere Modelle während BIC und ICOMP dazu tendieren, einfachere Modelle auszuwählen. Diese Aussagen werden durch Simulationsstudien bestätigt (BOZDOGAN, 2000; MYANG, 2000; BURNHAM und ANDERSON, 2004). AIC wählte in den Simulationsstudien im Vergleich zu BIC mit größerer Häufigkeit komplexere Modelle als das wahre Modell aus, während BIC gegenüber AIC einfachere Modelle mit größerer Häufigkeit bevorzugte. Die Simulationsergebnisse beziehen sich hauptsächlich auf Regressionsanalysen und zeigen starke Abhängigkeit vom Stichprobenumfang, von dem als wahr angesehenem Modell und von der Menge der konkurrierenden Alternativmodelle.

In dieser Studie sind die Ergebnisse von BIC und ICOMP im Vergleich zu den Resultaten von AIC besser interpretierbar. Bei den hierarchischen Modellen müssen innerhalb der HYS-Klassen Polynome angepasst werden. Teilweise besitzen Klassen nur 40 oder geringfügig mehr Beobachtungen, die aus einem relativ geringen Zeitabschnitt stammen können. In diesem Fall scheint es wenig sinnvoll zu sein, Polynome 3. oder höheren Grades anzupassen. Wählt man dagegen Polynome 4. Grades, deren Koeffizienten den HYS-Klassen nicht untergeordnet sind, so unterscheiden sich beispielsweise die zu erwartenden Zunahmekurven für die Herden und die stark besetzten HYS-Klassen nur um einen konstanten Faktor. Für die Anpassung von Zunahmekurven innerhalb der schwach und stark besetzten HYS-Klassen muss ein Kompromiss gefunden werden, der gemäß BIC und ICOMP in der hierarchischen Unterordnung von Polynomen 2. Grades besteht.

Die hier beschriebene Vorgehensweise zur Modellwahl ist nur ein möglicher erster Schritt. Im zweiten Schritt muss die Struktur der Matrix Σ bei gegebener Struktur für die festen Effekte optimiert werden. Wie in der Tierzüchtung üblich, wird bei dieser Optimierung die Schätzung der Varianzkomponenten mit der REML-Methode erfolgen. Durch Berücksichtigung der verwandtschaftlichen Beziehungen zwischen den Tieren lässt sich die Matrix Σ in genetische Komponenten und in permanente und temporäre Umweltkomponenten weiter zerlegen. Der Übergang zu Random-Regression Modellen ermöglicht es, die Kovariablen zum tatsächlichen Alter am Kontrolltag und nicht nur zu den Mittelpunkten der acht Zeitintervalle aufzustellen. Dadurch lassen sich so genannte Kovarianzfunktionen (analog zu Darstellung (PLGn)) aufstellen, welche die phänotypischen Kovarianzen zwischen Beobachtungen am gleichen Tier zu beliebigen Zeitpunkten t_k und $t_{k'}$ wiedergeben. Der grundlegende Konflikt bleibt jedoch bestehen. Basiert die Schätzung der Varianzkomponenten auf der

REML-Methode, so kann nur die Struktur der Kovarianzfunktionen optimiert werden. Erst der Übergang zur ML-Methode, welche für die Schätzung der Varianzkomponenten eigentlich nicht angebracht ist, gestattet die gleichzeitige Optimierung der fixen Effekte und die Auswahl einer geeigneten Struktur für Σ. Zwangsläufig ist man auf Zweischrittverfahren angewiesen. Natürlich besteht die Möglichkeit zuerst Σ zu optimieren und im zweiten Schritt den besten Ansatz für die fixen Effekte auszuwählen (vgl. LONG und BRAND, 1997). Allerdings muss dann im ersten Schritt ein möglichst komplexer Ansatz für die fixen Effekte verwendet werden. Beispielsweise hierarchisch untergeordnete Polynome 4. oder höheren Grades, wobei die Saison weniger als drei Monate einschließen könnte. Unabhängig vom gewählten Vorgehen sollte letztendlich ein Modell ausgewählt werden, dessen fixe und zufällige Parameter gut interpretierbar sind und dessen geschätzte Varianz- und Korrelationsfunktionen sich beispielsweise im Einklang mit entsprechenden OLS-Residuenanalysen befinden. Abschließend sei bemerkt: Wesentlich für eine gute Modellwahl ist die Festlegung der konkurrierenden Modelle, d.h. der so genannten Priormodelle aus denen die analytischen Kriterien das „beste“ Modell auswählen. Hierzu gehört Verständnis für das Datenmaterial, Kenntnisse über mögliche Auswertungsmodelle als Ergebnis einer sorgfältigen Literaturanalyse und die Auslotung der numerischen Umsetzbarkeit.

Literatur

ALBUQUERQUE, L.G.; MEYER, K.:
Estimates of covariance functions for growth from birth to 630 days of age in Nelore cattle. J. Anim. Sci. **79** (2001), 2776-2789

AKAIKE, H.:
Information theory and an extension of the maximum likelihood principle. 2nd Int. Symp. Information theory. B.N. Petrov and F. Csaki, ed. Akademiai Kiado, Budapest, Hungary (1973), 267-281

AKAIKE, H.:
A new look at the statistical model identification. IEEE Transactions on Automatic Control. **19** (1974),716-723

BOZDOGAN, H.:
On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. Communications in Statistics: Theory and Methods. **19** (1990), 221-278

BOZDOGAN, H.:
Akaike’s information criterion and recent developments in information complexity. Journal of mathematical psychology. **44** (2000), 62-91

BURNHAM, K.P.; ANDERSON, D.R.:
Multimodel inference: understanding AIC and BIC in model selection. Sociological methods and research **33** (2004), 261-304

GROENEVELD, E.; BRADE, W.:
Rechentechnische Aspekte der multivariaten REML Kovarianzkomponentenschätzung, dargestellt an einem Anwendungsbeispiel aus der Rinderzüchtung. Arch. Tierz., Dummerstorf **39** (1996), 81-87

HUISMAN, A. E.:
Genetic analysis of growth and feed intake pattern in pigs. (2002), Diss. Wageningen

KREJČOVÁ, H.; MIELENZ, N.; PŘIBYL, J.; SCHÜLER, L.:
Estimation of genetic parameters for daily gains of bulls with multi-trait and random regression models. Arch. Tierz., Dummerstorf **50** (2007) 1, 000-000

KIRKPATRICK, M.; LOFSVOLD, D.; BULMER, M.:
Analysis of inheritance, selection and evolution of growth trajectories. Genetics **124** (1990), 979-993

LONG, N.; BRAND, R.:
Model selection in linear mixed effects models using SAS PROC MIXED. SAS, SUGI 22 Proceedings, Statistics, Data Analysis, and Modelling, paper **284** (1997) San Diego, California

MALOVRH, S.:
Genetic evaluation using random regression models for longitudinal measurements of body weight in animals. (2003), Diss. Ljubljana

MEYER, K.:
Estimates of direct and maternal covariance functions for growth of Australian beef calves from birth to weaning. Genet. Sel. Evol. **33** (2001a), 487-514

MEYER, K.:
Estimating genetic covariance functions assuming a parametric correlation structure for environmental effects. Genet. Sel. Evol. **33** (2001b), 557-585

MIELENZ, N.; SPILKE, J.; KRECJOVA, H.; SCHÜLER, L.:
Statistical Analysis of Test-Day Milk Yields Using Random Regression Models for the Comparison of Feeding Groups during the Lactation Period. Arch. Anim. Nutr. **60** (2006), 341-357

MYANG, I. J.:
The importance of complexity in model selection. Journal of mathematical psychology **44** (2000), 190-204

NOBRE, P.R.; MISZTAL, I.; TSURUTA, S.; BERTRAND, J.K.; SILVA, L.O.; LOPES, P.:
Analyses of growth curves of Nellore cattle by multiple-trait and random regression models. J. Anim. Sci. **81** (2003), 918-926

PTAK, E.; SCHAEFFER, L.R.:
Use of test day yields for genetic evaluation of dairy sires and cows. Livest. Prod. Sci., **34** (1993), 23-34

RÖHE, R.; KRIETER, J.; PREISINGER, R.:
Bedeutung der Varianzkomponentenschätzung für die Zucht von landwirtschaftlichen Nutztieren – eine Übersicht. Arch. Tierz., Dummerstorf **43** (2000), 523-534

SCHAEFFER, L.R.; DEKKERS, J.C.M.:
Random regressions in animal models for test-day production in dairy cattle. Proc. 5th WCGALP, Guelph, Canada (1994)

SCHWARZ, G.:
Estimating the dimension of a model. Annals of Statistics **6** (1978), 461-464

SWALVE, H.H.:
Test day models in the analysis of dairy production data - a review. Arch. Tierz., Dummerstorf, **38** (1995) 6, 591-612

VERBEKE, G.; LESAFFRE, E.; BRANDT, L.J.:
The detection of residual serial correlation in linear mixed models. Statistics in medicine **17** (1998), 1391-1402

Eingegangen: 01.08.2006

Akzeptiert: 05.12.2006

Autor für Korrespondenz
Dr. NORBERT MIELENZ
Institut für Agrar- und Ernährungswissenschaften
der Martin-Luther-Universität Halle-Wittenberg
Adam-Kuckhoff-Straße 35
06108 HALLE

E-Mail: norbert.mielenz@landw.uni-halle.de