Elektronische Datenverarbeitung in der Tierzüchtung

Das Rechenzentrum der Humboldt-Universität zu Berlin beging in diesem Jahr sein 30jähriges Gründungsjubiläum und hat in der Zeit seines Bestehens viele Bereiche der Universität nachhaltig unterstützt. Das trifft auch für das ehemalige Institut für Tierzüchtung und Haustiergenetik der Landwirtschaftlich-Gärtnerischen Fakultät zu. Die beiden Einrichtungen verbindet eine nahezu 30jährige Zusammenarbeit. Der nachfolgende Beitrag soll die Notwendigkeit der engen Kooperation eines züchterisch orientierten Institutes mit dem Rechenzentrum unterstreichen. Er ist gleichzeitig als Dank an die Mitarbeiter des Rechenzentrums für den großen Einsatz bei der Lösung inhaltlicher und organisatorischer Probleme gedacht, die in den vielen Jahren intensiver Zusammenarbeit anstanden.

Einleitung

Seit der Wiederentdeckung der Mendelschen Regeln im Jahre 1900 wurde die Züchtung in zunehmendem Maße wissenschaftlich auf der Basis der vorhandenen genetischen Erkenntnisse betrieben. Die Genetik hat seit Beginn des Jahrhunderts eine stürmische Entwicklung erfahren und umfaßt heute viele Teildisziplinen, deren Bedeutung für die Tierzucht sehr unterschiedlich ist. Die Klassische Genetik, auch als Faktoren- oder Mendelgenetik bezeichnet, vermochte die Züchtung nur in geringem Maße zu beeinflussen, da sie sich mit einzelnen Genen und deren Wirkung auf die Ausprägung der Merkmale auseinandersetzte. Nur wenige wirtschaftlich interessante Eigenschaften der Nutztiere werden von einzelnen oder wenigen Genen kontrolliert.

Die entscheidenden Leistungseigenschaften, wie Wachstum, Milch- oder Legeleistung, werden von einer Vielzahl im einzelnen nicht bekannter Gene beeinflußt. Ihre Wirkung wird darüber hinaus stark durch Umwelteinflüsse modifiziert, zu denen vor allem Fütterung, Haltung, Pflege und Klima gehören, um nur die wesentlichsten zu nennen. Mit der Vererbung dieser o.g. Merkmale und ihrer züchterischen Nutzung beschäftigt sich die quantitative Genetik. Sie ist ein Teil der Populationsgenetik und benutzt in starkem Maße Methoden der Mathematischen Statistik. Obwohl die Grundlagen der Populationsgenetik durch Hardy und Weinberg schon 1908 und die der quantitativen Genetik 1920 von R. A. Fisher entwickelt worden waren, ließ ihre praktische Anwendung noch einige Jahrzehnte auf sich warten. Die Erkenntnisse der quantitativen Genetik wurden in einigen entwickelten Tierzuchtländern in den 50er Jahren, verstärkt jedoch erst in den 60er Jahren genutzt.

Bedeutung der quantitativen Genetik in der Tierzüchtung

Die relativ späte Nutzung der Methoden der quantitativen Genetik in der Tierzüchtung hat mehrere Ursachen. Ein wesentlicher Grund war der Mangel an technischen Voraussetzungen, große Mengen an Daten zu verarbeiten. Es ist deshalb kein Zufall, daß der verstärkte Einsatz von Methoden der quantitativen Genetik in der Züchtung mit der Einführung der elektronischen Datenverarbeitung in weite Bereiche des gesellschaftlichen Lebens zusammenfällt.

Die quantitative Genetik an sich als auch ihre Bedeutung für die Tierzüchtung sind außerhalb der Fachwelt relativ unbekannt, sicherlich nicht zuletzt ihres abstrakten Charakters wegen. Das wird im Vergleich zur Molekulargenetik besonders deutlich, deren Bekanntheitsgrad bedeutend größer ist - wer hat nicht schon etwas von der Doppelhelix nach Crick und Watson gehört -, die aber in der Tierzüchtung erst seit wenigen Jahren auf begrenzten Gebieten praktische Bedeutung erlangt hat.

Die quantitative Genetik analysiert und charakterisiert die in der Tierzüchtung interessanten Merkmale (Wachstum, Milchleistung u.a.) auf der Populationsebene mit Hilfe von definierten Parametern. Dazu gehört vor allem der Heritabilitätskoeffizient, kurz auch Erblichkeitsgrad genannt, der das Verhältnis der genetisch bedingten Varianz des Merkmals in der Population zur gesamten (phänotypischen) Varianz zum Ausdruck bringt.

Weitere, genetisch wichtige Parameter für die quantitative Genetik und ihre Anwendung in der Züchtung sind die genetischen Korrelationskoeffizienten. Sie spiegeln die genetischen Abhängigkeiten zwischen den Merkmalen wider.

Für eine gründliche Zuchtplanung zur Maximierung des Zuchtfortschritts in der Population sind Schätzwerte dieser genetischen Parameter erforderlich, die bestimmte Genauigkeitsanforderungen erfüllen müssen. Als Maß der Genauigkeit wird z. B. der Standardfehler des Schätzwertes verwendet, der eine vorgegebene Höhe nicht überschreiten sollte. Der Standardfehler ist Grundlage für die Bildung eines Konfidenzintervalles, das die Genauigkeit noch anschaulicher zum Ausdruck bringt.

Da die Genauigkeit auch von der Größe des zu schätzenden Parameters abhängt, sind allgemein verbindliche Aussagen über den notwendigen Stichprobenumfang nicht möglich. Aus den Grundlagen der Versuchsplanung für die Schätzung genetischer Parameter (Rasch u. a. 1978) kann jedoch abgeleitet werden, daß für die Schätzung der Heritabilitätskoeffizienten bei günstiger Struktur des Materials ein Mindeststichprobenumfang von 2000 Versuchseinheiten, für die Schätzung genetischer Korrela-

tionskoeffizienten sogar mindestens 6000 Versuchseinheiten erforderlich sind. Die Versuchseinheit ist im allgemeinen ein Tier, an dem die Selektionsmerkmale gemessen bzw. festgestellt werden.

Für die Schätzung genetischer Parameter ist eine definierte und bekannte Verwandtschaftsstruktur unter den Tieren erforderlich, um die Varianz- und Kovarianzkomponenten als Anteile der genetischen Varianz- und Kovarianzkomponenten zu interpretieren und daraus die erforderlichen Schätzwerte abzuleiten. Berücksichtigt man, daß in ein Zuchtprogramm mindestens 10 bis 20 Selektionsmerkmale einbezogen werden müssen, ist leicht einzusehen, daß hier eine Fülle von Daten zu verarbeiten ist, die mit herkömmlicher Rechentechnik nicht zu bewältigen ist. Es ist noch hinzuzufügen, daß dieser Prozeß wiederholt durchzuführen ist, da die genetischen Parameter keine Konstanten sind, sondern sich im Laufe des Züchtungsprozesses verändern und neu geschätzt werden müssen. Auf die Kompliziertheit der Datenstruktur ist später noch einzugehen.

Merkmalsmodelle als Basis für Parameterschätzungen

Für jede Parameterschätzung ist ein Merkmalsmodell notwendig, dem eine Hypothese über das ursächlich bedingte Zustandekommen der auszuwertenden Beobachtungswerte zugrunde liegt. Das Modell muß alle wesentlichen Einflüsse auf die Beobachtungswerte erfassen, um seine Funktion zu erfüllen. Das bedeutet, daß neben den eigentlich interessierenden genetischen Einflußgrößen auch alle systematisch wirkenden Umwelteinflüsse in das Modell aufgenommen werden müssen. Sie zeichnen sich dadurch aus, daß sie auf eine ganze Gruppe von Tieren (allgemein Versuchseinheiten) in gleicher Richtung und Stärke wirken. Durch ihre Berücksichtigung im Modell wird angestrebt, daß die sogenannten "Restfehler", die im allgemeinen mit e bezeichnet werden, als "zufällige Fehler" mit dem "Erwartungswert" Null angesehen werden dürfen.

Als einfaches Beispiel möge das folgende Modell einer zweifaktoriellen, kreuzklassifizierten Varianzanalyse mit Wechselwirkung und festen Effekten (in klassischer Schreibweise) dienen

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk}$$

dabei ist

y_{ijk} = Merkmalswert des k-ten Nachkommen

μ = "mittlerer" Erwartungswert

a_i = Effekt des i-ten Vaters (als fester Effekt betrachtet) i= 1...A

b_j = Effekt des j-ten Betriebes j = 1...B

$(ab)_{ij}$ = Interaktion zwischen dem i-ten Vater und dem j-ten Betrieb

e_{ijk} = Zufallsfehler k = 1...N_{ij} , $N_{ij} > 1$ Sind alle N_{ij} gleich groß, liegt der balancierte Fall vor.

N = Summe aller N_{ij}

Neben $E(e_{ijk}) = 0$ sollen die Fehler unkorreliert sein und gleiche Varianz σ^2 besitzen. Für statistische Tests wird noch zusätzlich eine Normalverteilung der Fehler vorausgesetzt.

Das zugehörige Datenmaterial besteht (mindestens) aus einer Matrix vom Typ (N,3), d.h. N Zeilen und 3 Spalten (y_{ijk} , i, j). Kern der Berechnung ist die Methode der kleinsten Quadratsummen (LS), d.h. die Minimierung der Quadratsumme der e_{ijk} . Diese führt auf das lineare Gleichungssystem der Normalgleichungen für die gesuchten Schätzungen der Effekte. Die Normalgleichungen besitzen oft keine eindeutigen Lösungen, was man an obiger Modellgleichung erkennt, denn z.B. könnte man von allen a_i eine Konstante c subtrahieren und zu den b_j addieren, was ebenfalls eine Lösung liefert. Daher muß man, wenn man die Schätzung der einzelnen Effekte benötigt, den Normalgleichungen weitere, möglichst plausible Gleichungen (sogenannte Reparametrisierungsbedingungen) hinzufügen, um ein eindeutig auflösbares Gleichungssystem zu erhalten. Dessen Lösung hängt also in hohem Maße von den in den unbalancierten Fällen recht problematischen Reparametrisierungsbedingungen ab. Deshalb versucht man, mit "schätzbaren" linearen Funktionen auszukommen.

Dem aufmerksamen Leser wird nicht entgangen sein, daß der mütterliche Einfluß auf die Merkmalsbildung des Nachkommen unberücksichtigt bleibt, obwohl er sicher nicht unwesentlich ist. Es handelt sich hierbei um ein sogenanntes Halbgeschwistermodell, bei dem alle Nachkommen von verschiedenen Müttern abstammen und somit die Effekte der verschiedenen Mütter in e eingehen, ohne daß die o.g. Voraussetzungen verletzt werden.

Allgemein lassen sich diese linearen Modelle (mit festen Effekten) in Matrixschreibweise wie folgt darstellen

$$y = X\beta + e$$

Es sind:

y = Spaltenvektor der Beobachtungswerte (N,1)

X = Versuchsplanmatrix (N,p)

p = Anzahl der Effekte im Modell (im Beispiel des Halbgeschwistermodells ist p = 1+A+B+AB)

X' = transponierte Matrix von X

β = Spaltenvektor aller Effekte (p,1)

e = Spaltenvektor der Restfehler (N,1)

Diese Darstellung erlaubt einen hohen Verallgemeinerungsgrad und eine geschlossene Behandlung

der mathematischen Theorie zahlreicher Modellklassen. Die eindeutige Berechnung der Schätzung b aller Effekte β des Modells ist meist nicht möglich, da die Matrix des Normalgleichungssystems

$$X'Xb = X'y$$

im allgemeinen keinen vollen Rang p hat. Die Lösung kann man in der Form

$$b = (X'X)^{-}X'y$$

schreiben, wobei $(X'X)^{-}$ eine beliebige "Verallgemeinerte Inverse" von X'X ist.

Jede Lösung b für den Parametervektor β minimiert (e'e) , woraus sich eine Schätzung für die unbekannte Varianz σ^2 ergibt. Die dazu vorausgesetzte Unkorreliertheit und Varianzhomogenität läßt sich wie folgt schreiben:

$$D(e) = \sigma^2 I_N$$

mit D(e) = Dispersionsmatrix des Fehlervektors
I_N = Einheitsmatrix der Ordnung N

Prüfbare Hypothesen sind dann Gleichungen der Form

$$K\beta = a$$

Dabei sind:

K = gegebene Koeffizientenmatrix vom Typ (r,p)

a = gegebener Koeffizientenvektor vom Typ (r,1)

r = Anzahl der Hypothesen

Die Zeilen von Kβ müssen schätzbare lineare Kontraste seien, d.h. Kb ist unabhängig von der ausgewählten Lösung b . Dazu muß K folgende Bedingung erfüllen:

$$K = K(X'X)^{-}(X'X)$$

Die meisten in der Tierzüchtung anzuwendenden Modelle sind jedoch erheblich allgemeiner. Es sind "gemischte Modelle" mit der Modellgleichung

$$y = X\beta + Zu + e$$

mit

Z = Versuchsplanmatrix der zufälligen Effekte (N,q)

u = Vektor der zufälligen Effekte (q,1)

Neben der Unkorreliertheit von u und e muß für den Erwartungswert E(u) = 0 gelten. D(u) ist die Matrix der Kovarianzkomponenten.

In der Züchtung werden genetische Effekte von Individuen (Väter oder Mütter) wegen der zufälligen Weitergabe der Chromosomen bei der Keimzellenbildung als zufällig, die meisten Umwelteffekte dagegen als fix interpretiert.

Die Effekte, ihre Erwartungswerte, Varianzen und Kovarianzen sind an eine Reihe von Voraussetzungen gebunden, auf die nicht näher eingegangen werden soll. Wesentlich ist jedoch, daß die Kovarianzmatrix $D(y) = \sigma^2 V$ eine quadratische Matrix vom Typ (N,N) und keine Diagonalmatrix ist.
Eine Lösung für β lautet nun

$$b = (X'V^{-1}X)^{-}X'\,V^{-1}\,y$$

enthält also die Inverse von V .

Da N sehr groß sein kann, ist eine Berechnung von V^{-1} meist nicht realisierbar, wenn man nicht an V weitere Bedingungen stellt, wie z.B. in der von Henderson vorgeschlagenen gemischten Modellgleichung (MME). Trotzdem erfordert auch die Lösung dieser Gleichungssysteme eine leistungsfähige Rechentechnik und ein hohes Niveau der Software, besonders unter dem Gesichtspunkt der Behandlung multivariater Modelle, bei denen anstelle des Merkmals y dann ein Merkmalsvektor steht.

Methoden zur Schätzung genetischer Parameter

Die erwähnten genetischen Parameter sind Quotienten linearer Funktionen von Varianz- und Kovarianzkomponenten, so daß die meisten verwendeten Methoden der Mathematischen Statistik die Schätzung von Varianz- und Kovarianzkomponenten zum Gegenstand haben.

Wesentliche Methoden zur Schätzung genetischer Parameter gehen auf Henderson zurück, der insgesamt drei vorschlägt. Sie werden auch als ANOVA-Verfahren bezeichnet, weil sie auf dem Prinzip der Varianzanalyse basieren.

HENDERSON-1 entspricht dem Standardverfahren der Varianzanalyse und kann auf Modelle angewendet werden, die außer μ nur zufällige Effekte der Faktoren enthalten (random models). Von besonderer Bedeutung ist die Methode HENDERSON-3, da sie für Merkmalsmodelle geeignet ist, in denen sowohl zufällige als auch fixe Effekte enthalten sind. Sie ist auch auf unbalancierte Datenstrukturen anwendbar und liefert erwartungstreue Schätzwerte. Allerdings können auch Schätzwerte auftreten, die außerhalb des Definitionsbereiches des Parameters liegen.

Weitere Methoden zur Schätzung genetischer Parameter für genetische Modelle und unbalancierte Datenstrukturen gehen auf die Anwendung des Maximum-Likelihood-Prinzips (ML) zurück. Hier ist ein Verfahren besonders beachtenswert, das die Maximum-Likelihood-Eigenschaften nur auf die interessierenden Varianz- und Kovarianzkomponenten der genetischen Effekte bezieht, während die fixen Effekte, meist systematische Umwelteffekte, eliminiert werden. Das Verfahren ist unter dem Namen REML (restricted maximum likelihood) bekannt. Es hat den Vorteil, wie alle ML-Verfahren, daß nur Schätzwerte auftreten, die im Definitionsbereich der Parameter liegen. Für die Schätzung mehrerer Parameter (Heritabilitätskoeffizienten, genetische und phänotypische Korrelationskoeffizienten) ist gesichert, daß die

Schätzwerte positiv definiert sind. Das ist eine Voraussetzung für die Anwendung der gewonnenen Schätzwerte für die anschließende Zuchtplanung. Im Gegensatz zu den HENDERSON-Methoden ist eine Normalverteilung der untersuchten Merkmale Voraussetzung, was als nachteilig anzusehen ist. Noch schwerer fällt jedoch ins Gewicht, daß Schätzgleichungen nichtlinear sind und nur iterativ gelöst werden können. Der anfallende Rechenaufwand ist, immer unter Beachtung der erforderlichen Stichprobengröße, nur mit einer leistungsfähigen Computertechnik zu bewältigen.

Mit einem weiteren, MINQUE genannten Verfahren (minimum norm quadratic unbiased estimation), wird versucht, erwartungstreue Schätzwerte bei minimaler Stichprobenvarianz zu erhalten. Das Verfahren erfordert im Gegensatz zu REML keine Normalverteilung und keine iterative Lösung. Es ist jedoch für die Lösung die Vorgabe der "wahren" Parameter notwendig, die nicht bekannt sind, so daß beim gleichen Datenmaterial verschiedene Lösungen in Abhängigkeit vom Ausgangswert möglich sind. Ebenso können auch Schätzwerte außerhalb des theoretischen Parameterbereichs liegen.

Unter Berücksichtigung der Vor- und Nachteile verschiedener Verfahren kommt Searle (1989) zu dem Ergebnis, den REML- Methoden den Vorzug zu geben.

Simulationsstudien

Wenn über die Bedeutung der modernen Rechentechnik für die Züchtung diskutiert wird, dürfen Simulationsstudien nicht vollkommen unerwähnt bleiben. Neben den klassischen Methoden der Zuchtplanung, mit deren Hilfe unter Berücksichtigung der geschätzten Populationsparameter für verschiedene Bedingungen (Selektionsschärfe, Nutzungsdauer u.ä.) der erreichbare Zuchtfortschritt vorhergesagt werden kann, ist es mit Hilfe der modernen Computertechnik möglich, den Genotyp von Individuen auf dem Computer darzustellen und die Weitergabe der Erbinformation über die Gene von einer Generation zur anderen zu simulieren. Es versteht sich von selbst, daß die Zahl der zu berücksichtigenden Gene ebenso wie die Populationsgröße (Anzahl der Tiere) trotz großer Fortschritte in der Datenverarbeitung beschränkt bleiben muß, sich aber mit der weiteren Entwicklung der Rechentechnik ständig neue Möglichkeiten ergeben.

Die Simulationsstudien haben bisher sehr interessante wissenschaftliche Ergebnisse gebracht. Sie eignen sich dazu, die Theorie der quantitativen Genetik zu überprüfen und zu vervollkommnen. Eine Modellbildung, die die Individuen der Population mit ihren übertragbaren Genen abbildet, steht dem biologischen Prozeß der Vererbung weit näher als ein Modell, das die Merkmale auf Populationsebene widerspiegelt.

Ausblick

Die Tierzüchtung ist gegenwärtig dabei, an den Erkenntnissen und Methoden der molekularen Genetik zu partizipieren. Es zeigt sich, daß diese Ergebnisse weitgehend nur mit Hilfe von Methoden der Populations- und quantitativen Genetik zu nutzen sind. Auch die Methoden der quantitativen Genetik sind noch ständig in der Entwicklung begriffen, obwohl schon vor zwei Jahrzehnten deutliche Skepsis geäußert wurde, ob aus dieser Richtung noch weitere Fortschritte zu erwarten seien. Somit wird die elektronische Datenverarbeitung auch in Zukunft für die Tierzüchtungsforschung ihre Bedeutung behalten.

Gerhard Seeland *
Andreas Baudisch

* Herr Dr. Gerhard Seeland ist Professor am Institut für Angewandte Nutztierwissenschaften der Landwirtschaftlich-Gärtnerischen Fakultät der Humboldt-Universität zu Berlin