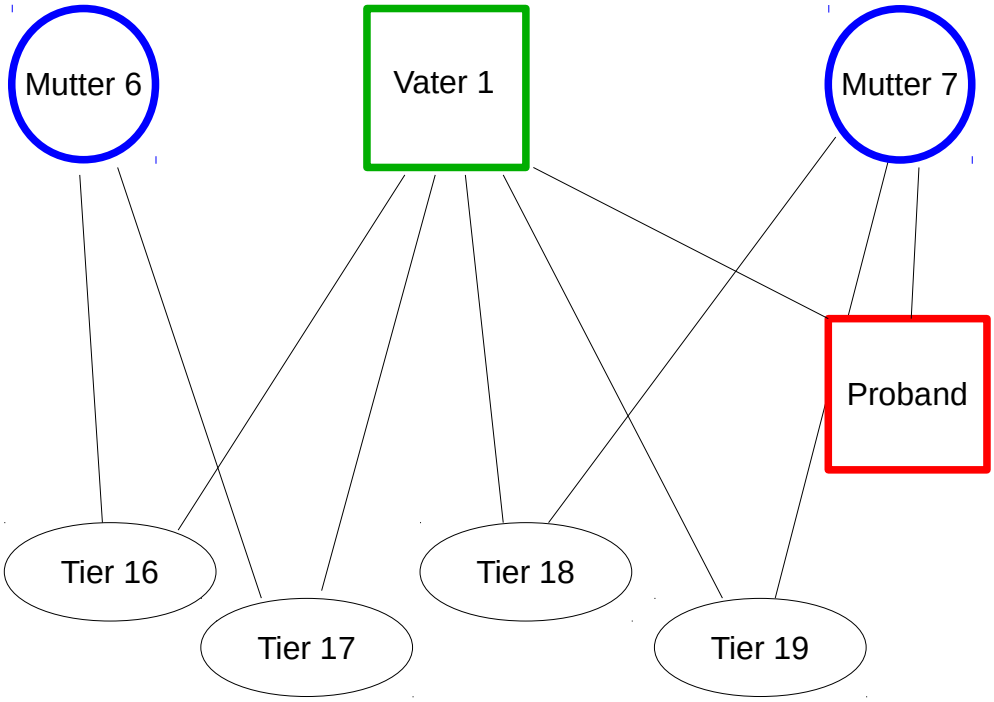

Abbildung 3.2: Informationsquellen bei Voll- und Halbgeschwister

3.4.3 Vollgeschwister

Falls wir den Zuchtwert für einen Probanden aus der Nachkommengeneration schätzen wollen, dann können seine Vollgeschwister und seine Halbgeschwister Informationen liefern. Die Abbilung 3.2 gibt eine Übersicht über die Verwandtschaftsstruktur zwischen dem Proband und den Informationsquellen. Der Proband hat die Tiere 18 und 19 als Vollgeschwister und die Tiere 16 und 17 sind Halbgeschwister zum Probanden. Der Proband selber hat keine Beobachtung.

Für die Schätzung des Zuchtwerts des Probanden mit der Indexgleichung müssen wir für die aktuelle Situation der verfügbaren Informationsquellen in die Matrizen P und G umsetzen. Schematisch sieht die Matrix P, wie folgt aus

$$P = \begin{bmatrix} var_P(\bar{VG}) & cov_P(HG) \\ cov_P(HG) & var_P(\bar{VG}) \end{bmatrix}$$

Die Varianz $var_P(\bar{VG})$ eines Vollgeschwisterdurchschnitts berechnen wir gemäss der schon bekannten Formel (3.18) berechnet, wobei bei Vollgeschwistern $t = h^2/2$. Somit ist

$$var_P(\bar{VG}) = \frac{1 + (n-1)\ h^2/2}{n}\ \sigma_x^2$$

Zwischen den Vollgeschwistergruppen sind alle Tiere Halbgeschwister. Somit entspricht die Kovarianz zwischen den Informationen der genetischen Kovarianz von Halbgeschwistern. Diese beträgt,

$$cov_P(HG) = \frac{1}{4}h^2\sigma_x^2$$

Somit können wir P aufstellen als

$$P = \begin{bmatrix} \frac{1+(n-1)\ h^2/2}{n}\sigma_x^2 & \frac{1}{4}h^2\sigma_x^2 \\ \frac{1}{4}h^2\sigma_x^2 & \frac{1+(n-1)\ h^2/2}{n}\ \sigma_x^2 \end{bmatrix}$$

Die rechte Seite stellt die genetischen Beziehungen zwischen den Informanten und dem Probanden. Die Tiere 16 und 17 sind Halbgeschwister zum Probanden, deshalb beträgt die genetische Kovarianz $r_{1\alpha} = 1/4\ \sigma_u^2$. Die Tiere 18 und 19 sind Vollgeschwister zum Probanden. Somit beträgt $r_{2\alpha} = 1/2\ \sigma_u^2$. Einsetzen der bis jetzt gefundenen Beziehungen in die Indexgleichung und Division beider Seiten der Gleichung durch σ_x^2 führt zu

$$\begin{bmatrix} \frac{1+(n-1)\ h^2/2}{n} & \frac{1}{4}h^2 \\ \frac{1}{4}h^2 & \frac{1+(n-1)\ h^2/2}{n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1/4\ h^2 \\ 1/2\ h^2 \end{bmatrix}$$

Setzen wir dafür die Zahlen ein, dann erhalten wir die folgende Beziehung

$$\begin{bmatrix} 0.5625 & 0.0625 \\ 0.0625 & 0.5625 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0.0625 \\ 0.1250 \end{bmatrix}$$

Als Lösung erhalten wir

$$b = \begin{bmatrix} 0.0875 \\ 0.2125 \end{bmatrix}$$

Die beiden Vollgeschwister zum Probanden erhalten erwartungsgemäss ein höheres Gewicht als die beiden Halbgeschwister. Unter angenommenen Leistungsabweichungen von -0.035 und 0.095 ergibt sich der geschätzte Zuchtwert (I) für den Probanden als

$$I = b^T * x = \begin{bmatrix} 0.0875 & 0.2125 \end{bmatrix} \begin{bmatrix} -0.04 \\ 0.10 \end{bmatrix} = 0.017125$$

Der geschätzte Zuchtwert des Probanden ist beträchtlich tiefer als die gegebenen Leistungsabweichungen. Dies wird durch die Gewichtung der Informationsquellen verursacht.

3.4.4 Vollgeschwister und Eigenleistung

In diesem Abschnitt gehen wir davon aus, dass der Proband zusätzlich zu den Informationen der Voll- und der Halbgeschwister auch noch eine Eigenleistung aufweist. Somit erweitern sich die Indexgleichungen um die folgenden Elemente. Die schematische Darstellung der erweiterten Matrix P sieht, wie folgt aus

$$P = \begin{bmatrix} var_P(\bar{VG}) & cov_P(HG) & cov_P(HG) \\ cov_P(HG) & var_P(\bar{VG}) & cov_P(VG) \\ cov_P(HG) & cov_P(VG) & var_p \end{bmatrix}$$

In dieser erweiterten Matrix ist das Element $cov_P(VG)$ neu dazugekommen. Das ist die Kovarianz zwischen der Eigenleistung und der durchschnittlichen Leistung der Vollgeschwister. Diese Kovarianz entspricht der genetischen Kovarianz zwischen dem Probanden und seinen Vollgeschwistern und beträgt somit die Hälfte der genetisch-additiven Varianz. Das Einsetzen der Formeln in die erweiterten Formen der Matrizen P und G führt zu folgendem Ergebnis

$$P = \begin{bmatrix} \frac{1+(n-1)\ h^2/2}{n}\sigma_x^2 & \frac{1}{4}h^2\sigma_x^2 & \frac{1}{4}h^2\sigma_x^2 \\ \frac{1}{4}h^2\sigma_x^2 & \frac{1+(n-1)\ h^2/2}{n}\ \sigma_x^2 & \frac{1}{2}h^2\sigma_x^2 \\ \frac{1}{4}h^2\sigma_x^2 & \frac{1}{2}h^2\sigma_x^2 & \sigma_x^2 \end{bmatrix}$$

und

$$G = \begin{bmatrix} 1/4\ h^2\sigma_x^2 \\ 1/2\ h^2\sigma_x^2 \\ h^2\sigma_x^2 \end{bmatrix}$$

Die numerische Lösung soll Teil einer Übung sein.

3.5 Sicherheit der geschätzten Zuchtwerte

Mittlerweilen haben wir verschiedene Zuchtwerte für den Probanden im Zahlenbeispiel berechnet. Es stellt sich nun die Frage, welcher der Schätzwerte ist genauer. Die Genauigkeit der geschätzten Zuchtwerte wurde schon in (3.14) definiert. Für komplexere Indices braucht es aber noch ein paar Elemente mehr. Ausgangspunkt bietet die Indexgleichung

$$Pb = Ga \tag{3.24}$$

Der Index entspricht

$$I = b^T * x \tag{3.25}$$

Die Varianz-Kovarianz-Matrix der Informationsquellen (x) entspricht der Matrix P und somit gilt

$$var(I) = \sigma_I^2 = b^T * P * b \tag{3.26}$$

Die Varianz ($var(T)$) des Zuchtziels (T) entspricht nicht der Matrix G, sondern dafür müssen wir eine neue Matrix (C) definieren. Es gilt also

$$C = var(u)$$

die genetische Varianz-Kovarianz-Matrix aller Merkmale im Zuchtziel. Die Varianz des Zuchtziels enspricht somit

$$var(T) = \sigma_T^2 = a^T * C * a \tag{3.27}$$

Daraus folgt das Bestimmtheitsmass (B) oder die Genauigkeit (r_{TI}) als

$$B = r_{TI}^2 = \frac{cov(T,I)^2}{var(T)} = \frac{\sigma_I^2}{\sigma_T^2} = \frac{b^T P b}{a^T C a} \tag{3.28}$$

Für die bis anhin geschätzten Zuchtwerte gilt: $C = \sigma_u^2 = h^2\sigma_x^2$, da wir nur ein Merkmal berücksichtigt haben. Da wir im Zuchtziel nur ein Merkmal haben, ist auch $a = 1$. Somit lauten die Genauigkeiten

Methode	Genauigkeit
ohne Eigenleistung	0.128
mit Eigenleistung	0.324

Die zusätzliche Berücksichtigung bringt somit einen beträchtlichen Anstieg des Bestimmtheismass.

3.6 Selektionserfolg

Wie schon früher gesehen wird der Selektionserfolg ΔG berechnet als

$$\Delta G = i * r_{TI} * \sigma_T \tag{3.29}$$

Im Fall eines einzelnen Merkmals im Zuchtziel ist $\sigma_T = \sigma_u$ und weiter gilt $r_{TI} = \sigma_I / \sigma_T$. Damit folgt

$$\Delta G = i * \sigma_I \tag{3.30}$$

Somit hängt der Selektion bei der Indexselektion nur von der Selektionsintensität und der Streuung des Index ab. Diese Streuung ist aber eine Funktion der Genauigkeit der Zuchtwertschätzung. Je ungenauer die Zuchtwerte geschätzt werden, desto kleiner wird die Streung des Index. Die Argumentation kann auch umgedreht werden. Der Selektionserfolg bei der Indexselektion hängt nur von der Genauigkeit der Zuchtwertschätzung ab.

3.7 Selektionserfolg in Einzelmerkmalen

Der nach der Gleichung (3.29) berechnete Selektionserfolg ergibt den Fortschritt auf der Basis des Gesamtzuchtwerts also in einer monetären Einheit. In der Praxis besteht aber oft ein Interesse am Selektionserfolg in den Einzelmerkmalen. Die Selektionsfortschritte der einzelnen Merkmale berechnen sich als

$$\Delta g = b^T \Delta G$$

wobei Δg ein Vektor mit den Selektionserfolgen der einzelnen Merkmale ist.

3.8 Anhang 1: Least Squares Lösungen für die Betriebe

Die Least Squares Lösungen der Betriebe werden aus dem einfachen Regressionsmodell mit dem Betrieb als einzigen fixen Effekt für die Beobachtungen. Das Modell für eine Beobachtung y_{ij} lautet somit

$$y_{ij} = \beta_j + \epsilon_{ij} \tag{3.31}$$

Die für das Relativieren gebrauchten Least Squares Lösungen der Betriebe erhalten wir mit der Funktion `lm()` in R. Dazu müssen wir zuerst einen dataframe vorbereiten, welche die Zunahmen der Nachkommen und die zugehörigen Betriebe enthält. Wir nennen den dataframe `tbl_beef_farm` und dieser sieht als Tabelle, wie folgt aus.

Betrieb	Zunahme
1	1.26
1	1.32
1	1.40
1	1.44
2	1.52
2	1.50
2	1.42
2	1.46
1	1.34
1	1.32
1	1.24
1	1.28
2	1.44
2	1.40
2	1.54
2	1.56

Bevor wir das Modell in `lm` angeben ist es wichtig, dass wir sicherstellen, dass die Variable Betrieb im dataframe `tbl_beef_farm` auch wirklich den Datentyp `factor` hat. Dies können wir mit folgendem Statement überprüfen.

```
is.factor(tbl_beef_farm$Betrieb)
```

```
## [1] TRUE
```

Falls die oben gezeigte Überprüfung nicht das Resultat `TRUE` zurückgibt, dann müssen wir die Variable Betrieb in den Datentypen `factor` umwandeln mit dem Befehl

```
tbl_beef_farm$Betrieb <- as.factor(tbl_beef_farm$Betrieb)
```

Das Modell wird dann spezifiziert mit

```
lm_beef_farm <- lm(Zunahme ~ 0 + Betrieb, data = tbl_beef_farm)
```

Die Resultate der Anpassung des Regressionsmodells erhalten wir mit der `summary()`-Funktion

```
summary(lm_beef_farm)
```

```
##
## Call:
## lm(formula = Zunahme ~ 0 + Betrieb, data = tbl_beef_farm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08500 -0.04875 -0.00500  0.04500  0.11500
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## Betrieb1  1.32500    0.02248   58.94   <2e-16 ***
## Betrieb2  1.48000    0.02248   65.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06358 on 14 degrees of freedom
## Multiple R-squared:  0.9982, Adjusted R-squared:  0.998
## F-statistic:  3904 on 2 and 14 DF,  p-value: < 2.2e-16
```

Die Least Squares Lösungen für die Betriebe erhalten wir mit dem Statement

```
coefficients(lm_beef_farm)
```

```
## Betrieb1 Betrieb2
##    1.325    1.480
```

Kapitel 4

BLUP Zuchtwertschätzung

Bei der Zuchtwertschätzung aufgrund des Selektionsindexes (siehe Abschnitt 3.4) besteht die Problematik der Bestimmung des Vergleichswertes mit welchem die beobachteten Phänotypen relativiert werden sollen. Diese Problematik war schon lange bekannt, aber erst die Einführung der Zuchtwertschätzung anhand des BLUP-Verfahrens brachte eine Lösung. Zunächst wollen wir die Problematik noch etwas genauer beschreiben.

4.1 Problematik des Vergleichswertes

In der Theorie ist das Populationsmittel (μ) der ideale Vergleichswert, da der mittlere Zuchtwert ($\bar{G}$) über eine Population gleich Null ist. Aus dem verwendeten genetischen Standardmodell ((4.1)) lässt sich die folgende Beziehung ableiten.

$$P \;=\; G \;+\; E \rightarrow \bar{P} = \bar{G} + \bar{E} = \bar{E} = \mu \tag{4.1}$$

Das Populationsmittel misst also den mittleren Einfluss der Umwelt (E). Streng genommen gilt das aber nur in einer idealisierten Population, in welcher alle Selektionskandidaten und alle Leistungen der anderen Informationsquellen zur selben Zeit erbracht werden. In der Praxis ist das aber nicht realistisch, weil bei der Kombination von Informationen aus verschiedenen Generationen (z. Bsp. Eigenleistung und Nachkommenleistungen) werden diese Leistungen sicher nicht gleichzeitig erbracht. Ausserdem werden Populationen in der Praxis an verschiedenen Standorten auf verschiedenen Betrieben gehalten. Diese Unterschiede haben einen erheblichen Einfluss auf die Leistung. Als Beispiel dafür ist in Abbildung 4.1[1] die Milchproduktion im Berg- und Talgebiet über verschiedene Betriebsgrössen in der Schweiz dargestellt.

Alle die erwähnten Faktoren können als Umweltfaktoren zusammengefasst werden, welche die genetischen Unterschiede zwischen den Tieren verzerren. Geschätzte Zuchtwerte sollten aber nicht von den Umweltfaktoren abhängig sein. Deshalb hat man schon früh gleiche Umweltbedingungen wie z. Bsp. in der Milchviehhaltung Herde, Kalbejahr, Kalbesaison und das Erstkalbealter zu Klassen von systematisierbaren **Umwelteinflüssen** zusammengefasst. Neben Umwelteinflüssen gibt es auch noch andere Faktoren, wie zum Beispiel Laktationsnummer oder Alter, welche nichts mit der Umwelt zu tun haben, aber trotzdem einen Einfluss auf die Leistung haben. Diese sollten eigentlich besser als systematisierbare **fixe Effekte** bezeichnet werden. Vereinfachend werden hier alle systematisierbaren Einflussfaktoren auf die Leistung als fixe Effekte bezeichnet. Um Verzerrungen der geschätzten Zuchtwerte zu vermeiden, unterteilt man die Population in Vergleichsgruppen, die in allen fixen Effekten der gleichen Klasse angehören.

Je mehr fixe Effekte in der Zuchtwertschätzung berücksichtigt werden und je feiner die Einteilung in Vergleichsgruppen ist, desto besser können Einflüsse der fixen Effekte auf die geschätzten Zuchtwerte berücksichtigt werden. Auf der anderen Seite führt eine sehr feine Einteilung der Population in sehr viele Ver-

[1] Grafik von https://www.swissmilk.ch/de/produzenten/milchmarkt/marktakteure-strukturen/milchproduzenten/

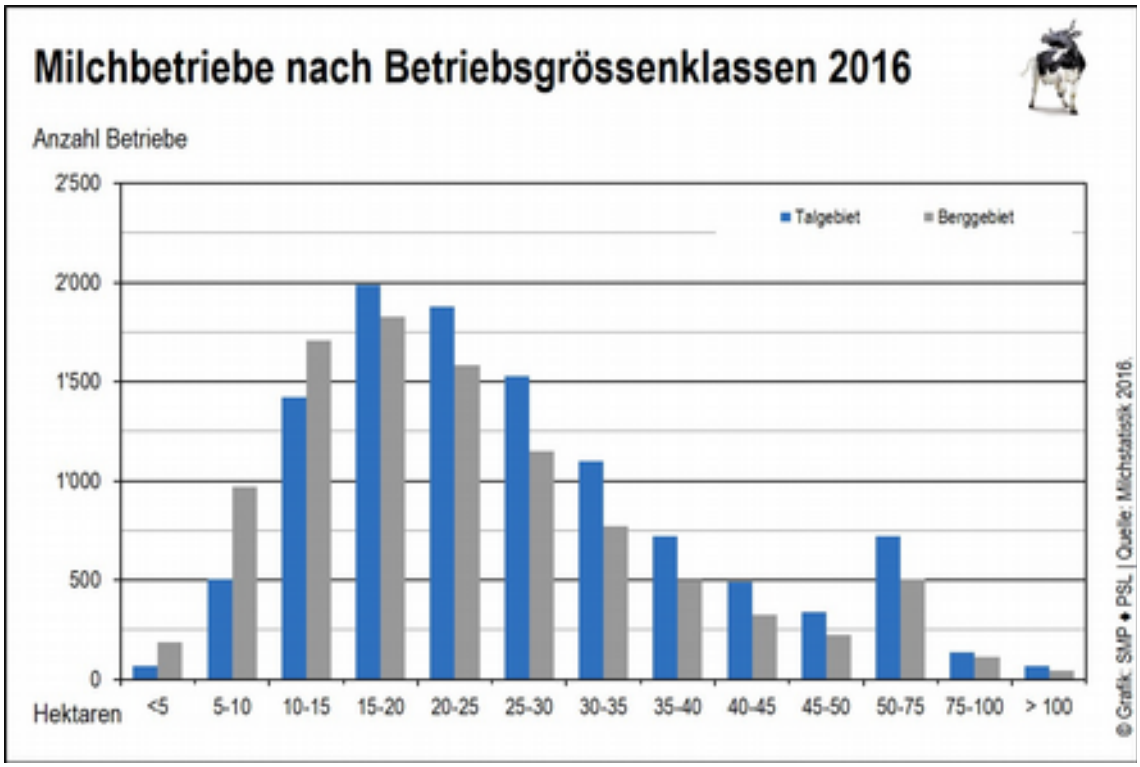

Abbildung 4.1: Milchproduktion im Berg- und Talgebiet

gleichsgruppen zu einer schwachen Gruppenbesetzung, d.h. die Anzahl Tiere in einer Gruppe kann sehr tief sein. Dies hat statistiche Konsequenzen, welche nicht wünschenswert sind. Diese Konsequenzen werden als **Verzerrungen** bezeichnet und werden im nächsten Abschnitt beschrieben.

4.1.1 Verzerrung der geschätzten Zuchtwerte

Mit schwächer werdenden Vergleichsgruppen steigt die Wahrscheinlichkeit, dass der Mittelwert der Zuchtwerte nicht mehr bei Null liegt. Für die Verteilung des Mittelwerts ($\bar{u}$) der Zuchtwerte einer Gruppe von n zufällig ausgewählten Vergleichstiere aus der Population gilt

$$\bar{u} \sim \mathcal{N}(0, \frac{\sigma_u^2}{n}) \tag{4.2}$$

Verglichen mit der phänotypischen Varianz (σ_x^2) ist additiv-genetische Varianz σ_u^2 in der Regel klein. Damit besteht zwar nur eine kleine Wahrscheinlichkeit, dass der mittlere Zuchtwert in einer Vergleichsgruppe von Null abweicht. Sinkt aber die Gruppengrösse sehr stark ab, dann steigt dieses Risiko trotzdem wieder an.

Wenn der mittlere geschätzte Zuchtwert in einer Vergleichsgruppe nicht mehr Null ist, dann tritt das Phänomen der **Verzerrung** (auch "Bias" genannt) auf. Die Auswirkung der Verzerrung können anhand der folgenden kurzen Herleitung gezeigt werden. Betrachten wir zunächst den Mittelwert $\bar{x}_{VT}$ der phänotypischen Leistungen in einer Vergleichsgruppe (VT)

$$\bar{x}_{VT} = \bar{P}_{VT} = \bar{G}_{VT} + \bar{U}_{VT} \tag{4.3}$$

Dabei wird angenommen, dass der mittlere Zuchtwert $\bar{G}_{VT} = 0$ ist und dass der mittlere Umwelteffekt den Einfluss aller fixen Effekte misst. Die Auswirkungen auf den geschätzten Zuchtwert I, wenn $\bar{G}_{VT} \neq 0$ ist, sehen wir in der folgenden Gleichungssequenz

$$\begin{aligned} I &= b * (x - \bar{x}_{VT}) \\ &= b * (x - (\bar{G}_{VT} + \bar{U}_{VT})) \\ &= b * (x - \bar{U}_{VT}) - b * \bar{G}_{VT} \\ &= Z\hat{W} - b * \bar{G}_{VT} \end{aligned} \tag{4.4}$$

Der Index (I) schätzt also nicht den Zuchtwert ($Z\hat{W}$) des Tieres sondern etwas anderes. Der Term $b * \bar{G}_{VT}$ bezeichnet man als Verzerrung. Somit ist die Verzerrung abhängig vom mittleren geschätzten Zuchtwert der Vergleichstiere.

4.2 Lösung der Vergleichsproblematik mit BLUP

Wie im Abschnitt 4.1 gezeigt wurde, stellt die Vergleichsproblematik einen so genannten "Teufelskreis" dar. Für die Berechnung eines Vergleichswertes, der eine Schätzung von unverzerrten ("un-biased") Zuchtwerten erlaubt, müssten die Zuchtwerte der Vergleichstiere bekannt sein. Für die Schätzung der Zuchtwerte der Vergleichstiere brauchen wir aber wieder die Vergleichswerte.

Erst die Arbeiten von Charles R. Henderson und seinem Team zwischen 1949 und 1973, welche zur Entwicklung des BLUP-Verfahrens für die Zuchtwertschätzung führten, konnten das Problem nachhaltig lösen. Dabei werden im BLUP-Verfahren die fixen Effekte und die Zuchtwerte simultan geschätzt. Neben der Entwicklung des BLUP-Verfahrens für die Zuchtwertschätzung hat Henderson und seine Mitarbeiter auch Rechenvorschriften (Algorithmen) vorgeschlagen, welche eine sehr effiziente Umsetzung des BLUP-Verfahrens in der praktischen Zuchtwertschätzung erlauben und somit der ganzen Methodik zum Durchbruch verholfen hat.

Aus Sicht der Statistik besteht das BLUP-Verfahren in der Anwendung von linearen gemischten Modellen zur Auswertung der Daten aus der Tierzucht. Lineare gemischte Modelle sind eine Unterklasse der linearen Modelle, welche neben den fixen Effekten auch zufällige Effekte (abgesehen von den zufälligen Resteffekten) im Modell mit einschliesst. Wie diese Effekte unterschieden werden, beschreibt der nächste Abschnitt.

4.3 Fixe und zufällige Effekte

Die Unterscheidung in fixe und zufällige Effekte basiert nicht auf einer soliden universell gültigen Definition und ist oft auch Gegenstand von Diskussionen. In der praktischen Zuchtwertschätzung ist die Einteilung fixe und zufällige Effekte auch von den verwendeten Softwareprogrammen abhängig. Ein Beispiel dafür sind die Herde-Jahr-Saison-Effekten, welche man aufgrund vom gängigen Verständnis als fixe Effekte auffassen würde. Diese Effekte haben sehr viele Effektstufen und da gewisse Programme damit nicht umgehen können, werden diese Effekte als zufällige Effekte modelliert.

Allgein lassen sich folgende Unterscheidungskriterien zwischen fixen und zufälligen Effekten auflisten.

fixer Effekt	zufälliger Effekt
Klassen sind exakt und reproduzierbar definiert	Realisierungen stammen aus einer Verteilung, welche durch Erwartungswert und Varianz definiert ist
Für den Wert einer Klasse kann a priori kein Erwartungswert angegeben werden	Phänotypische Beobachtungen werden durch Varianz der zufälligen Effekts beeinflusst
In der Statistik interessiert der Mittelwert der Effektklassen	In der Statistik interessiert die Varianz aller Realisierungen, nicht der Mittelwert
Fixe Effekte können für die Korrektur von Beobachtungen verwendet werden	

Bestimmte Effekte lassen sich eindeutig zuordnen. So sind Geschlecht, Rasse oder Betrieb sicher Effekte, welche aufgrund der genannten Unterscheidungskriterien als fix eingestuft werden können. Auf der anderen Seite ist auch klar, dass Zuchtwerte als zufällige Faktoren modelliert werden. Dabei wissen wir schon aufgrund der Annahmen aus dem grundlegenden genetischen Modells (siehe Gleichung (4.1)), dass der Erwartungswert der Zuchtwerte gleich Null ist und die Streuung der Zuchtwerte der additiv-genetischen Varianz entspricht.

4.4 Lineares gemischtes Modell

In einem linearen gemischten Modell gibt es neben dem zufälligen Resteffekt noch andere zufällige Effekte. In unserem Zahlenbeispiel werden die Betriebe als fixe Effekte und die Väter als zufällige Effekte modelliert. Für eine einzelne Beobachtung (y_{ijk}) lässt sich das lineare gemischte Modell, wie folgt notieren.

$$y_{ijk} = \beta_i + u_j + e_{ijk} \tag{4.5}$$

dabei steht β_i für den fixen Effekt des i-ten Betriebs, u_j der zufällige Zuchtwert des j-ten Vaters und e_{ijk} für den zufälligen Resteffekt des k-ten Nachkommen des j-ten Vaters auf dem i-ten Betrieb. Notiert man diese Modellformulierung für jede Beobachtung und verwandelt diese in Matrix-Vektor-Schreibweise, so erhalten wir das folgende Modell

$$y = X\beta + Zu + e \tag{4.6}$$

wobei

- y Vektor der Länge n mit Beobachtungen
- b Vektor der Länge p mit Klassen der fixen Effekte
- X $n \times p$ - Designmatrix als Verknüpfung zwischen fixen Effekten und Beobachtungen
- u Vektor der Länge q mit Klassen der zufälligen Effekte
- Z $n \times q$ - Designmatrix als Verknüpfung zwischen zufälligen Effekten und Beobachtungen
- e Vektor der Länge n mit zufälligen Resteffekten

In der Anwendung des linearen gemischten Modells in der Zuchtwertschätzung werden alle fix klassierbaren Effekte (Umwelt, etc.) als fixe Effekte (β) modelliert und die Zuchtwerte werden als zufällige Effekte (u) im Modell berücksichtigt. Das Resultat der Lösung des linearen gemischten Modells beinhaltet Schätzer ($\hat{\beta}$) für die fixen Effekte und geschätzte Zuchtwerte ($\hat{u}$). Im Gegensatz um Selektionsindex müssen wir beim linearen gemischten Modell nicht zuerst die Beobachtungen relativieren und die relativierten Beobachtungen dann gewichten. Die Schätzung der fixen und der zufälligen Effekte geschieht im BLUP-Verfahren in einem Schritt.

4.5 Vatermodell

Wir wenden das lineare gemischte Modell nun für unser Zahlenbeispiel an. Zuerst betrachten wir den analogen Fall zur Schätzung der Zuchtwerte für die drei Väter anhand der Nachkommenleistungen. Das resultierende

Modell wurde anfangs bei der Einführung des BLUP-Verfahrens sehr häufig eingesetzt und bei einzelnen Auswertungen (z. Bsp. bei der Zuchtwertschätzung Geburtsverlauf beim Milchvieh in der Schweiz) wird dieses Modell immer noch verwendet. Dieses Modell wird als **Vatermodell** bezeichnet und ist in der folgenden Gleichung für unser Zahlenbeispiel gezeigt.

$$\begin{bmatrix} 1.26 \\ 1.32 \\ 1.40 \\ 1.44 \\ 1.52 \\ 1.50 \\ 1.42 \\ 1.46 \\ 1.34 \\ 1.32 \\ 1.24 \\ 1.28 \\ 1.44 \\ 1.40 \\ 1.54 \\ 1.56 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \end{bmatrix}$$

Wobei β_1 und β_2 die unbekannten Effekte der beiden Betriebe und u_1, u_2 und u_3 die ebenfalls unbekannten Zuchtwerte der drei Väter darstellen. Die obigen Gleichungssystem ausformulierten Designmatrizen X und Z zeigen, wie die Beobachtungen zu den fixen Effekten und den Zuchtwerten verknüpft werden.

Da die zufälligen Effekte u und e als Abweichungen definiert sind, sind deren Erwartungswerte Null, somit gilt, dass

$$E(u) = 0 \text{ und } E(e) = 0$$

Somit gilt

$$E(y) = E(X\beta + Zu + e) = X\beta$$

In einem linearen gemischten Modell spielen neben den Erwartungswerten auch die Varianzen eine Rolle. Dabei gilt es zu beachten, dass die fixen Effekte (β) keinen Beitrag zur Varianz leisten. Bei der Betrachtung der Varianzen beginnen wir bei den Resteffekten (e). Im einfachsten Fall, wenn keine besondere Struktur, wie wiederholte Messungen, in den Daten vorhanden sind, dann können wir die Varianz der Resteffekte einfach, wie folgt definieren.

$$var(e) = I * \sigma_e^2 = R$$

Für die Zuchtwerte (u) ist die Varianz ($var(u)$) als Funktion der genetisch-additiven Verwandtschaftsmatrix definiert.

$$var(u) = A * \sigma_u^2 = G$$

Beim Vatermodell gilt es zu beachten, dass nur die Verwandtschaft zwischen den Vätern berücksichtigt werden. Für unser Zahlenbeispiel sind die Väter untereinander nicht verwandt und somit gilt

$$G = I * \sigma_u^2/4$$

Die Varianz ($var(y)$) der Beobachtungen können wir nun zusammensetzen. Damit erhalten wir

$$var(y) = var(X\beta + Zu + e) = var(Zu + e) = ZGZ^T + R = V$$

wobei wir annehmen, dass $cov(u, e) = 0$

In der praktischen Zuchtwertschätzung ist die Kovarianz zwischen den Beobachtungen (y) und den Zuchtwerten (u) oft von Bedeutung. Diese kann wie folgt berechnet werden.

$$\begin{aligned} cov(y, u) &= cov(Zu + e, u) && (4.7)\\ &= cov(Zu, u) + cov(e, u) && (4.8)\\ &= Zcov(u, u) && (4.9)\\ &= ZG && (4.10) \end{aligned}$$

4.5.1 Schätzung der unbekannten Effekte

Bis hierher haben wir die Eigenschaften des linearen gemischten Modells für die Anwendung der Zuchtwertschätzung beschrieben. Nun geht es darum für die unbekannten Effekte β und u Schätzwerte aus den Daten zu ermitteln. Die ausführliche Herleitung der Schätzwerte würde den Rahmen dieser Veranstaltung sprengen. Deshalb gehen wir direkt zu den Ergebnissen der Arbeiten von Henderson und seinem Team über.

Die mit dem BLUP-Verfahren ermittelten Schätzer für die Zuchtwerte (u) betragen

$$\hat{u} = BLUP(u) = GZ^T V^{-1}(y - X\hat{\beta}) \tag{4.11}$$

wobei $\hat{\beta}$ dem Schätzer der fixen Effekte entspricht. Dieser Schätzer entspricht dem allgemeinen Least-Squares-Schätzer (GLS) der fixen Effekte. Somit gilt

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{4.12}$$

Die Matrix V enthält hier auch die genetischen Kovarianzen.

4.5.2 Eigenschaften von BLUP

Die Methoden, welche auf dem BLUP-Verfahren (auch kurz BLUP-Methoden genannt) haben wichtige Eigenschaften und diese sind in der Abkürzung BLUP versteckt. Die Abkürzung BLUP steht nämlich für **B**est **L**inear **U**nbiased **P**rediction. Die einzelnen Bestandteile haben die folgende Bedeutung

- *Best*: unter allen lineare Schätzern ist der Schätzer aus BLUP der "beste" im Sinne, dass er die tiefste Fehlervarianz ($var(u - \hat{u})$) hat. Dies ist gleichbedeutend mit der Eigenschaft, dass die Korrelation zwischen wahrem Effekt (u) und dem Schätzer ($\hat{u}$) maximal ist.
- *Linear*: der Schätzer ist eine lineare Funktion der Beobachtungen.
- *Unbiased*: die Schätzwerte sind unverzerrt, das heisst es gilt $E(u) = E(\hat{u})$.
- *Prediction*: die Unterscheidung zwischen Vorhersage und Schätzung wird im deutschen Sprachgebrauch eigentlich nicht gemacht. Deshalb sprechen wir hier allgemein von Schätzung oder Schätzwert.

Bei der Zuchtwertschätzung mit der BLUP-Methode nehmen wir an, dass die Varianzkomponenten fehlerfrei bekannt sind. Dies ist in der Praxis nicht der Fall. Die Varianzkomponenten sind unbekannt und müssen auch aus den Daten geschätzt werden.

4.6 Hendersons Mischmodellgleichungen

Die Formeln zur Berechnung der Schätzer für die Zuchtwerte ((4.11)) und die fixen Effekte ((4.12)) beinhalten beide die Inverse (V^{-1}) der Kovarianzmatrix V der phänotypischen Beobachtungen. In den praktischen Zuchtwertschätzungen hat diese Matrix eine Dimension in der Grössenordnung von $10^6 \times 10^6$. Die Inversion einer solch grossen Matrix ist nicht praktikabel. Als Lösung diese Problems entwickelte Henderson die sogenannten **Mischmodellgleichungen**, welche die Inversion von V vermied. Für das verwendete Zahlenbeispiel sehen die Mischmodellgleichungen, wie folgt aus

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{4.13}$$

Sofern die Matrix R eine einfache Struktur hat und deshalb gilt, dass $R = I * \sigma_e^2$ können die Mischmodellgleichungen vereinfacht geschrieben werden als

$$\begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + G^{-1} * \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \end{bmatrix} \tag{4.14}$$

Die neu eingeführte Variable λ enspricht dem Varianzverhältnis $\lambda = \sigma_e^2 / \sigma_u^2$.

4.7 Zuchtwertschätzung mit dem BLUP-Tiermodell

Die Mischmodellgleichungen stellen ein allgemeines Werkzeug für Fragestellungen, welche mit gemischten linearen Modellen bearbeitet werden können. Die bisherige Verwendung der Mischmodellgleichungen und der BLUP-Methode ermöglichte uns die Schätzung der Zuchtwerte der Väter. Es ist allerdings unbefriedigend, dass weder die Mütter noch die Nachkommen einen geschätzten Zuchtwert erhielten. Für die praktische Anwendung in der Zuchtarbeit wären aber geschätzte Zuchtwerte für diese Tiere als Selektionskriterium auch wünschenswert.

Genau dieses Bedürfnis, dass alle Tiere in einem Pedigree (Stammbaum, welcher die Abstammungen der Tiere aufzeigt) geschätzte Zuchtwerte erhalten, wird mit dem BLUP-Tiermodell befriedigt. Das statistische Modell sieht gleich aus, wie das Modell des Vatermodells (siehe Gleichung (4.6)). Unterschiedlich ist allerdings der Vektor (u) der Zuchtwerte und die Matrix Z. Der Vektor der Zuchtwerte enhält nun Zuchtwerte aller Tiere im Pedigree.

Durch die Veränderung des Vektors u ändert sich auch die Varianz ($var(u)$) der Zuchtwerte. Diese enspricht nun der vollständigen genetisch-additiven Verwandschaftsmatrix A mal der genetisch-additiven Varianz.

$$G = A * \sigma_u^2$$

Die Schätzwerte der fixen Effekte und der Zuchtwerte erhalten wir auch aus den Mischmodellgleichungen.