

Angewandte Statistische Methoden in den Nutztierwissenschaften

Peter von Rohr

20 February 2017

Administration

- ▶ Veranstaltung: 2 V im Vorlesungsverzeichnis
- ▶ Plan: 2 V \rightarrow 1 U + 1V
- ▶ Übungen: Beispiele in R
- ▶ Unterlagen: Folien, Übungen, Lösungen
- ▶ Prüfung: schriftlich, Termin: 29.5.2017, Dauer: 1 Stunde
- ▶ Prüfungsstoff: Lösungen der Übungen, Folien

Lernziele

Die Studierenden . . .

- ▶ kennen die Eigenschaften der multiplen linearen Regression und
- ▶ können einfache Datensätze mithilfe der Regressionsmethode analysieren
- ▶ wissen wieso multiple linear Regressionen bei der genomischen Selektion nicht brauchbar ist
- ▶ kennen die in der genomischen Selektion verwendeten statistischen Verfahren, wie
- ▶ BLUP-basierte Verfahren,
- ▶ Bayes'sche Verfahren und
- ▶ die LASSO Methode
- ▶ können einfach Übungsbeispiele mit der Statistiksoftware R erfolgreich bearbeiten.

Programm

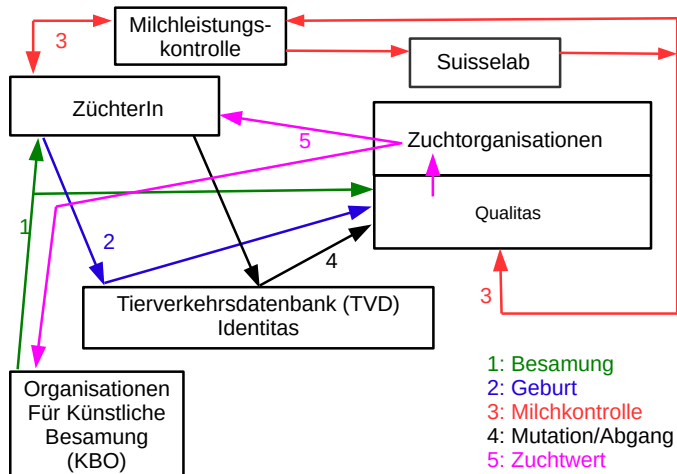
Woche	Datum	Thema
1	20.02	Einführung, Statistik in der Genomik
2	27.02	Problem der multiplen lineare Regression
3	06.03	Alternative Ansätze - LASSO
4	13.03	BLUP
5	20.03	Einführung in Bayes'sche Statistik
6	27.03	Bayes'sche Verfahren in der Genomik
15	29.05	Prüfung

Thema

- ▶ Mit **Genomischer Selektion** (GS) kam Paradigmentwechsel in der Tierzucht
- ▶ Bedeutung dieser Veränderung mit Fokus auf die verwendeten statistische Methoden
- ▶ Grundstein für GS war das Paper

“Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829”

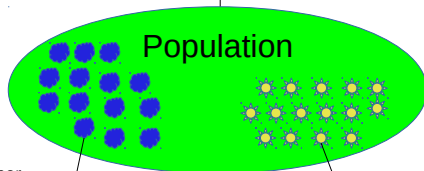
Informationsfluss in einem Zuchtprogramm



Genomische Selektion

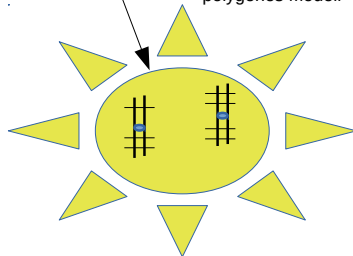
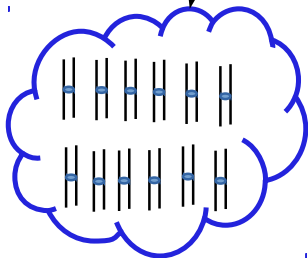
Tiermodell

Genomische
Selektion



Unendlich viele
unbekannte
„Gene“ mit kleiner
Wirkung ==>
Infinitesimalmodell

Begrenzte Anzahl
Loci mit
geschätzter
Wirkung ==>
polygenes Modell



Vor Einführung von GS

- ▶ Informationsquellen für Zuchtwertschätzung
- ▶ phänotypische Leistungen
- ▶ verwandtschaftliche Beziehungen / Abstammungen / Pedigree
- ▶ Varianzkomponenten aus periodischen Schätzungen
- ▶ BLUP Tiermodell zur Schätzung der Zuchtwerte
- ▶ ab ca 1990 einzelne genetische Marker als fixe oder zufällige Effekte ins Tiermodell integriert
- ▶ Problem: einzelne Marker werden sehr schnell fixiert
- ▶ Konsequenzen der Fixation für das Zuchtziel unbekannt?
- ▶ Uneinigkeit, welches die beste Strategie sein könnte
- ▶ Durch technologischen Fortschritt wurde Problem hinfällig

Modellierung vor GS

- ▶ BLUP Tiermodell

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

wobei:

- y**: Vektor mit phänotypischen Beobachtungen
- b**: Vektor mit fixen Effekten
- X**: Inzidenzmatrix, welche fixe Effekte den Beobachtungen zuordnet
- u**: Vektor mit Zuchtwerten (zufällig)
- Z**: Inzidenzmatrix der Zuchtwerte
- e**: Vektor mit Residuen (zufällig)

- ▶ Varianzen: $Var(\mathbf{e}) = \mathbf{R} = \mathbf{I} * \sigma_e^2$, $Var(\mathbf{u}) = \mathbf{G} = \mathbf{A} * \sigma_g^2$,
 $Cov(\mathbf{u}, \mathbf{e}^T) = Cov(\mathbf{e}, \mathbf{u}^T) = \mathbf{0}$, $\rightarrow Var(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$

Mit GS

- ▶ Genomische Selektion ist Methode der Wahl in modernen Zuchtprogrammen
- ▶ **Theorie:** Genomische Zuchtwerte basieren nicht mehr auf dem BLUP-Tiermodell mit einem “infinitesimal model” für die Genwirkung sondern auf einem Modell einer endlichen Anzahl Genorten mit hauptsächlich additiver Wirkung.
- ▶ Häufig verwendet wird eine Prozedur bestehend aus zwei Schritten
 1. Schätzung der additiven Substitutionseffekte (a -Werte im Substitutionsmodell)
 2. Schätzung der genomischen Zuchtwerten aufgrund der unter 1) geschätzten Substitutionseffekte und aufgrund der Typisierungsergebnisse
- ▶ NB: Es gibt auch Verfahren, welche beide Schritte zu einem kombinieren, sogenannte “single step” Verfahren

Modellierung mit GS

- ▶ Annahme: Betrachtung der zwei-Schritt Prozedur
- ▶ Da Genorte bekannt und SNP-Genotypen beobachtet werden können, braucht es kein Tiermodell mit zufälligen Zuchtwerte mehr
- ▶ Somit brauchen Zuchtwerte nicht mehr als zufällige Effekte eines gemischten linearen Modells geschätzt zu werden.
- ▶ Genetische Komponenten können als additiv genetische Effekte nach dem Gen-Substitutionsmodell aus der quantitativen Genetik geschätzt werden
- ▶ Was bleibt ist ein Modell mit nur fixen Effekten und einem zufälligen Rest

Modelle in GS

- ▶ Wie sieht der Schritt 1 aus?
- ▶ Idealfall: Summe aller additiven Genwirkungen und SNP Genotypen als Beobachtungen, daraus können Substitutionseffekte mit folgendem Modell geschätzt werden

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{M}\mathbf{a} + \epsilon$$

wobei:

- g** Vektor von wahren Zuchtwerten
- μ Achsenabschnitt
- a** Vektor mit Gensubstitutionseffekten
- M** Inzidenzmatrix als Verknüpfung zwischen **a** und **g**
- ϵ Vektor von zufälligen Residuen

Modelle in GS II

- ▶ Wahre Zuchtwerte können nicht beobachtet werden
- ▶ Alternativ dazu können phänotypische Beobachtungen verwendet werden
- ▶ Individuelle Beobachtung beim Tier

$$\mathbf{y} = (\mathbf{1}\mu + \mathbf{Xb}) + \mathbf{Ma} + (\epsilon + \mathbf{e})$$

wobei:

- y** Vektor der phänotypischen Beobachtungen
- b** Vektor der fixen Umweltfaktoren
- X** Inzidenzmatrix der fixen Effekte
- e** Vektor von nicht-genetische Residuen

Modelle in GS III

- ▶ BLUP Zuchtwerte $\hat{\mathbf{g}}$ werden wie Beobachtungen behandelt
- ▶ Idee: Geschätzter Zuchtwert = wahrer Zuchtwert plus Abweichung

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

Probleme mit BLUP Zuchtwerten als Beobachtungen:

1. Addition der Abweichung zu \mathbf{g} führt zu einer Reduktion der Varianz
2. BLUP Zuchtwerte werden gegen das Mittel der Eltern gedrückt (shrinkage estimator)

→ **Deregression** der Zuchtwerte

Reduktion der Varianz

- ▶ Reduktion der Varianz heisst: $var(\hat{g}_i) \leq var(g_i)$, obwohl $var(\hat{g}_i - g_i) \geq 0$
- ▶ Addition der Abweichung $(\hat{g}_i - g_i)$ zum wahren Zuchtwert g_i reduziert die Varianz $var(g + (\hat{g}_i - g_i))$ der Summe
- ▶ Grund bei BLUP gilt: $cov(\hat{g}_i, g_i) = var(\hat{g}_i)$
- ▶ Allgemein gilt: $var(a - b) = var(a) + var(b) - 2 * cov(a, b)$
- ▶ Anwendung auf BLUP Zuchtwerte:

$$var(\hat{g}_i - g_i) = var(\hat{g}_i) + var(g_i) - 2 * cov(\hat{g}_i, g_i) = var(g_i) - var(\hat{g}_i) \geq 0$$

$$\rightarrow var(\hat{g}_i) \leq var(g_i)$$

Schrumpfen (shrinkage) zum Elterndurchschnitt

- ▶ Dies gilt für alle Nachkommen gleich unabhängig der Allele, welche sie erhalten haben
- ▶ Das Ausmass der Schrumpfung ist abhängig von der Genauigkeit des geschätzten Zuchtwerts
- ▶ Unser Interesse ist aber der Einfluss von Markerallelen auf den Phänotyp und der ist unabhängig von der Genauigkeit der Zuchtwerte
- ▶ → Deregression hebt diese Effekte auf

Deregression von Zuchtwerten

- ▶ Wie können wir die geschätzten Zuchtwerte anpassen, damit die genannten Probleme nicht mehr bestehen
- ▶ Gesucht ist eine Matrix **K**, welche mit den geschätzten Zuchtwerten multipliziert wird, so dass Probleme behoben werden

Modelle in GS Zusammenfassung

- ▶ Zusammenfassung der Modelle zur Effektschätzung

wahre Zuchtwerte

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{M}\mathbf{a} + \epsilon$$

phänotypische Beobachtungen

$$\mathbf{y} = (\mathbf{1}\mu + \mathbf{X}\mathbf{b}) + \mathbf{M}\mathbf{a} + (\epsilon + \mathbf{e})$$

geschätzte Zuchtwerte

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

- ▶ Alle diese Modelle sehen eigentlich aus wie eine ...
- ▶ Aber ...

Probleme bei der Modellierung mit GS

- ▶ Statistische Modelle in genomischer Zuchtwertschätzung haben mehr Parameter (p) als Beobachtungen (n)
- ▶ Konsequenz: least squares funktioniert nicht zur Parameterschätzung in den verwendeten Regressionsmodellen
- ▶ Frage: welche Methoden stehen zur Auswahl
- ▶ Multiple lineare Regression
- ▶ LASSO
- ▶ BLUP - single step Verfahren
- ▶ Bayes'sche Verfahren

Multiple lineare Regression

- ▶ Einfachheit des Modells
- ▶ Least squares kann nicht verwendet werden, da $n \ll p$
- ▶ kein Ersatz der multiplen linearen Regression durch mehrere Regressionen mit weniger Parameter
- ▶ Forward-selection, d.h. schrittweise Berücksichtigung von signifikanten SNPs im Modell ist keine stabile Prozedur, da diese vom Startpunkt abhängig ist
- ▶ Referenz

“> Kapitel 1 aus den Vorlesungsunterlagen zu: Computational Statistics. Peter Bühlmann und Martin Mächler. Seminar für Statistik ETHZ. Version 2014”

LASSO

- ▶ LASSO bedeutet Least Absolute Shrinkage and Selection Operator
- ▶ Veränderung der Zielfunktion von den quadrierten Residuen zum Absolutbetrag der Residuen führt zu einer Selektion der Effekte
- ▶ Referenzen:

“> Kap 6.2.2 von: Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning. ISBN 978-1-4614-7137-0 ISBN 978-1-4614-7138-7 (eBook) DOI 10.1007/978-1-4614-7138-7 Springer New York Heidelberg Dordrecht London”

BLUP - single step Verfahren

- ▶ SNP-Effekte und genomische Zuchtwerte in einem Schritt geschätzt
- ▶ Tiere mit phänotypischen Leistungen und/oder genomischer Information in einer Auswertung
- ▶ → bessere Berücksichtigung der genetischen Vorselektion in der Auswertung
- ▶ Referenzen:

“> Z. Liu, M. E. Goddard, F. Reinhardt, and R. Reents. A single-step genomic model with direct estimation of marker effects. *J. Dairy Sci.* 97:5833–5850 (2014) <http://dx.doi.org/10.3168/jds.2014-7924>.
<http://www.sciencedirect.com/science/article/pii/S0022030214004895>”

“> Ignacy Misztal, Samuel E. Aggrey, and William M. Muir. Experiences with a single-step genome evaluation. *2013 Poultry Science* 92 :2530–2534 <http://dx.doi.org/10.3382/ps.2012-02739>.
<http://ps.oxfordjournals.org/content/92/9/2530.full.pdf>”

Bayes'sche Verfahren

- ▶ a priori Information so gewählt, dass nur wenige SNP einen Einfluss, d.h. einen Effekt $\neq 0$ haben
- ▶ Referenz:

“> Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829”

“> Kapitel 10 und 11 von: Genome-Wide Association Studies and Genomic Prediction. Cedric Gondro, Julius van der Werf, Ben Hayes. ISSN 1064-3745 ISSN 1940-6029 (electronic) ISBN 978-1-62703-446-3 ISBN 978-1-62703-447-0 (eBook) DOI 10.1007/978-1-62703-447-0 Springer New York Heidelberg Dordrecht London”