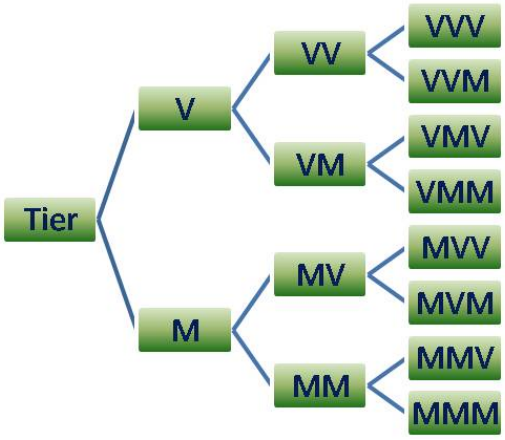
QUALITAS

Zusammenfassung

Deregression von geschätzen Zuchtwerten gemäss GTF2009



Peter von Rohr
FB EDV, Qualitas AG
Chamerstrasse 56, 6300 Zug
http://www.qualitasag.ch
peter.vonrohr@qualitasag.ch

Contents

Dokumentenstatus

Version	Datum	Wer	Änderung
0.0.0.900		pvr	Erstellung
0.0.0.901	16.03.2016	pvr	Ideales Modell, Zuchtwert-Modell
0.0.0.902	17.03.2016	pvr	Erklärungen zur Varianzreduktion, Start des Beweises, Deregression
0.0.0.903	29.03.2016	pvr	Gewichtung der deregressierten Zuchtwerte, Entfernen der Elterninformation
0.0.0.904	12.04.2016	pvr	Anhang mit Beweis abgeschlossen

Abkürzungen

Abkürzung	Bedeutung
PE	prediction error
PEV	prediction error variance
PA	parental average
EBV	estimated breeding value

Erklärung

Dieses Dokument gibt eine Zusammenfassung vom Paper von Garrick, Taylor, and Fernando (2009). Der Fokus der Zusammenfassung liegt auf dem Abschnitt der Deregression von geschätzten Zuchtwerten (EBV), dies deshalb, weil wir dafür ein direktes Interesse für eine Anwendung haben. Die anderen Abschnitte insbesondere das Material zum Hintergrund des Papers werden hier kurz gehalten.

In runden Klammern () werden die ursprünglichen englischen Begriffe aufgeführt, wo immer dies sinnvoll erscheint. Die geschweiften Klammern {} enthalten zusätzliche erklärende Informationen, welche nicht im Paper sind.

Hintergrund (Background)

Genomische Effektschätzung (Genomic prediction) verwendet geschätzte Regressionskoeffizienten von Leistungsdaten auf Genotypen einer dichten Markerkarte aus einer Trainingspopulation für die Vorhersage der genomischen Zuchtwerte in einer Zielpopulation. In der Trainingspopulation befinden sich genotypisierte Tiere mit sehr unterschiedlichen Informationen, wie z. Bsp. wiederholte Messungen, Nachkommeninformationen oder traditionell geschätzte Zuchtwerte.

In der Literatur gibt es keine einhellige Meinung, wie die verschiedenen Informationsquellen behandelt werden sollten. {Der Überblick im Paper über die verschiedenen Arten, wie die Informationen berücksichtigt werden, wird hier nicht angeführt}

Die Effektschätzung in der genomischen Selektion ist abhängig davon, welche Daten als abhängige Beobachtungen verwendet werden.

Methoden (Methods)

Ein ideales Modell

Idealerweise hätten wir für das Training wahre genotypische Werte (g) von unverwandte Individuen ohne Selektion. Dann würde das Modell so aussehen

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{Ma} + \epsilon \tag{1}$$

wobei: $\mathbf{g}$ dem Vektor mit wahren Zuchtwerten mit $var(\mathbf{g}) = \mathbf{T}\sigma_g^2$ entspricht und das Skalar σ_g^2 die genetische Varianz darstellt, wobei $\mathbf{T}$ aufgrund der Theorie zu Kopplung und Kopplungsungleichgewicht aufgestellt werden kann. Die Grösse μ stellt den Achsenabschnitt (intercept) dar. Die Inzidenz-Matrix $\mathbf{M}$ verbindet die wahren Zuchtwerte in $\mathbf{g}$ mit den entsprechenden Substitutionseffekten in $\mathbf{a}$. Die Varianz $var(\mathbf{Ma}) = \mathbf{G}\sigma_M^2$, wobei $\mathbf{G}$ der genomischen Verwandtschaftsmatrix entspricht. Der Vektor ϵ steht für die Residuen mit $var(\epsilon) = \mathbf{E}\sigma_\epsilon^2$.

{Die Diskussion, welche Effekte vom Modell in Gleichung (1) zufällig oder fix sind, und ob ein polygener Effekt im Modell berücksichtigt werden soll, wird hier nicht im Detail ausgeführt. Die Autoren stellen fest, dass die Varianz basierend auf den Markereffekten nicht die gesamte genetische Varianz erklären kann. Somit wird ein gewisser polygener Anteil im Effekt ϵ zu finden sein. Da die Individuen im Training-Set aber miteinander verwandt sind, können die ϵ Effekte nicht als unkorreliert modelliert werden. Als Folge davon entspricht die Matrix $\mathbf{E}$ nicht der Einheitsmatrix, sondern wird durch die additive genetische Verwandtschaftsmatrix $\mathbf{A}$ approximiert. Alternativ zur Verwendung von $\mathbf{A}$ können auch explizit korrelierte Residuen gefittet werden.

Die zufälligen Effekte (g, M und ϵ) des Modells in Gleichung (1) werden als unkorreliert angenommen. Daraus folgt, dass $\sigma_g^2 = \sigma_M^2 + \sigma_\epsilon^2$. Der Anteil der genetischen Varianz, welche nicht von den Markern erklärt wird ist somit

$$c = \frac{\sigma_\epsilon^2}{\sigma_g^2} = 1 - \frac{\sigma_M^2}{\sigma_g^2}$$

Die Abschnitte zu den Modellen mit phänotypischen Records, wiederholten Messungen und mit den Nachkommendurchschnitten werden hier vorerst nicht behandelt.}

Geschätzte Zuchtwerte als Trainingsdaten

Ein mit BLUP geschätzter Zuchtwert kann als Summe des wahren Zuchtwertes plus ein Schätzfehler (prediction error) betrachtet werden. Als Formel heisst das $\hat{\mathbf{g}} = \mathbf{g} + (\hat{\mathbf{g}} - \mathbf{g})$. Setzt man diese Beziehung in Gleichung (1) ein, dann folgt

$$\hat{\mathbf{g}} = \mathbf{g} + (\hat{\mathbf{g}} - \mathbf{g}) = \mathbf{1}\mu + \mathbf{Ma} + (\epsilon + (\hat{\mathbf{g}} - \mathbf{g})) \tag{2}$$

{Gleichung (2) hier in der Zusammenfassung entspricht Gleichung (8) im Paper.}

Die Verwendung von geschätzten Zuchtwerten für die Effektschätzung führt zu mindestens zwei Problemen, welche beide mit den Eigenschaften von BLUP-Schätzern zusammenhängen.

1. Problem: Reduktion der Varianz

{Diese Abschnittsgliederung und dieser Titel ist nicht im Paper, wird aber aus Gründen der Übersichtlichkeit hier eingefügt.}

Die Addition des Schätzfehlers in Gleichung (2) dazu führt, dass die Varianz reduziert und nicht erhöht wird, obwohl die Diagonalelemente von $var(\hat{\mathbf{g}} - \mathbf{g})$ positiv sein müssen, da es sich um Varianzen handelt.

{Der Punkt hier, welcher nicht offensichtlich klar ist, rührt daher, dass im allgemeinen, wenn zu einer bestimmten Zufallsvariablen X mit einer Varianz σ_X^2 eine zweite Zufallsvariable Y mit Varianz σ_Y^2 addiert, dann ist die Varianz $var(X+Y)$ der Summe von X und Y im allgemeinen grösser als die einzelnen Varianzen σ_X^2 und σ_Y^2, insbesondere, wenn X und Y als unkorreliert angenommen werden. Dieser Sachverhalt trifft aber bei der Situation mit den geschätzten Zuchtwerten als Beobachtungen nicht zu. Wenn wir zu einem wahren Zuchtwert (g_i) eines Tieres i den Schätzfehler ($\hat{g}_i - g_i$) addieren, dann ist die Varianz ($var(\hat{g}_i)$) der Summe aus wahren Zuchtwerten und Schätzfehlern kleiner als die Varianz der wahren Zuchtwerte ($var(g)).

Was hier nicht von Bedeutung ist, aber im Paper erwähnt wird, dass bei der Verwendung von phänotypischen Beobachtungen y_i die Varianz steigt, da im Vergleich zum idealen Modell eine nicht genetische Komponente hinzukommt. Als Formel heisst das, dass $var(g_i) > var(\hat{g}_i)$, während $var(g_i) < var(y_i)$, was durch die Schrumpfung (shrinkage) der BLUP-Schätzung verursacht wird.}

Im allgemeinen ist die Varianz einer Differenz wie folgt definiert: $var(\hat{g}_i - g_i) = var(g_i) + var(\hat{g}_i) - 2cov(\hat{g}_i, g_i)$. Aber für BLUP gilt $cov(\hat{g}_i, g_i) = var(\hat{g}_i)$ {der Beweis dafür ist im Anhang} und somit ist $var(\hat{g}_i - g_i) = var(g_i) - var(\hat{g}_i) > 0$ und somit ist $var(g_i) > var(\hat{g}_i)$.

Die Reduktion der Varianz der Trainingsdaten rührt von der negativen Korrelation zwischen wahren Zuchtwerten g_i und Schätzfehlern ($\hat{g}_i - g_i$), d.h. $cov(g_i, \hat{g}_i - g_i) = cov(g_i, \hat{g}_i) - var(g_i) = var(\hat{g}_i) - var(g_i) < 0$. Das heisst Tiere mit hohem wahren Zuchtwert haben tendenziell zu tiefe Schätzwerte, da ihre Schätzfehler negative sind und Tiere mit tiefem wahren Zuchtwert werden eher überschätzt, da deren Schätzfehler eher positiv sind. Dies ist eine Konsequenz der Schrumpfung (shrinkage) von BLUP und von der Tatsache, dass Schätzfehler und geschätzte Zuchtwerte unkorreliert sind, d.h. $cov(\hat{g}_i, \hat{g}_i - g_i) = var(\hat{g}_i) - cov(\hat{g}_i, g_i) = 0$.

Damit die Covarianz zwischen wahren Zuchtwerten und Schätzfehlern berücksichtig werden können, müsste ein komplexeres Modell als jenes von Gleichung (2) verwendet werden. Dies wäre auch rein rechnerisch viel aufwändiger. {Wie ein solches Modell aussehen würde und welchen rechnerischen Mehraufwand anfallen würde, wird im Paper nicht weiter erläutert.}

2. Problem: Schrumpfung (Shrinkage) der Schätzung zum Mittel

{Dieser Titel ist nicht im Paper, wird aber aus Gründen der Übersichtlichkeit hier eingefügt.}

Die mit BLUP geschätzten Zuchtwerte sind zum Mittel hin geschrumpft (shrinked), wobei das Ausmass der Schrumpfung vom Informationsgehalt (Genauigkeit) abhängig ist. Das wird offensichtlich begründet dadurch dass die Regression vom wahren Zuchtwert auf den Phänotyp gleich 1 ist und die Regression vom geschätzten Zuchtwert auf den wahren Zuchtwert ist gleich $r_i^2 < 1$, wobei r_i^2 der Genauigkeit des geschätzten Zuchtwertes entspricht und als das Quadrat der Korrelation zwischen wahrem und geschätztem Zuchtwert definiert ist. An einem bestimmten Markerlocus ist also die Differenz zwischen geschätzten Zuchtwerten von verschiedenen Genotypen kleiner, als das der Fall wäre, wenn wir wahre Zuchtwerte oder Phänotypen als Beobachtungen verwenden würden. Das Ausmass der Reduktion der Differenzen hängt von den Genauigkeiten r_i^2 der geschätzten Zuchtwerte ab. Da uns der Effekt von unterschiedlichen Markergenotypen auf die phänotypische Leistung interessiert und da Tiere in einem Trainingsdatensatz unterschiedlich genau geschätzte Zuchtwerte haben, müssen wir die geschätzten Zuchtwerte de-regressieren, bevor wir diese zur Effektschätzung verwenden können.

Deregression von geschätzten Zuchtwerten

Die Problem im Zusammenhang mit der Verwendung von geschätzten Zuchtwerten als Trainingsdaten können gelöst werden indem die geschätzten Zuchtwerte mit einer Diagonalmatrix $\mathbf{K}$ skaliert werden. Somit verwenden wir anstelle des Modells in Gleichung (2) das folgende Model für die genomische Effektschätzung

$$\mathbf{K}\hat{\mathbf{g}} = \mathbf{g} + (\mathbf{K}\hat{\mathbf{g}} - \mathbf{g}) = \mathbf{1}\mu + \mathbf{M}\mathbf{a} + (\epsilon + (\mathbf{K}\hat{\mathbf{g}} - \mathbf{g})) \tag{3}$$

für eine Matrix $\mathbf{K}$, welche so bestimmt wird, dass $cov(g_i, k_i\hat{g}_i - g_i) = 0$ und $cov(k_i\hat{g}_i, g_i) = \rho$, wobei ρ eine Konstante ist. Da die $cov(g_i, k_i\hat{g}_i - g_i) = k_i * var(\hat{g}_i) - var(g_i)$ dann 0 ist, sobald $k_i = var(g_i)/var(hatg_i)$. Für diesen Wert von k_i, ist $cov(k_i\hat{g}_i, g_i) = var(g_i)$, was eine Konstante ist für alle Tiere unabhängig von deren Genauigkeit r_i^2. Somit ist die Matrix $\mathbf{K}$ bestimmt als $\mathbf{K} = diag(r_i^{^2})$.

Wichtig hier ist die Feststellung, dass allfällige post-analytische Basiskorrekturen einen Einfluss auf die dregressierten Werte haben. Dies sieht man eindrücklich beim Betrachten des Kontrasts zwischen deregressierten geschätzten Zuchtwerten von zwei Tieren i und j. Ohne Entfernung der Basiskorrektur sieht der Kontrast im allgemeinen, wie folgt aus:

$$\frac{\hat{g}_i - b}{r_i^2} - \frac{\hat{g}_j - b}{r_j^2} \neq \frac{\hat{g}_i}{r_i^2} - \frac{\hat{g}_j}{r_j^2}$$

Somit müssen diese Basisanpassungen wieder rückgängig gemacht werden.

Deregressierte Beobachtungen {wahrscheinlch sind hier mit "Beobachtungen" die deregressierten geschätzten Zuchtwerte $\hat{g}_i/r_i$ gemeint} entsprechen Einzelwerten, welche die gesamte Information zum Individuum und dessen Verwandten enthalten, als wären es phänotypische Beobachtungen mit $h^2 = r^2$. Dies kann gezeigt werden, da h^2 die Regression des Genotyps auf den Phänotypen darstellt. Nehmen wir die deregressierten Zuchtwerte als Phänotypen, dann folgt, dass

$$h^2 = \frac{cov(\hat{g}_i/r_i^2, g)}{var(\hat{g}_i/r_i^2)} = \frac{1/r_i^2 var(\hat{g}_i)}{1/r_i^4 var(\hat{g}_i)} = r_i^2$$

Werden in der genomischen Effektschätzung für den Trainings-Schritt deregressierte geschätzte Zuchtwerte verwendet, dann ist das äquivalent zu einem Training mit Phanotypen mit variablem h^2. Vorausgesetzt, dass $r_i^2 > h^2$ ist, ist die Verwendung von deregressierten geschätzten Zuchtwerten äquivalent zu Merkmalen mit höherer Heritabilität. Dieser Effekt der gesteigerten Heritabilität wird später bei der Diskussion um den Ausschluss von Ahnen-Informationen wieder verringert, oder wird ganz verschwinden. {Somit hat diese

Diskussion um die Analogie der degressierten Zuchtwerte mit Merkmalen mit einer erhöhten Heritabilität wohl keine grosse Bedeutung und hat eher akademischen Charakter}

Gewichtung von deregressierten Informationen

Deregressierte Beobachtungen haben heterogene Varianz, wenn r^2 zwischen den Individuen variiert. Die Varianz der Residuen für eine bestimmte deregressierte Beobachtung entspricht:

$$var(\epsilon_i + k_i\hat{g}_i - g_i) = var(\epsilon_i) + var(k_i\hat{g}_i - g_i) = var(\epsilon_i) + k_i^2 var(\hat{g}_i) + var(g_i) - 2k_i var(\hat{g}_i)$$

wobei $var(\hat{g}_i) = r_i^2 var(g_i)$ und $k_i r_i^2 = 1$. Somit vereinfacht sich die Varianz der Residuen zu

$$var(\epsilon_i + k_i\hat{g}_i - g_i) = var(\epsilon_i) + \frac{1 - r_i^2}{r_i^2} var(g_i)$$

Werden wie zuvor die Off-Diagonalelemente ignoriert, dann entsprechen die Diagonalelemente der Inversen der Covarianzmatrix der Residuen nach der Faktorierung von σ_e^2 dem folgenden Ausdruck:

$$\frac{\sigma_e^2}{[c + (1 - r_i^2)/r_i^2]\sigma_g^2}$$

was sich vereinfachen lässt zu

$$w_i = \frac{1 - h^2}{[c + (1 - r_i^2)/r_i^2]h^2}$$

Entfernung der Effekte der Elterndurchschnitte

Zuchtwerte, welche mit dem BLUP Tiermodell geschätzt werden, schrumpfen (shrink) die Informationen der Individuen und der Nachkommen zum Mittel der Elternzuchtwerte (parental average PA EBV) hin. {Im Paper wird hier Mrode (2005) als Referenz angegeben.} Somit erscheint es aus zwei Gründen sinnvoll die Effekte der Elterndurchschnitte abzuziehen.

1. Individuen ohne Information

Individuen mit Zuchtwerten ohne eigene Information oder ohne Nachkommeninformationen können keinen Beitrag zur Schätzung der genomischen Zuchtwerte liefern. Angenommen eine Gruppe von Halbgeschwistern mit individuellen Markergenotypen und deregressierte Ahnenzuchtwerte, dann können diese keine zusätzliche Information liefern, welche nicht schon in den Elterngenotypen und den Elternzuchtwerten enthalten sind.

2. Eltern mit segregierendem Haupteffekt-Allelen

Falls bei den Eltern ein Haupteffekt (major effect) {Wahrscheinlich ist hier gemeint, dass ein Locus mit einem sehr grossen Effekt segregiert.} segregiert, dann erhalten etwa die Hälfte der Nachkommen das Allel mit positiver Wirkung und die andere Hälfte der Nachkommen das Allel mit negativer Wirkung. Die deregressierten Zuchtwerte aber werden zum gleichen Elterndurchschnitt hin geschrumpft (shrunken towards the same PA EBV). Elterndurchschnittseffekte können durch direktes speichern deregressierten Informationen des Individuums und der Nachkommen während der iterativen Lösung der Gleichungssysteme, eliminiert werden. {Hier wird auf das Paper von VanRaden and Wiggans (1991) verwiesen.} Falls kein Zugang zum

Routine-Prozess der Zuchtwertschätzung möglich ist, dann ist es nötig, die Schätzgleichungen zu approximieren und dann daraus die deregressierten Werte ohne die Elterndurchschnitte abzuleiten.

{Was ab hier folgt, ist eine ziemlich lange und verästelte Herleitung von verschiedenen Resultaten. Mangels besserer Kenntnis wird hier angenommen, dass die ganzen Ausführungen dazu dienen, die Gleichungen der Zuchtwertschätzung zu approximieren und daraus dann die deregressierten Zuchtwerte ohne Elterndurchschnitte abzuleiten. Diese Ableitung wird offenbar so gemacht, dass ein lineares Gleichungssystem aufgestellt wird, welches die Komponente des Elterndurchschnitts $\hat{g}_{PA}$ im geschätzten Zuchtwert $\hat{g}$ von den Komponente des Individuums $\hat{g}_i$ trennt.}

Diese Approximation der Schätzgleichungen kann für ein Individuum i mit Eltern s und d aufgrund folgender gegebener Grössen gemacht werden:

- h^2
- geschätzte Zuchtwerte für Individuum i und seine Eltern s und d
- r^2 für i, s und d.

{Die Herleitung der degressierten Zuchtwerte, welche um den Elterndurchschnitt korrigiert sind, wird in drei Schritten gemacht. Im Paper ist diese Aufteilung der drei Schritte nicht so explizit mit Untertiteln gemacht, wie hier, aber die um die Untertitel erweiterte Gliederung sollte zu einem besseren Verständnis beitragen.}

Schritt 1: Aufstellen der Mischmodellgleichungen

Die Elterndurchschnitte (PA) für Zuchtwert $\hat{g}$ und das Bestimmtheitsmass r^2 lassen sich berechnen als:

$$\hat{g}_{PA} = \frac{\hat{g}_s + \hat{g}_d}{2}$$

und

$$r^2_{PA} = \frac{r^2_s + r^2_d}{4}$$

Angenommen die Eltern sind nicht verwandt und nicht ingezüchtet, dann entspricht die additiv genetische Covarianzmatrix zwischen dem Elterndurchschnitt und dem Nachkommen

$$\mathbf{G} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 1 \end{bmatrix} \sigma^2_g$$

{Die Bedeutung der Komponenten der angegebenen additiv genetischen Covarianzmatrix wird nicht genau spezifiziert. Wahrscheinlich wird implizit auch der Elterndurchschnitt für die wahren Zuchtwerte g gebildet. Somit ist dann $g_{PA} = (g_s + g_d)/2$ und $var(g_{PA}) = var((g_s + g_d)/2) = (var(g_s) + var(g_d))/4 = 0.5\sigma^2_g$, wobei $var(g_s) = var(g_d) = \sigma^2_g$ und $cov(g_s, g_d) = 0$, falls s und d nicht verwandt und nicht ingezüchtet sind. Die Off-Diagonalelemente berechnen sich als: $cov(g_{PA}, g_i) = cov((g_s + g_d)/2, g_i) = (cov(g_s, g_i) + cov(g_d.g_i))/2$. Sind s und d weder verwandt noch ingezüchtet, dann folgt, dass $cov(g_s, g_i) = cov(g_d.g_i) = 0.5\sigma^2_g$. Somit sind auch die Off-Diagonalelemente gleich $0.5\sigma^2_g$. Das zweite Diagonalelement entspricht $var(g_i) = \sigma^2_g$, falls i nicht ingezüchtet ist.}

Die Inverse Matrix $\mathbf{G}^{-1}$ der Matrix $\mathbf{G}$ berechnet sich aus den vereinfachten Invertierungsregeln für 2×2 Matrizen und ist somit:

$$\mathbf{G}^{-1} = \begin{bmatrix} 4 & -2 \\ -2 & 2 \end{bmatrix} \sigma^{-2}_g$$

{Die Regeln zur Invertierung einer 2×2 Matrix $\mathbf{A}$ sehen wie folgt aus:

- Auf der Hauptdiagonalen werden die Elemente vertauscht
- Die Elemente der Nebendiagonalen werden mit -1 multipliziert
- Die ganze Matrix wird mit dem Kehrwert $1/det(A)$ der Determinanten multipliziert

Die Determinante einer 2×2 Matrix $\mathbf{A}$ berechnet sich aus der Differenz zwischen dem Produkt der Elemente auf der Hauptdiagonalen minus dem Produkt der Elemente der Offdiagonalen, oder anders

$$det(A) = a_{11} * a_{22} - a_{12} * a_{21}$$

}

Das Resultat der Matrix $\mathbf{G}^{-1}$ wird verwendet um das Gleichungssystem aufzustellen, welches den geschätzten Zuchtwert in die Komponente des Elterndurchschnitts und die Komponente des Individuums aufteilt. Dieses Gleichungssystem sieht wie folgt aus.

$$\begin{bmatrix} Z'_{PA}Z_{PA} + 4\lambda & -2\lambda \\ -2\lambda & Z'_iZ_i + 2\lambda \end{bmatrix} \begin{bmatrix} \hat{g}_{PA} \\ \hat{g}_i \end{bmatrix} = \begin{bmatrix} y^*_{PA} \\ y^*_i \end{bmatrix} \tag{4}$$

wobei y^*_i der Information, welche zum Individuum gehört und äquivalent einer rechten Handseite (right-hand-side) eines Gleichungssystems ist. $Z'_{PA}Z_{PA}$ und Z'_iZ_i entsprechen den unbekannten Anteilen der Komponente des Elterndurchschnitts und der Komponente des Individuums (plus Information von Nachkommen). $\lambda = (1 - h^2)/h^2$ wird als bekannt angenommen.

Wir definieren nun

$$\begin{bmatrix} Z'_{PA}Z_{PA} + 4\lambda & -2\lambda \\ -2\lambda & Z'_iZ_i + 2\lambda \end{bmatrix}^{-1} = \begin{bmatrix} c^{PA,PA} & c^{PA,i} \\ c^{i,PA} & c^{i,i} \end{bmatrix} = \mathbf{C} \tag{5}$$

Basierend auf der Tatsache aus Henderson (1975) ist $r^2_i = \frac{var(\hat{g}_i)}{var(g_i)}$ und $var(\hat{\mathbf{g}}) = \mathbf{G} - \mathbf{C}\sigma^2_e$. Daraus folgt, dass $r^2_{PA} = 0.5 - \lambda c^{PA,PA}$ und $r^2_i = 1 - \lambda c^{i,i}$. Umgeformt und nach den c-Komponenten aufgelöst, erhalten wir $c^{PA,PA} = (0.5 - r^2_{PA})/\lambda$ und $c^{i,i} = (1 - r^2_i)/\lambda$.

{Die Herleitung der beiden Bestimmtheitsmasse erscheint hier nicht ganz so offensichtlich, wie das im Paper dargestellt ist, vor allem beim Ausdruck für r^2_{PA} kann der Hinweis mit der Definition von r^2_i irreführend sein.

Zunächst folgt die Definition von r^2_i aus Henderson (1975) dadurch, dass bei BLUP gilt, dass $cov(\hat{g}_i, g_i) = var(\hat{g}_i)$ dies ist im Anhang gezeigt. Somit lässt sich die Definition des Bestimmtheitsmasses vereinfachen zu:

$$r^2_i = \frac{cov(\hat{g}_i, g_i)^2}{var(\hat{g}_i)var(g_i)} = \frac{var(\hat{g}_i)^2}{var(\hat{g}_i)var(g_i)} = \frac{var(\hat{g}_i)}{var(g_i)}$$

Tier i: Für Tier i ist die Varianz des geschätzten Zuchtwertes $var(\hat{g}_i) = \mathbf{G}_{22} - \mathbf{C}_{22}\sigma^2_e = \sigma^2_g - c^{ii}\sigma^2_e$. Das Bestimmtheitsmass ist dann

$$r^2_i = \frac{var(\hat{g}_i)}{var(g_i)} = \frac{\sigma^2_g - c^{ii}\sigma^2_e}{\sigma^2_g} = 1 - \lambda c^{ii}$$

wobei verwendet wurde, dass

$$\lambda = \frac{1 - h^2}{h^2} = \frac{1 - \sigma^2_g/\sigma^2_p}{\sigma^2_g/\sigma^2_p} = \frac{\sigma^2_p - \sigma^2_g}{\sigma^2_g} = \frac{\sigma^2_e}{\sigma^2_g}$$

Elterndurchschnitt PA: Beim Elterndurchschnitt kann das Bestimmtheitsmass nicht direkt notiert werden, sondern es müssen dabei folgende Beziezungen berücksichtigt werden:

$$r^2_{PA} = \frac{r^2_s + r^2_d}{4} = 1/4 \left(\frac{var(\hat{g}_s)}{var(g_s)} + \frac{var(\hat{g}_d)}{var(g_d)} \right) = \frac{var(\hat{g}_{PA})}{\sigma^2_g}$$

An dieser Stelle verwenden wir, dass $var(\hat{g}_{PA}) = \mathbf{G}_{11} - \mathbf{C}_{11}\sigma^2_e = 0.5\sigma^2_g - c^{PA,PA}\sigma^2_e$

Das eingesetzt in die Formel für das Bestimmtheitsmass r^2_{PA} ergibt

$$r^2_{PA} = \frac{0.5\sigma^2_g - c^{PA,PA}\sigma^2_e}{\sigma^2_g} = 0.5 - \lambda c^{PA,PA}$$

Lösen wir diese Gleichungen nach den Elementen $c^{PA,PA}$ und $c^{i,i}$ der Matrix $\mathbf{C}$ auf so erhalten wir

$c^{PA,PA} = (0.5 - r^2_{PA})/\lambda$ und $c^{i,i} = (1 - r^2_i)/\lambda$

}

Basierend auf der Formel zur Invertierung einer 2×2-matrix können die Komponenten $c^{PA,PA}$ und $c^{i,i}$ auch geschrieben werden als

$$c^{PA,PA} = (Z_i'Z_i + 2\lambda)/det(\mathbf{C}) = (0.5 - r^2_{PA})/\lambda \tag{6}$$

$$c^{i,i} = (Z'_{PA}Z_{PA} + 4\lambda)/det(\mathbf{C}) = (1 - r^2_i)/\lambda \tag{7}$$

wobei $det(\mathbf{C}) = (Z'_{PA}Z_{PA} + 4\lambda)(Z_i'Z_i + 2\lambda) - 4\lambda^2$

{Gleichungen (6) und (7) entsprechen den Gleichungen (12) und (13) im Paper.}

Schritt 2: Lösungen für $Z_i'Z_i$ und $Z'_{PA}Z_{PA}$

In einem zweiten Schritt werden die Gleichungen (6) und (7) nach $Z_i'Z_i$ und $Z'_{PA}Z_{PA}$ aufgelöst. Eine direkte Lösung erhält man durch Division von Gleichung (6) durch Gleichung (7). Setzen wir $\delta = (0.5 - r^2_{PA})/(1 - r^2_i)$ und vereinfachen den Gleichungsquotienten, so erhalten wir

$$Z_i'Z_i = \delta Z'_{PA}Z_{PA} + 2\lambda(2\delta - 1) \tag{8}$$

Setzen wir (8) in (7) ein und definieren $\alpha = 1/(0.5 - r^2_{PA})$ erhalten wir folgende quadratische Gleichung für $Z'_{PA}Z_{PA}$:

$$0.5(Z'_{PA}Z_{PA})^2 + \lambda(4 - 0.5\alpha)(Z'_{PA}Z_{PA}) + 2\lambda^2(4 - \alpha - 1/\delta) = 0 \tag{9}$$

{Im Paper wird ohne genauere Begründung einfach die positive Lösung weiter verwendet.}

Die quadratische Gleichung (9) hat folgende positive Lösung

$$Z'_{PA}Z_{PA} = \lambda(4 - 0.5\alpha) + 0.5\lambda\sqrt{(\alpha^2 + 16/\delta)} \tag{10}$$

Durch Einsetzen der Beziehung für $Z'_{PA}Z_{PA}$ aus Gleichung (10) in Gleichung (8) erhalten wir die Lösung für $Z_i'Z_i$. Die beiden Lösungen für $Z'_{PA}Z_{PA}$ und $Z_i'Z_i$ ermöglicht uns die Koeffizientenmatrix aus Gleichung (4) aufzustellen.

Schritt 3: Deregressierter Zuchtwert

Die rechte Handseite der Mischmodellgleichungen in (4) resultiert aus der Multiplikation der Koeffizientenmatrix mit dem bekannten Vektor der Zuchtwerte des Elterndurchschnitts und des Individuums. Die Komponente der rechten Handseite, welche befreit ist vom Elterndurchschnitt entspricht y_i^*. Die Gleichung für den geschätzten Zuchtwert ($\hat{g}_{i-PA}$) des Individuums i, welcher befreit ist vom Elterndurchschnitt lautet:

$$[Z_i'Z_i + \lambda]\,[\hat{g}_{i-PA}] = [y_i^*]$$

Das zugehörige Bestimmtheitsmass lautet:

$$r_i^{2*} = 1 - \lambda/(Z_i'Z_i + \lambda)$$

Der deregressierte Zuchtwert frei vom Elterndurchschnitt lautet dann

$$\frac{\hat{g}_{i-PA}}{r_i^{2*}} = \frac{y_i^*}{Z_i'Z_i}$$

Anhang

Im Paper wird verwendet, dass $cov(\hat{g}_i, g_i) = var(\hat{g}_i)$. Hier wird diese Beziehung bewiesen. Der Beweis verwendet einfach die Definitionen der verwendeten Grössen und vereinfacht die resultierenden Ausdrücke bis zu deren Gleichheit.

{In diesem Abschnitt wurde die Notation vereinfacht. Alle Matrizen und Vektoren sind nicht in boldface geschrieben.}

Das Modell

Das folgende gemischte linear Modell wird angenommen:

$$y = Xb + Zg + e \tag{11}$$

wobei:

y	Vektor der Beobachtungen
b	Vektor der fixen Effekte
X	Inzidenzmatrix zwischen fixen Effekten und Beobachtungen
g	Vektor der zufälligen Effekte
Z	Inzidenzmatrix zwischen zufälligen Effekten und Beobachtungen
e	ein Vektor von zufälligen Resteffekten

Die Varianzen für die zufälligen Effekte lauten: $var(g) = G$, $var(e) = R$ daraus folgt dann, dass

$$var(y) = var(Xb + Zg + e) = ZGZ^T + R$$

Analog dazu können wir ableiten, dass

$$cov(y, g^T) = cov(Xb + Zg + e, g^T) = ZG$$

Die Schätzer

Für die Parameter b und g verwenden wir den BLUE Schätzer ($\hat{b}$) und den BLUP Schätzer ($\hat{g}$). Somit gilt, dass

$$\hat{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{12}$$

und

$$\hat{g} = GZ^T V^{-1} (y - X\hat{b}) \tag{13}$$

Der Beweis

Der Beweis läuft so, dass für beide Terme $cov(\hat{g}_i, g_i)$ und $var(\hat{g}_i)$ die Definitionen eingesetzt werden und dann die Terme so vereinfacht werden, bis sie gleich sind.

$$var(\hat{g}) = var(GZ^T V^{-1}(y - X\hat{b})) = GZ^T V^{-1} * var(y - X\hat{b}) * V^{-1} ZG \tag{14}$$

In Gleichung (14) berechnen wir vorerst einmal nur den Term $var(y - X\hat{b})$ und setzen anschliessend das Resultat wieder ein.

$$\begin{aligned} var(y - X\hat{b}) &= var(y) + var(X\hat{b}) - cov(y, (X\hat{b})^T) - cov(X\hat{b}, y^T) \\ &= V + X * var(\hat{b}) * X^T - cov(y, \hat{b}^T) * X^T - X * cov(\hat{b}, y^T) \end{aligned} \tag{15}$$

Aus (15) berechnen wir $var(\hat{b})$ und $cov(y, \hat{b}^T)$ wobei gilt, dass $cov(y, \hat{b}^T) = (cov(\hat{b}, y^T))^T$

$$\begin{aligned} var(\hat{b}) &= var((X^T V^{-1} X)^{-1} X^T V^{-1} y) \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} * var(y) * V^{-1} X (X^T V^{-1} X)^{-1} \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} * V * V^{-1} X (X^T V^{-1} X)^{-1} \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} X (X^T V^{-1} X)^{-1} \\ &= (X^T V^{-1} X)^{-1} \end{aligned} \tag{16}$$

Die Kovarianz zwischen der Beobachtung y und dem Schätzer $\hat{b}$ lautet

$$\begin{aligned} cov(\hat{b}, y^T) &= cov((X^T V^{-1} X)^{-1} X^T V^{-1} y, y^T) \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} * cov(y, y^T) \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} * V \\ &= (X^T V^{-1} X)^{-1} X^T \end{aligned} \tag{17}$$

Das Resultate aus (17) und (16) werden in (15) eingesetzt.

$$\begin{aligned} var(y - X\hat{b}) &= var(y) + var(X\hat{b}) - cov(y, (X\hat{b})^T) - cov(X\hat{b}, y^T) \\ &= V + X * var(\hat{b}) * X^T - cov(y, \hat{b}^T) * X^T - X * cov(\hat{b}, y^T) \\ &= V + X * (X^T V^{-1} X)^{-1} * X^T - X (X^T V^{-1} X)^{-1} * X^T - X * (X^T V^{-1} X)^{-1} X^T \\ &= V - X * (X^T V^{-1} X)^{-1} * X^T \end{aligned} \tag{18}$$

Somit vereinfacht sich $var(\hat{g})$ zu:

$$\begin{aligned} var(\hat{g}) &= var(G Z^T V^{-1} (y - X\hat{b})) \\ &= G Z^T V^{-1} * var(y - X\hat{b}) * V^{-1} Z G \\ &= G Z^T V^{-1} * (V - X * (X^T V^{-1} X)^{-1} * X^T) * V^{-1} Z G \\ &= G Z^T V^{-1} * V * V^{-1} Z G - G Z^T V^{-1} * X * (X^T V^{-1} X)^{-1} * X^T * V^{-1} Z G \\ &= G Z^T V^{-1} (I - X (X^T V^{-1} X)^{-1} X^T V^{-1}) Z G \end{aligned} \tag{19}$$

Analog vefahren wir mit $cov(\hat{g}, g^T)$. Als erstes setzen wir die Definition von $\hat{g}$ ein und vereinfachen bis wir das gewünschte Resultat erhalten.

$$
\begin{aligned}
cov(\hat{g}, g^T) &= cov(GZ^T V^{-1}(y - X\hat{b}), g^T) \\
&= GZ^T V^{-1} cov(((y - X\hat{b})), g^T) \\
&= GZ^T V^{-1} \Big(cov(y, g^T) - X * cov(\hat{b}, g^T) \Big) \\
&= GZ^T V^{-1} \left(cov(y, g^T) - X * cov((X^T V^{-1} X)^{-1} X^T V^{-1} y, g^T) \right) \\
&= GZ^T V^{-1} \left(cov(y, g^T) - X (X^T V^{-1} X)^{-1} X^T V^{-1} * cov(y, g^T) \right) \\
&= GZ^T V^{-1} \left(ZG - X (X^T V^{-1} X)^{-1} X^T V^{-1} * ZG \right) \\
&= GZ^T V^{-1} \left(I - X (X^T V^{-1} X)^{-1} X^T V^{-1} \right) ZG \qquad (20)
\end{aligned}
$$

Ein Vergleich zwischen (19) und (20) zeigt, dass die beiden Terme gleich sind und somit gilt, dass

$$var(\hat{g}) = cov(\hat{g}, g^T)$$

References

Garrick, Dorian, Jeremy Taylor, and Rohan Fernando. 2009. "Deregressing Estimated Breeding Values and Weighting Information for Genomic Regression Analyses." *Genetics Selection Evolution* 41 (1): 55. doi:10.1186/1297-9686-41-55.

Henderson, CR. 1975. "Best Linear Unbiased Estimation and Prediction Under a Selection Model." *Biometrics* 31: 423–49.

Mrode, R. 2005. *Linear Models for the Prediction of Animal Breeding Values*. Edited by Cambridge: CABI.

VanRaden, PM, and GR Wiggans. 1991. "Derivation, Calculation, and Use of National Animal Model Information." *J Dairy Sci* 74 (8): 2737–46. http://www.pubmed.org/display.cgi?uids=1918547.