Angewandte Statistische Methoden in den Nutztierwissenschaften

Peter von Rohr

2017-03-20

Contents

Vorwort

Dieses Dokument umfasst die kompletten Unterlagen zur Vorlesung **Angewandte Statistische Methoden in den Nutztierwissenschaften**. Der Titel dieser Vorlesung ist sehr allgemein gehalten. Dies würde es erlauben einen grosszügigen Überblick über eine breite Palette an statistischen Methoden, welche in den Nutztierwissenschaften eingesetzt werden, zu geben.

Wir schlagen an dieser Stelle aber einen anderen Weg ein, und fokussieren uns auf die statistischen Methoden in der genomischen Selektion. Nur diese bewusste Wahl eines spezifischen Gebietes ermöglicht es uns, den behandelten Stoff angemessen zu vertiefen. Im anschliessenden Unterabschnitt wollen wir die hier getroffene Entscheidung der Fokusierung auf die genomische Selektion motivieren. Dabei wird klar, dass wir mit der Wahl des Themas der multiplen linearen Regression als Ausgangspunkt auch eine Leserschaft ansprechen, welche nicht primär an der Tierzucht interessiert ist.

Motivation

Vom Standpunkt der statistischen Modellierung, ist das einfache lineare Modell mit fixen Effektstufen für den Einsatz in der genomischen Selektion ausreichend. Diese Art von Modellen werden auch als Regressionsmodelle bezeichnet. Die Problematik entsteht erst bei der Technik, welche wir für die Schätzung der unbekannten Parameter verwenden können. In der klassischen Regressionsanalyse ist die Methode der kleinsten Quadrate (Least Squares) die Methode der Wahl. Least Squares können wir aber für die genomische Selektion nicht verwenden, da die Anzahl unbekannter Parameter (p) grösser ist als die Anzahl Beobachtungen (n).

Mit der steigenden Grösse und Komplexität von aktuellen Datensätzen tritt das soeben beschriebene Problem nicht nur in der Tierzucht auf, sondern es gibt eine breite Palette von Anwendungen. In der Vorlesung beschrieben wir diese Problematik am Beispiel der genomischen Selektion und es werden alternative Techniken zur Schätzung von Parametern vorgeschlagen. Da die Methode der multiplen Regressionsanalyse in früheren Vorlesungen behandelt wurde, bietet diese ein idealer Ausgangspunkt für den in dieser Veranstaltung präsentierten Stoffinhalt.

Einordnung

Die Vorlesung **Angewandte Statistische Methoden in den Nutztierwissenschaften** ist eine halbsemestrige Veranstaltung und wird im Masterstudiengang Agrarwissenschaften der ETH Zürich angeboten.

Lernziele

Für die Verwendung des hier präsentierten Stoffs schlagen wir die folgenden Lernziele vor.

Die Studierenden …

- kennen die Eigenschaften der multiplen linearen Regression und

- können einfache Datensätze mithilfe der Regressionsmethode analysieren
- wissen wieso multiple lineare 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 einfache Übungsbeispiele mit der Statistiksoftware R erfolgreich bearbeiten.

Chapter 1

Einführung

Den in dieser Vorlesung präsentierte Stoff kann aus mehreren Gesichtspunkten betrachtet werden. Aus Sicht der Tierzucht behandeln wir die statistischen Methoden, welche in der **genomischen Selektion** angewendet werden. Für Statistiker stellen wir verschiedene Methoden der Regularisierung in hoch-parametrischen Modellen vor. In der sehr populären Disziplin des **Machine Learnings** wird das hier besprochene Problem als die Selektion von relevanten Features im Kontext des Supervised Learnings dargestellt.

1.1 Beschreibung des Problems

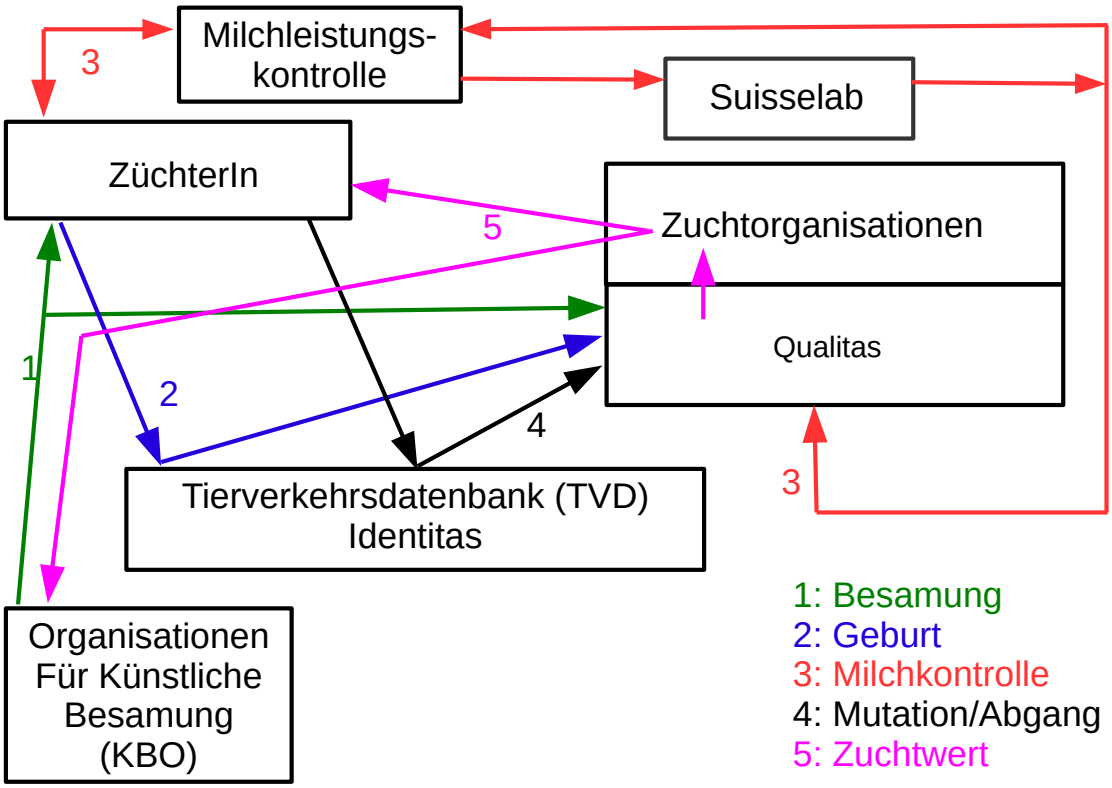
Alle die soeben genannten Formulierungen beschreiben das gleiche Problem. Wir gehen von einem Datensatz aus, welcher aus Beobachtungen besteht. Jede Beobachtung ist charakterisiert durch sehr viele unabhängige Grössen. Die Gewichtung der zu einer Beobachtung gehörenden Grössen wird über unbekannte Parameter erreicht.

Als Beispiel für einen solchen Datensatz können wir eine Population mit SNP-typisierten Tieren betrachten. Das Typisierrungsergebnis für ein bestimmtes Tier enthält die Genotypen an den Genorten, welche bei der Typisierung untersucht werden. Die einzelnen Genorte werden als sogenannte Single Nucleotide Polymorphisms (SNP) bezeichnet. In Abhängigkeit des anbietenden Labors gibt es verschiedene Optionen für die gewünschte Typisierung. Die Optionen unterscheiden sich vor allem in der Dichte der untersuchten Genorte. Das heisst bei einer grösseren Dichte werden mehr SNPs untersucht. Typische Werte von gängigen Anbietern bewegen sich im Bereich zwischen 50000 (50K) bis rund 800000 (800K) untersuchte SNPs pro untersuchtes Genom. Die totale Anzahl an SNP im Genom beträgt rund 20 Millionen. Somit ist ein Typisierungsergebnis eine vom Anbieter gemachte Auswahl aller verfügbaren SNPs.

1.2 Rückblick

Bis Anfangs des 21. Jahrhundert wurden eigentlich keine genomischen Informationen in Zuchtprogrammen berücksichtigt. Mit genomischer Information ist hier die Genotyp-Varianten einer grosser Anzahl von Genorten, welche über das ganze Genom verteilt ist. Um die Jahrtausendwende waren sehr viele ForscherInnen in einem Gebiet aktiv, welches damals als Mapping von sogenannten `Quatitative Trait Loci` (QTL) bezeichnet wurde. Eine Übersicht zu QTL ist im Buch (Balding et al., 2009). Das Ziel der Untersuchungen im Bereich QTL-Mapping war das Finden von Regionen im Genom, welche wichtig sind für die Ausprägung von spezifischen Phänotypen. Heute spricht man nicht mehr QTL-Mapping sondern heute wird die Suche von genetischen Orten, welche einen wichtigen Einfluss auf die Ausprägung eines Phänotyps haben, mit `Genome Wide Association Study` (GWAS) bezeichnet.

Trotz umfangreicher Forschungstätigkeit auf dem Gebiet des QTL-Mappings, fanden keine Resultate aus diesen Arbeiten den Weg in die praktische Zuchtarbeit. Somit verläuft die Zuchtarbeit bis vor kurzem nach dem klassischen Schema, welches nachfolgend gezeigt ist.



1.2.1 Paradigmenwechsel

Die Publikation (Meuwissen et al., 2001) gilt als Grundstein für eine neue Ära in der praktischen Zuchtarbeit. Die Autoren haben gezeigt, wie genomische Information, welche in genügender Dichte vorliegen muss, zur Schätzung von Zuchtwerten verwendet werden kann. Sie konnten auch statistische Methoden zeigen, mit welchen die Parameter in verwendeten Modell geschätzt werden können. Wir werden zu einem späteren Zeitpunkt noch genauer auf den Inhalt des Papers von (Meuwissen et al., 2001) zurückkommen.

1.2.2 Vor der genomischen Selektion

Von Anfangs der 1980-er Jahre wurden die statistischen Auswertungen in den Zuchtprogrammen auf das BLUP-Tiermodell abgestellt. In dieser Zeit wurden die einfachen Modelle auch durch verschiedene Erweiterungen ausgebaut. Bei der Milchproduktion wurde von einfachen Laktationsleistungen auf Testtagesmodelle umgestellt. Bei der Wurfgrösse beim Schwein oder anderen diskreten Merkmalen wurden auch `Generalized Linear Mixed Models` (GLMM) verwendet. Unabhängig von den verwendeten Modellen wurden in allen Auswertungen die gleichen Informationen berücksichtigt. - phänotypische Leistungen - Pedigree - Varianzkomponenten aus periodischen Schätzungen

Versuchsweise wurde ab den 1990-er Jahren erste genetische Marker mit in den Zuchtprogrammen berücksichtigt. Das Problem war dass diese wenigen Markern sehr schnell auf einer bestimmten Variante fixiert

war. Nach der Fixierung lieferten diese Genorte keine zusätzliche Information zur Auswahl von potentiellen Zuchttieren. Es war zu dieser Zeit nicht klar, wie das Problem der Fixierung von einzelnen Genorten behandelt werden soll und es gab auch keine wirklich gute Strategie für die Berücksichtigung von genetischen Informationen in Zuchtprogrammen.

1.2.3 Modellierung vor der genomischen Selektion

Vor der Einführung der genomischen Selektion war das BLUP-Tiermodell die Methode der Wahl für die Auswertung von Leistungsdaten in der Tierzucht. In seiner einfachsten Form sieht dieses Modell wie folgt aus.

$$y = Xb + Zu + e \tag{1.1}$$

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)

Die Co-Varianzen der zufälligen Komponenten sind definiert als:

$$Var(\mathbf{e}) = \mathbf{R} = \mathbf{I} * \sigma_e^2$$

$$Var(\mathbf{u}) = \mathbf{G} = \mathbf{A} * \sigma_g^2$$

$$Cov(\mathbf{u}, \mathbf{e}^T) = Cov(\mathbf{e}, \mathbf{u}^T) = \mathbf{0}$$

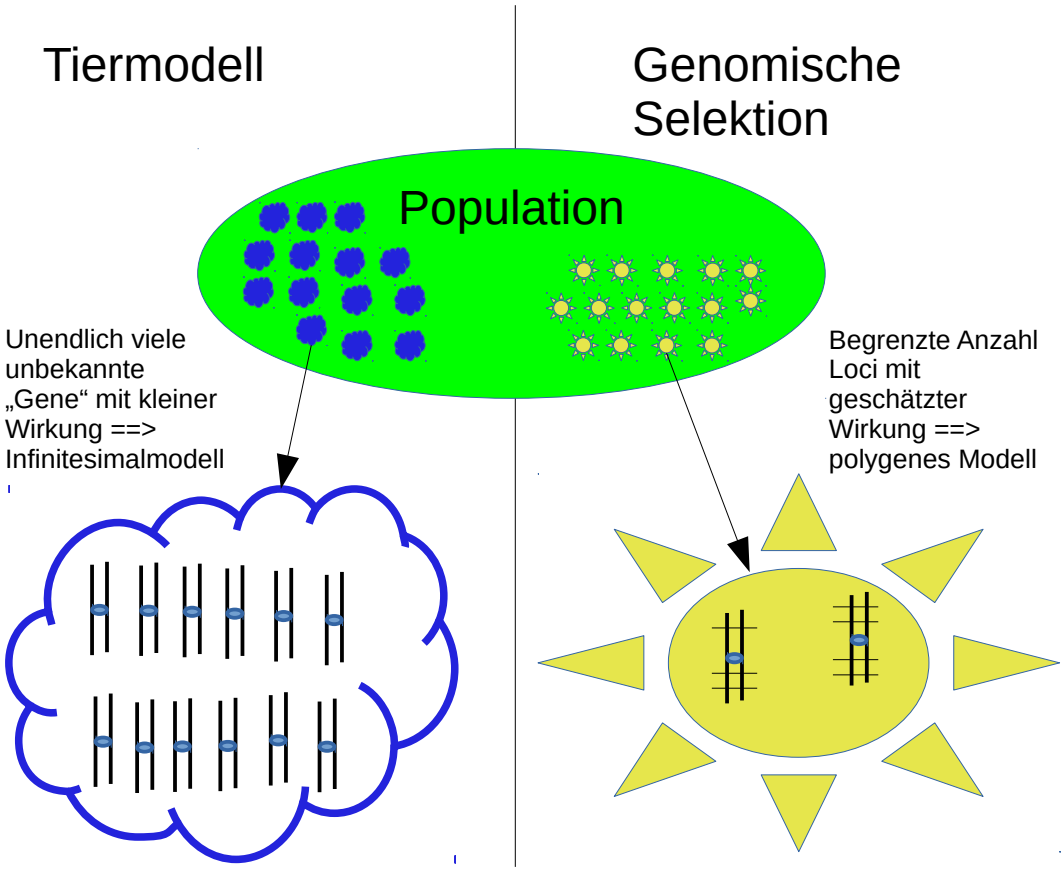
$$\rightarrow Var(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R}$$

1.3 Genomische Selektion

Vom Standpunkt der Genetik aus basiert das BLUP-Tiermodell auf dem sogenannten Infinitesimalmodell. In diesem Modell wird angenommen, dass die phänotypische Ausprägung eines Merkmals durch die Summe von unendlich vielen Genorten mit undendlich kleiner Wirkung verursacht wird. Durch diese Annahme lässt sich dem einzelnen Tier kein fix definierter Genotyp mehr zuordnen. Diese fehlende Zuordnung der einzelnen Genotypen wird über die Modellierung der Zuchtwerte als zufällige Effekte gelöst. Die zufälligen Effekte der Zuchtwerte entsprechen dabei Realisierungen einer Zufallsvariablen mit vorgegebener Verteilung.

In der genomischen Selektion verwenden wir das polygene Modell. Dabei werden die phänotypischen Leistungen als Summe von bekannten Genorten zusammengesetzt. Die konkrete Umsetzung des polygenen Modells wurde zum ersten Mal im Paper von (Meuwissen et al., 2001) gezeigt. Diese Autoren haben aufgrund von simulierten Daten gezeigt, dass es mit Hilfe einer sehr dichten Markerkarte möglich ist, die phänotypischen Leistungen alleine aufgrund der geschätzten Wirkungen an den Markergenorten zu modellieren.

Die folgende Abbildung fasst die Unterschiede zwischen dem Infinitesimalmodell und dem polygenen Modell zusammen.



1.3.1 Modellierung

Im Zusammenhang mit der genomischen Selektion besteht die Modellierung der Daten aus zwei Komponenten

1. Die Schätzung der Gen-Wirkungseffekte (a)
2. Die Schätzung der genomischen Zuchtwerte

Die Umsetzung der beiden Komponenten wird in zwei verschiedenen Verfahren gemacht. Im Zwei-Schritt-Verfahren werden beide Komponenten einzeln an verschiedenen Teilen der Zuchtpopulation ausgeführt. Im Gegensatz dazu werden im Single-Step-Verfahren beide Komponenten im gleichen Schritt realisiert.

1.3.2 Zwei-Schritt-Verfahren

Beim Zwei-Schritt-Verfahren wird die Population in ein Trainings- und ein Testset unterteilt. Im Trainingsset werden aufgrund von Typisierungsergebnissen und Beobachtungen die Gen-Wrkungseffekte (a) geschätzt. Sobald die Schätzwerte für die a-Effekte bekannt sind können diese für die Schätzung der genomischen Zuchtwerte verwendet werden.

Da aufgrund der Typisierungsergebnisse die Genotypen an den SNP-Genorten bekannt sind, brauchen wir kein gemischtes lineares Modell mehr. Im Gegensatz zur BLUP-Zuchtwertschätzung, ist in der genomischen Selektion beim Zwei-Schritt-Verfahren ein einfaches lineares Modell ausreichend. Im Idealfall, wenn die komplette Information zu allen Gen-Wirkungseffekten (a) bekannt sind, dann setzen sich die genotypischen Werte einfach zusammen aus den aufsummierten a-Werten. In Matrix-Vektor-Schreibweise können wir die folgende Modellgleichung aufstellen.

$$g = 1\mu + Ma + \epsilon \tag{1.2}$$

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

Die Matrix M ist eine Inzidenzmatrix, welche die genotypischen Werte im Vektor g mit den Gen-Wirkungseffekten a verknüpft. Die Matrix M hat die Dimension $n \times p$ wobei n der Anzahl Individuen mit einem Typisierungsergebnis entspricht und p gleich der Anzahl SNP-Genorte ist.

In der Realität im ersten Schritt des Zwei-Schritt-Verfahrens kennen wir aber weder die Komponenten des Vektors g noch die Gensubstitutionseffekte a. Somit müssen wir das Modell zur Schätzung der a-Effekte modifizieren. Bei der aktuellen Modifikation ersetzen wir den Vektor g durch die phänotypischen Beobachtung y.

$$y = (1\mu + Xb) + Ma + (\epsilon + e) \tag{1.3}$$

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

Das Modell mit den phänotypischen Beobachtungen erlaubt eine Schätzung der a-Effekte. Mit diesem Ansatz gibt es aber zwei Probleme.

1. **Verfügbarkeit**: wirtschaftliche Merkmale wie Milchleistung sind nur beim weiblichen Geschlecht beobachtbar. Somit müsste für die Selektion auf der männlichen Seite wieder auf Nachkommenleistungen zurückgegriffen werden. Dies verlängert aber das Generationenintervall.
2. **Vergleichbarkeit**: Beim Austausch von Information zwischen verschiedenen Ländern sind die phänotypischen Leistungen nicht unbedingt vergleichbar.

Diese beiden Probleme können gelöst werden, wenn anstelle von phänotypischen Leistungen y, geschätzte Zuchtwerte $\hat{g}$ verwendet werden. Das entsprechende Modell sieht dann wie folgt aus.

$$\hat{g} = g + (\hat{g} - g) = 1\mu + Ma + (\epsilon + (\hat{g} - g)) \tag{1.4}$$

1.3.3 Eigenschaften von BLUP-Zuchtwerten

Aufgrund der Eigenschaften von den BLUP-Zuchtwerten $\hat{g}$ führt die Addition der Abweichung $(\hat{g} - g)$ zu einer Reduktion der Varianz. Die Reduktion der Varianz bedeutet, dass $var(\hat{g}) \leq var(g)$ ist. Für BLUP-Zuchtwerte gilt, dass die Covarianz zwischen wahrem und geschätztem Zuchtwert gleich der Varianz der geschätzten Zuchtwerte ist. In Formeln geschrieben bedeutet dass,

$$cov(\hat{g}, g) = var(\hat{g}) \tag{1.5}$$

Setzen wir diese Beziehung in die Varianz der Abweichung $(\hat{g} - g)$ ein, dann erhalten wir

$$var(\hat{g} - g) = var(\hat{g}) + var(g) - 2cov(\hat{g}, g) = var(g) - var(\hat{g}) \geq 0 \tag{1.6}$$

Somit gilt, dass $var(g) \geq var(\hat{g})$ und somit ist die Reduktion der Varianz gezeigt. Im Zusammenhang mit der Varianzreduktion steht auch die zweite Eigenschaft von BLUP-Zuchtwerten, welche uns hier Schwierigkeiten bereitet und zwar handelt es sich dabei um den sogenannten Shrinkage-Effekt. Für einen geschätzten

Zuchtwert eines Tieres i bedeutet das, dass dieser zum Durchschnitt der geschätzten Zuchtwerte der Eltern regressiert wird. Das Ausmass dieses Regressions-Effektes hängt davon ab, aufgrund welcher Informationen der Zuchtwert von Tier i geschätzt wurde. Diese Abhängigkeit wird in der Zerlegung des geschätzten BLUP-Zuchtwertes des Tieres i in seine Komponenten sichtbar. Diese Zerlegung ist in (Hofer, 1990) und in (von Rohr, 2016) erklärt. Das Resultat der Zerlegung ist in der nachfolgenden Formel zusammengefasst.

$$\hat{g}_i = \frac{1}{1+\alpha\delta^{(i)}+\frac{\alpha}{4}\sum_{j=1}^{n}\delta^{(k_j)}}\left[y_i-\hat{\mu}+\frac{\alpha}{2}\left\{\delta^{(i)}(\hat{g}_s+\hat{g}_d)+\sum_{j=1}^{n}\delta^{(k_j)}(\hat{g}_{k_j}-\frac{1}{2}\hat{g}_{l_j})\right\}\right] \tag{1.7}$$

Die Zerlegung des geschätzen Zuchtwertes $\hat{g}_i$ für Tier i zeigt die Abhängigkeit des Ausmasses der Regression von $\hat{g}_i$ auf den Durchschnitt der geschätzten Elternzuchtwerte $\hat{g}_s$ und $\hat{g}_d$. Hat das Tier i keine Eigenleistung y_i, keine Nachkommen und keine Paarungspartner, so ist $\hat{g}_i$ vollständig durch $\hat{g}_s$ und $\hat{g}_d$ bestimmt. Sobald aber Tier i eine Eigenleistung hat und später dann noch Nachkommenleistungen dazukommen, nimmt der Einfluss von $\hat{g}_s$ und $\hat{g}_d$ auf $\hat{g}_i$ ab. Damit verringert sich auch das Ausmass des Regressions-Effektes von $\hat{g}_i$ auf den Durchschnitt der geschätzten Elternzuchtwerte.

Durch die Berücksichtigung zusätzlicher Informationen, wie Eigenleistung und Leistungen von Nachkommen und Paarungsparter, bei der Schätzung des Zuchtwertes für Tier i steigt auch die Genauigkeit oder das Bestimmtheitsmass (B) des geschätzten Zuchtwertes. Wir können aufgrund der Eigenschaften von BLUP-Zuchtwerten können wir folgende Zusammenhänge aufstellen. Je grösser die verfügbare Information für die Schätzung eines Zuchtwertes für Tier i, desto grösser ist das Bestimmtheitsmass des geschätzten Zuchtwertes und je tiefer ist der Regressions-Effekt des geschätzten Zuchtwertes auf den Durchschnitt der geschätzten Zuchtwerte der Eltern und je geringer ist auch die Varianzreduktion.

1.3.4 Einsatz von BLUP-Zuchtwerten in der genomischen Selektion

Eigenschaften von BLUP-Zuchtwerten führen zu Varianzreduktion und dazu dass geschätzte Zuchtwerte zum Durchschnitt der geschätzten Zuchtwerte der Eltern regressiert werden. Diese beiden Effekte sind problematisch bei der Verwendung von BLUP-Zuchtwerten für die Schätzung der a-Effekte in der genomischen Selektion. Ein bestimmtes Tier i hat immer die gleichen SNP-Genotypen und wir gehen davon aus, dass diese auch immer die gleiche Wirkung auf die Ausprägung eines Phänotyps haben. Der mit BLUP geschätzte Zuchtwert eines Tieres ändert sich aber während seines Lebens. In der Zeitperiode der Geburt bis zur Beobachtung einer Eigenleistung ist der geschätzte Zuchtwert durch die geschätzten Zuchtwerte der Eltern bestimmt. Mit zunehmendem Alter werden für Tier i mehr Informationen in der Zuchtwertschätzung berücksichtigt. Somit ändert sich der geschätzte Zuchtwert und damit würde sich auch die aufgrund der BLUP-Zuchtwerte geschätzten a-Effekte ändern. Das ist aufgrund von unserer Annahme der konstanten Wirkung der a-Effekte ein unerwünschtes Verhalten.

Die unerwünschten Veränderungen der geschätzten BLUP-Zuchtwerte werden durch eine Prozedur namens **Deregression** korrigiert. Da sich die Veränderungen der Zuchtwerte im wesentlichen durch eine Funktion der Änderungen im Bestimmtheitsmass beschrieben werden kann, ist die Deregression als Korrektur von geschätzten Zuchtwerten aufgrund deren Bestimmtheitsmass definiert. Einzelheiten zur Deregression können dem Paper (Garrick et al., 2009) entnommen werden.

1.4 Zusammenfassung

Die deregressierten Zuchtwerten werden als Beobachtunen für die Schätzung der a-Effekte im ersten Schritt des Zwei-Schritt-Verfahrens verwendet. Die geschätzen a-Werte werden dann verwendet um im zweiten Schritt die genomischen Zuchtwerte der restlichen Population zu berechnen.

Die im Zwei-Schritt-Verfahren verwendeten Modelle zur Schätzung der a-Effekte sind einfache lineare Modelle. Die Anzahl der Parameter p in diesen Modellen entspricht der Anzahl zu schätzender a-Werte und somit

der Anzahl an SNPs pro Typisierung. Diese Anzahl ist typischerweise bei 50K kann aber auch bis 800K anwachsen. In den meisten Fällen ist $p >> n$, wenn n die Anzahl typisierter Tiere ist. Somit können wir das klassische Least Squares Verfahren für die Schätztung der Parameter nicht verwenden.

1.5 Ausblick

Das Problem $p >> n$ kommt heutzutage in sehr vielen Anwendungen vor. In den nachfolgenden Kapiteln wollen wir uns ein paar Lösungsansätze anschauen, welche uns trotz der spärlich verfügbaren Informationen in den hoch-dimensionalen Parameterräumen, sinnvolle Schätzwerte für die Parameter im Modell liefern kann.

Chapter 2

Multiple Lineare Regression

Das Material dieses Kapitels ist eine Zusammenfassung aus den Vorlesungsunterlagen von (Bühlmann and Mächler, 2014).

Die multiple lineare Regression ist wie folgt definiert. Jedes Individuum i oder jedes Objekt i in einem Datensatz ist charakterisiert durch eine **Zielgrösse** y_i und durch eine Menge von **erklärenden Variablen** $\{x_{i,1}, x_{i,2}, \ldots, x_{i,p}\}$. Zusammengefasst besteht die bekannte Information für jedes Individuum oder jedes Objekt i aus einem Datensatz aus der folgenden Menge

$$\{x_{i,1}, x_{i,2}, \ldots, x_{i,p}, y_i\}$$

Das multiple lineare Regressionsmodell versucht die Zielgrösse bis auf einen zufälligen Restterm ϵ als lineare Funktion der erklärenden Variablen auszudrücken. Unser Ziel besteht in der Schätzung der unbekannten Parameter, welche im Regressionsmodell enthalten sind. Die nachfolgend gezeigte Modellformel soll die Unterscheidung zwischen erklärenden Variablen und unbekannten Parametern verdeutlichen.

$$y_i = \beta_i x_{i,1} + \ldots + \beta_p x_{i,p} + \epsilon_i \qquad (i = 1, \ldots, n) \tag{2.1}$$

Fassen wir die Gleichungen über alle $(i = 1, \ldots, n)$ zusammen und verwenden die Matrix-Vektor-Notation, so sieht das lineare Modell in (2.1) wie folgt aus.

$$y = X\beta + \epsilon \tag{2.2}$$

wobei

y	Vektor der Länge n mit allen Zielgrössen
β	Vektor der Länge p mit unbekannten Parametern
X	Matrix der Dimension $n \times p$ mit erklärenden Variablen
ϵ	Vektor der Länge n mit zufälligen Resteffekten

Die Reste ϵ_i im Modell (2.1) haben wir als zufällige Effekte definiert. Somit müssen wir geeignete Annahmen zur Dichteverteilung der ϵ_i treffen. Meistens gehen wir davon aus, dass die ϵ_i unabhängig sind und der gleichen Verteilung folgen. In der englischsprachigen Literatur wird das mit dem Begriff `independent, identically distributed` (i.i.d.) bezeichnet. Der Erwartungswert und die Varianz der Zufallsvariablen ϵ sind $E[\epsilon_i] = 0$ und $Var(\epsilon_i) = \sigma^2$.

2.1 Beispiele für Lineare Regressionen

2.1.1 Regression mit Achsenabschnitt

Die erste erklärende Variable wir oft als eine Konstante angenommen. Das bedeutet, dass der erste Kolonnenvektor in der Matrix X gleich dem Eins-Vektor ist. Die konstante erklärende Variable erlaubt es einen sogenannten **Achsenabschnitt** anzupassen. In skalarer Schreibweise hat das lineare Modell mit Achsenabschnitt die folgende Form

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \qquad (i = 1, \ldots, n) \tag{2.3}$$

2.1.2 Regression durch den Ursprung

Im Gegensatz zur Regression mit Achsenabschnitt steht die Regression durch den Ursprung. Diese kennt keine konstante erklärende Variable. Das Modell ohne Achsenabschnitt sieht dann wie folgt aus.

$$y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \qquad (i = 1, \ldots, n) \tag{2.4}$$

2.1.3 Regression mit transformierten Variablen

Regressionen können auch auf Transformationen der erklärenden Variablen oder auf transformierte Zielgrössen angepasst werden. Als Beispiel verwendet die sogenannte "quadratische" Regression die x_{ij} und die x_{ij}^2 als erklärende Variablen. Das Modell entspricht dann einer quadratischen Funktion in den x_j ist aber immer noch eine lineare Funktion im Bezug auf die Parameter β_j.

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i2}^2 + \epsilon_i \qquad (i = 1, \ldots, n) \tag{2.5}$$

Abgesehen von der quadratischen Regression sind auch andere Arten von Transformationen der erklärenden Variablen denkbar. Ein Beipsiel ist in der folgenden Gleichung gezeigt.

$$y_i = \beta_i + \beta_2 \log(x_{i2}) + \beta_3 sin(\pi x_{i3}) + \epsilon_i \qquad (i = 1, \ldots, n) \tag{2.6}$$

Auch dieses Modell ist *linear* in den Parametern β_j und wird somit als lineare Regression bezeichnet.

2.1.4 Anwendungen in den Nutztierwissenschaften

Eine Anwendung der linearen Regression in den Nutztierwissenschaften ist die Schätzung vom Lebendgewicht von Tieren aufgrund des Brustumfangs. Dafür werden Messbänder verwendet, welche auf der einen Seite den Brustumfang angeben und auf der anderen Seite das geschätzte Körpergewicht. Diese Anwendung macht eine Voraussage der Zielgrösse `Körpergewicht` aufgrund der beobachteten erklärenden Variablen `Brustumfang`.

Damit eine Voraussage für die Zielgrösse aufgrund der erklärenden Variablen möglich ist, muss zuerst ein angemessener Datensatz vorliegen, in welchem man für jedes Tier beide Informationen, also sowohl Körpergewicht als auch Brustumfang bekannt ist. Aufgrund dieser Informationen können dann die unbekannten Parameter geschätzt werden. Die geschätzten Parameter werden dann für die Vorhersagen verwendet.

Bei diesem ersten Beispiel handelt es sich um eine einfache lineare Regression. Das verwendete Regressionsmodelle hat nur eine erklärende Variable (`Brustumfang`) und eine Zielvariable (`Gewicht`). Das zu dieser Anwendung zugehörige Modell lautet

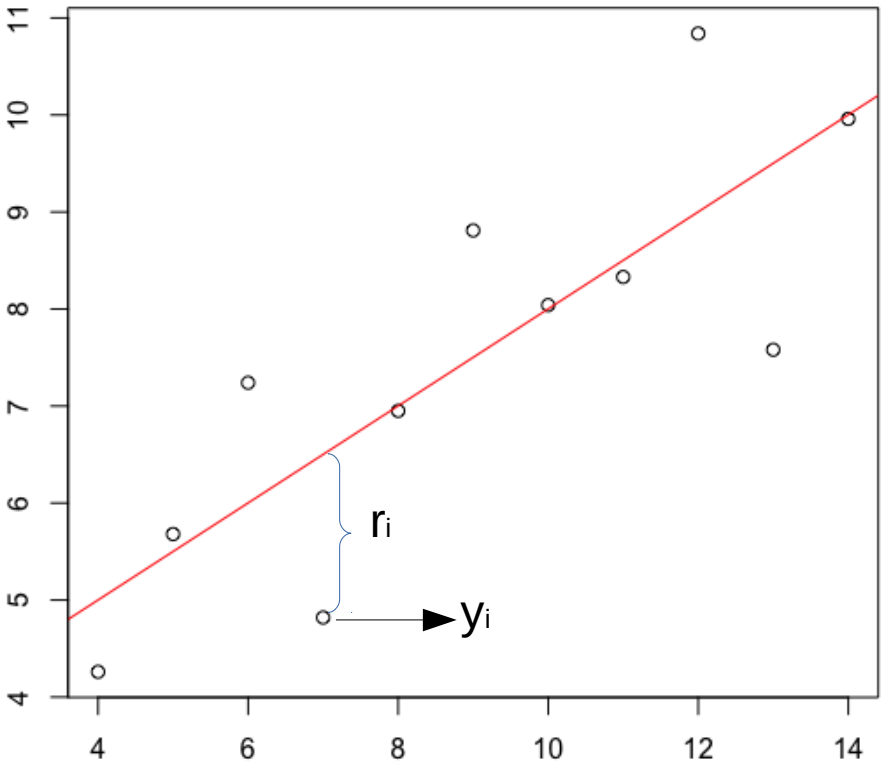
$$y_{G,i} = \beta_1 + \beta_2 x_{B,i} + \epsilon_i \tag{2.7}$$

2.1.5 Ziele der linearen Regression

- **Gute Anpassung**: das Modell soll so sein, dass die erklärenden Variablen möglichst präzise Voraussagen zu den Zielvariablen machen. Das Standardtool für die Anpassung ist die Methode der kleinsten Quadrate (`Least Squares`).
- **Parameterschätzung**: die unbekannten Parameter sollen so geschätzt sein, dass eine Veränderung der erklärenden Variablen in einer entsprechenden Veränderung der Zielgrösse führt.
- **Vorhersage**: noch nicht beobachtete Zielgrössen sollen als Funktionen von erklärenden Variablen vorhergesagt werden können
- **Fehler und Signigikanz**: werden durch Vertrauensintervalle und statistische Tests beurteilt
- **Modellentwicklung**: ist ein interaktiver Prozess, welche durch die oben genannten Ziele beeinflusst wird

2.2 Methode der kleinsten Quadrate (Least Squares)

Gegeben sei das lineare Modell $y = X\beta + \epsilon$. Wir wollen eine, gemäss den oben formulierten Zielen, möglichst gute Schätzung für β finden. Die folgende Darstellung erklärt, wie die Methode der kleinsten Quadrate funktioniert.



Die Punkte stehen für die Beobachtungen y_i. Die rote Linie steht für die Regressionsgerade. Die Distanz des Punktes zur Projektion in Richtung der y-Achse auf der Regressionslinie entspricht dem Residuum $r_i = y_i - x_i^T \hat{\beta}$. Für eine bestimmte Regressionsgerade (rote Linie im Diagramm) wird für jeden Punkt y_i das entsprechende Residuum r_i berechnet. Die Residuen r_i werden quadriert und addiert. Diese summierten

Quadrate der Residuen stellt ein Mass dar, wie gut die Regressionsgerade an die Beobachtungspunkte y_i angepasst ist.

Position und Verlauf der Regressionsgeraden können durch die Wahl des Vektors β beeinflusst werden. Gemäss der Methode der kleinsten Quadrate soll β so bestimmt werden, dass die Summe der quadrierten Residuen minimal wird. Der so bestimmte Vektor β wird dann als Least-Squares-Schätzer bezeichnet. In einer Formel können wir die Berechnung des Least-Squares-Schätzers ($\hat{\beta}$), wie folgt ausdrücken.

$$\hat{\beta} = argmin_{\beta} \|y - X\beta\|^2 \tag{2.8}$$

wobei $\|.\|$ für die Euklidsche Norm oder die Euklidsche Distanz steht. In einem ersten Schritt geht es darum das Minimum für den Ausdruck $\|y - X\beta\|^2$ zu finden. Dabei ist es einfacher, wenn wir folgende Umformung verwenden.

$$\|y - X\beta\|^2 = (y - X\beta)^T(y - X\beta) = y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \tag{2.9}$$

Leiten wir diesen Ausdruck nach β ab und setzen die erste Ableitung gleich 0, dann erhalten wir eine Gleichung für den Least-Squares-Schätzer $\hat{\beta}$.

$$-y^T X - y^T X + 2\hat{\beta}^T X^T X = 0 \tag{2.10}$$

Aus der obigen Formel können wir die sogenannte **Normalgleichung** herleiten. Diese lautet

$$X^T X\hat{\beta} = X^T y \tag{2.11}$$

Unter der Annahme, dass die Matrix X vollen Kolonnenrang p hat, können wir explizit nach $\hat{\beta}$ auflösen.

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2.12}$$

Die Residuen $r_i = y_i - x_i^T\hat{\beta}$ sind Schätzungen für die Resteffekte ϵ_i und können somit für die Schätzung von σ^2 verwendet werden.

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 \tag{2.13}$$

Der Faktor $1/(n-p)$ scheint ungewöhnlich, aber es kann gezeigt werden, dass die Wahl dieses Faktors zur Erwartungstreue von $\hat{\sigma}^2$ führt. Das heisst, es gilt $E\left[\hat{\sigma}^2\right] = \sigma^2$.

2.2.1 Annahmen hinter dem linearen Modell

Abgesehen davon, dass die Matrix X vollen Kolonnenrang $p < n$ haben muss, wurden für die erklärenden Variablen keine Annahmen getroffen. Insbesondere können die erklärenden Variablen kontinuierlich oder diskret sein. Kontinuierliche Variablen sind typischerweise Messgrössen, welche als reelle Zahlen (Gleitkommazahlen) erhoben werden. Diskrete Grössen können nur bestimmte Werte, wie zum Beispiel 0 oder 1 annehmen.

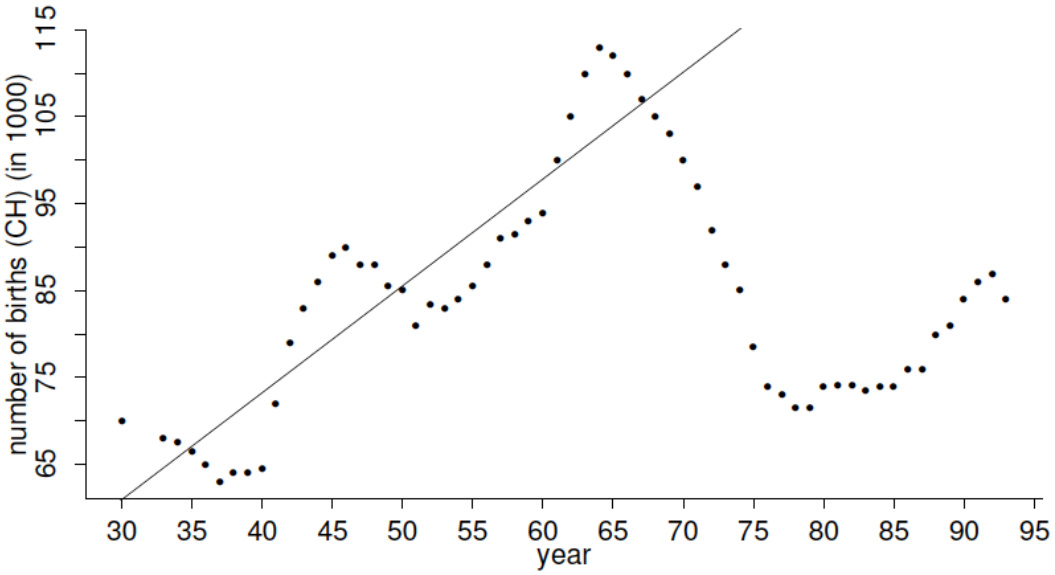
Damit die Anpassung eines linearen Modells mit der Methode der kleinsten Quadrate Sinn macht und die Tests und Vertrauensintervalle gültig sind, müssen wir gewisse Annahmen treffen.

1. **Korrektheit des linearen Modells**: Das heisst $E[\epsilon_i] = 0$ für alle i. Das heisst aber auch, dass die Zielgrössen und die erklärenden Variablen nicht gemischt werden dürfen.
2. **Alle x_i sind exakt**: Es wird angenommen, dass die Werte für x_i ohne Fehler beobachtet werden können.

3. **Konstante Varianz der Resteffekte**: $Var(\epsilon_i) = \sigma^2$ für alle i
4. **Resteffekte sind unkorreliert**: $Cov(\epsilon_i, \epsilon_j) = 0$ für alle $i \neq j$
5. **Resteffekte folgen Normalverteilung**: Der Vektor ϵ der Resteffekte folgt einer multivariaten Normalverteilung.

Falls diese Annahmen verletzt sind, gibt es eine Reihe von Massnahmen, welche getroffen werden können. Bei Verletzung der Annahme 3, können "weighted least squares" Methoden verwendet werden. Ähnlich bei Verletzung der Annahme 4, können wir "generalized least squares" verwenden. Ist die Annahme 5 der Normalverteilung nicht erfüllt, können wir auf sogenannte "robuste Methoden" ausweichen. Falls Annahme 2 nicht zutrifft, braucht es Korrekturen, welche als "errors in variables" bezeichnet wird. Falls die Annahme 1 nicht stimmt, braucht es nicht-lineare Modelle.

Die folgende Grafik zeigt das Beispiel des sogenannten "Pillen-Knicks". Dabei werden die Anzahl Geburten seit 1930 in der Schweiz gezeigt. Hier sind die Annahmen 1 und 4 verletzt. Dieses Beispiel zeigt auch die Gefahr bei Vorhersagen in Bereiche, wo keine erklärende Variablen vorliegen.



2.2.2 Kein Ersatz der multiplen Regression durch mehrere einfache Regressionen

Eine multiple Regression (mit mehreren erklärenden Variablen) soll nicht durch mehrere einfache Regressionen (mit nur einer erklärenden Variablen) ersetzt werden. Das folgende simulierte Beispiel zeigt weshalb.

Wir betrachten die folgenden erklärenden Variablen $x^{(1)}$ und $x^{(2)}$ und die Zielgrösse y mit folgenden Werten

x1	x2	y
0	-1	1
1	0	2
2	1	3
3	2	4
0	1	-1
1	2	0
2	3	1
3	4	2

Die multiple Regression führt zur Lösung der kleinsten Quadrate, welche die Daten exakt beschreibt, so wie diese erzeugt wurden.

$$y_i = \hat{y}_i = 2x_{i1} - x_{i2} \qquad \text{für alle } i \text{ mit } \hat{\sigma}^2 = 0 \tag{2.14}$$

Wird an die Daten nur eine einfache Regression mit der erklärenden Variablen $x^{(2)}$ und ignoriert $x^{(1)}$, so erhalten wir das folgende Resultat

$$\hat{y}_i = \frac{1}{9}x_{i2} + \frac{4}{3} \qquad \text{für alle } i \text{ mit } \hat{\sigma}^2 = 1.72 \tag{2.15}$$

Der Grund dafür ist, dass die erklärenden Variablen $x^{(1)}$ und $x^{(2)}$ korreliert sind. Falls $x^{(1)}$ steigt, dann steigt auch $x^{(2)}$. Da aber in der multiplen Regression $x^{(1)}$ einen grösseren Koeffizienten hat als $x^{(2)}$, muss dieser Effekt in der einfachen Regression durch $x^{(2)}$ kompensiert werden. Dies führt zur Abweichung zwischen den Resultaten der beiden Analysen.

2.3 Eigenschaften der Schätzungen

Die Least-Squares-Schätzer sind Zufallsvariablen, da für jeden Datensatz den wir vom gleichen unterliegenden Prozess beobachten, andere Werte resultieren. Damit ändern sich auch die Least-Squares-Schätzer. Da die Schätzer Funktionen der beobachteten Daten sind, haben die Schätzer auch einen zufälligen Charakter. Somit können wir Eigenschaften betreffend den Verteilungen und den Momenten für die Least-Squares-Schätzer herleiten. Die Ergebnisse sind hier nur kurz zusammengefasst.

2.3.1 Momente der Least-Squares Schätzungen

Wir nehmen das folgende lineare Modell an

$$y = X\beta + \epsilon,\ E[\epsilon] = 0,\ Cov(\epsilon) = E\left[\epsilon\epsilon^T\right] = \sigma^2 I_{n\times n} \tag{2.16}$$

Zusammen mit den oben getroffenen Annahmen können wir folgende Aussagen machen

1. $E\left[\hat{\beta}\right] = \beta$, das heisst, $\hat{\beta}$ ist unverzerrt
2. $E[\hat{y}] = E[y] = X\beta$, was aus 1. folgt. Zudem ist, $E[r] = 0$
3. $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$
4. $Cov(\hat{y}) = \sigma^2 P$, $Cov(r) = \sigma^2(I - P)$

Die Matrix P ist definiert als Projektionsmatrix aus $\hat{y} = Py$. Setzen wir die Least-Squares-Schätzer ein, dann folgt

$$\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Py \tag{2.17}$$

Somit ist die Matrix P definiert als $P = X(X^TX)^{-1}X^T$.

2.3.2 Verteilung der Least-Squares-Schätzer unter normalverteilten Fehlern

Zusätzlich zum linearen Modell nehmen wir an, dass $\epsilon_i, \ldots, \epsilon_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$, dann können wir zeigen, dass

1. $\hat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2(X^TX)^{-1}\right)$
2. $\hat{y} \sim \mathcal{N}_n\left(X\beta, \sigma^2 P\right)$, $r \sim \mathcal{N}_n\left(0, \sigma^2(I - P)\right)$
3. $\hat{\sigma}^2 \sim \frac{n}{n-p}\chi^2_{n-p}$

Die Annahme der Normalverteilung ist oft (annähernd) erfüllt und kann durch den zentralen Grenzwertsatz bei grösseren Datensätzen begründet werden. Diese Eigenschaften im Bezug auf die Verteilung der Schätzer führt zur Herleitung von Vertrauensintervallen und statistischen Tests für die geschätzten Parameter. Sind die Annahmen der Normalverteilung nicht erfüllt, müssen wir auf sogenannte robuste Methoden ausweichen. Diese werden hier nicht behandelt.

2.4 Tests und Vertrauensintervalle

2.4.1 Einzeltests

Wir nehmen an, dass das lineare Modell korrekt ist und dass die Resteffekte $\epsilon_1, \ldots, \epsilon_n$ i.i.d. $\sim \mathcal{N}\left(0, \sigma^2\right)$. Dann haben wir gesehen gemäss den Eigenschaften aus dem vorherigen Abschnitt ist dann $\hat{\beta}$ normalverteilt.

Im Allgemeinen sind wir daran interessiert, ob ein bestimmter Parameter β_j einen Einfluss hat. Dies lässt sich mit der Nullhypothese $H_{0,j} : \beta_j = 0$ gegenüber der Alternativen $H_{A,j} : \beta_j \neq 0$ überprüfen. Da $\hat{\beta}$ einer Normalverteilung folgt, können wir herleiten, dass unter der Nullhypothese $H_{0,j}$ gilt

$$\frac{\hat{\beta}_j}{\sqrt{\sigma^2 (X^T X)^{-1}_{jj}}} \sim \mathcal{N}(0, 1) \tag{2.18}$$

Da σ^2 unbekannt ist, ist die obige Teststatistk in der Praxis nicht brauchbar. Ersetzen wir σ^2 durch den Schätzwert $\hat{\sigma}^2$ so erhalten wir die sogenannte t-Teststatistik.

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{jj}}} \sim t_{n-p} \tag{2.19}$$

Anhand dieses Tests können wir die Relevanz der erklärenden Variablen quantifizieren, indem wir die Teststatistiken T_j für $j = 1, \ldots, p$ analysieren. Die Beurteilung der Relevanz der erklärenden Variablen aufgrund dieser einzelnen t-Tests birgt zwei Probleme.

1. **Multiples Testen**: Werden sehr viele Tests durchgeführt, dann sind bei einem angenommenen Signifikanz-Niveau von α automatisch ein Anteil α aller Tests signifikant. Werden beispielsweise 100 Tests auf dem Niveau $\alpha = 0.05$ durchgeführt, dann sind automatisch 5 Tests signifikant.
2. **Korrelation der erklärenden Variablen**: Falls die erklärenden Variablen untereinander korreliert sind, dann beeinflusst dies auch die Testergebnisse und kann diese verzerren.

2.4.2 Globaler Test

Wenn wir testen wollen, ob (abgesehen vom Achsenabschnitt) überhaupt eine erklärende Variable einen Einfluss auf die Zielgrösse hat, dann können wir diese mit folgender Nullhypothese $H_0 : \beta_2 = \ldots = \beta_p = 0$ versus die Alternative $H_A : \beta_j \neq 0$ für $j = 2, \ldots, p$ tun. Solch ein Test kann mit der Zerlegung der Varianz der Beobachtungen y_i um das globale Mittel $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ konstruiert werden. In Vektor-Schreibweise sieht diese Zerlegung wie folgt aus

$$\|y - \bar{y}\|^2 = \|\hat{y} - \bar{y}\|^2 + \|y - \hat{y}\|^2 \tag{2.20}$$

Diese Zerlegung teilt die quadrierten Abweichungen der Beobachtungen y vom allgemeinen Mittel $\bar{y}$ in die quadrierten Abweichungen der gefitteten Werte $\hat{y}$ vom allgemeinen Mittel plus die quadrierten Residuen $y - \hat{y}$ auf. Eine solche Zerlegung lässt sich am einfachsten in einer Varianzanalysetabelle zusammenfassen.

	Summenquadrate	Freiheitsgrade	mittlere Summenquadrate
Regression	$\|\hat{y} - \bar{y}\|^2$	$p - 1$	$\|\hat{y} - \bar{y}\|^2/(p-1)$
Rest	$\|y - \hat{y}\|^2$	$n - p$	$\|y - \hat{y}\|^2/(n-p)$
Total	$\|y - \bar{y}\|^2$	$n - 1$	

Im Falle der globalen Nullhypothese haben die erklärenden Variablen keinen Einfluss auf die Zielgrösse. Somit ist $E[y] = const. = E[\bar{y}]$. Daraus folgt, dass der Erwartungswert der mittleren Summenquadrate der Regression gleich σ^2 ist. Teilen wir die mittleren Summenquadrate der Regression durch die mittleren Summenquadrate des Rests (Schätzung von σ^2) erhalten wir eine dimensionslose Grösse, welcher einer F-Statistik entspricht. Unter der Nullhypothese gilt, dass

$$F = \frac{\|\hat{y} - \bar{y}\|^2/(p-1)}{\|y - \hat{y}\|^2/(n-p)} = F_{p-1,n-p} \tag{2.21}$$

Dies wird als globaler F-Test der Regression bezeichnet.

Abgesehen von der Bewertung der statistischen Signifikanz mit dem globalen F-Test, sind wir auch daran interessiert, wie gut die Anpassung des Modells an die Daten ist. Eine mögliche Grösse für die Qualität der Anpassung ist das sogenannten R^2. Dies ist definiert als das folgende Verhältnis.

$$R^2 = \frac{\|\hat{y} - \bar{y}\|^2}{\|y - \bar{y}\|^2} \tag{2.22}$$

Das R^2 entspricht dem Verhältnis der Variation der Beobachtungen um das globale Mittel, welcher durch die Regression erklärt werden kann. Aus dieser Definition ist klar, dass wir nach Modellen suchen mit einem möglichst grossen R^2.

2.4.3 Vertrauensintervalle

In Anlehnung an den t-Test der einzelnen Parameter β_j können wir Vertrauensintervalle ableiten. Das zwei-seitige Vertraunensintervall auf dem Niveau $1 - \alpha$ für β_j ist definiert als

$$\hat{\beta}_j \pm \sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{jj}} \cdot t_{n-p;1-\alpha/2} \tag{2.23}$$

Hier $t_{n-p;1-\alpha/2}$ ist das $1 - \alpha/2$-Quantil der t_{n-p}-Verteilung.

2.5 Output von R

In R wird eine lineare Regression mit der Funktion `lm()` angepasst. Die Zusammenfassung der Resultate von `lm()` ist in der nachfolgenden Diagramm gezeigt.

```
Call:
lm(formula = LOGRUT ~ ., data = asphalt1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48348 -0.14374 -0.01198  0.15523  0.39652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.781239   2.459179  -2.351 0.027280 *
LOGVISC     -0.513325   0.073056  -7.027 2.90e-07 ***
ASPH         1.146898   0.265572   4.319 0.000235 ***
BASE         0.232809   0.326528   0.713 0.482731
RUN         -0.618893   0.294384  -2.102 0.046199 *
FINES        0.004343   0.007881   0.551 0.586700
VOIDS        0.316648   0.110329   2.870 0.008433 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2604 on 24 degrees of freedom
Multiple R-Squared: 0.9722,     Adjusted R-squared: 0.9653
F-statistic: 140.1 on 6 and 24 DF,  p-value: < 2.2e-16
```

1 2 3 4

Die verschiedenen Bereiche der Resultate sind numeriert durch farbige Rechtecke gekennzeichnet. Im ersten Bereich ist unter der Überschrift `Call` der Funktionsaufruf nochmals aufgeführt. So ist dokumentiert, wie die nachfolgenden Resultate zustande kamen. Der zweite Bereich gibt einige Informationen zur empirischen Verteilung der Residuen. Diese Kennzahlen der Residuen-Verteilung sind nützlich um gewisse Annahmen bezüglich der Residuen im Modell grob überprüfen zu können. Block drei enthält die Resultate der Parameterschätzungen. Abgesehen von den Schätzwerten sind auch die Standardfehler und die Quantile des entsprechenden t-Tests für jeden Parameter enthalten. Die Kolonne ganz links im Block drei zeigt das Signifikanz-Niveau der t-Tests für jeden Parameterschätzwert. Der vierte und letzte Block enthält die Schätzung der Rest-Standardabweichung (residual standard error) und das Testergebnis des globalen F-Tests. Zur Beurteilung der Anpassungsqualität ist das R^2 und eine korrigierte Version des R^2 aufgeführt. Die korrigierte Version des R^2 berücksichtigt die Anzahl der geschätzten Parameter und ist definiert als

$$\bar{R}^2 = R^2 - (1 - R^2)\frac{p-1}{n-p} \tag{2.24}$$

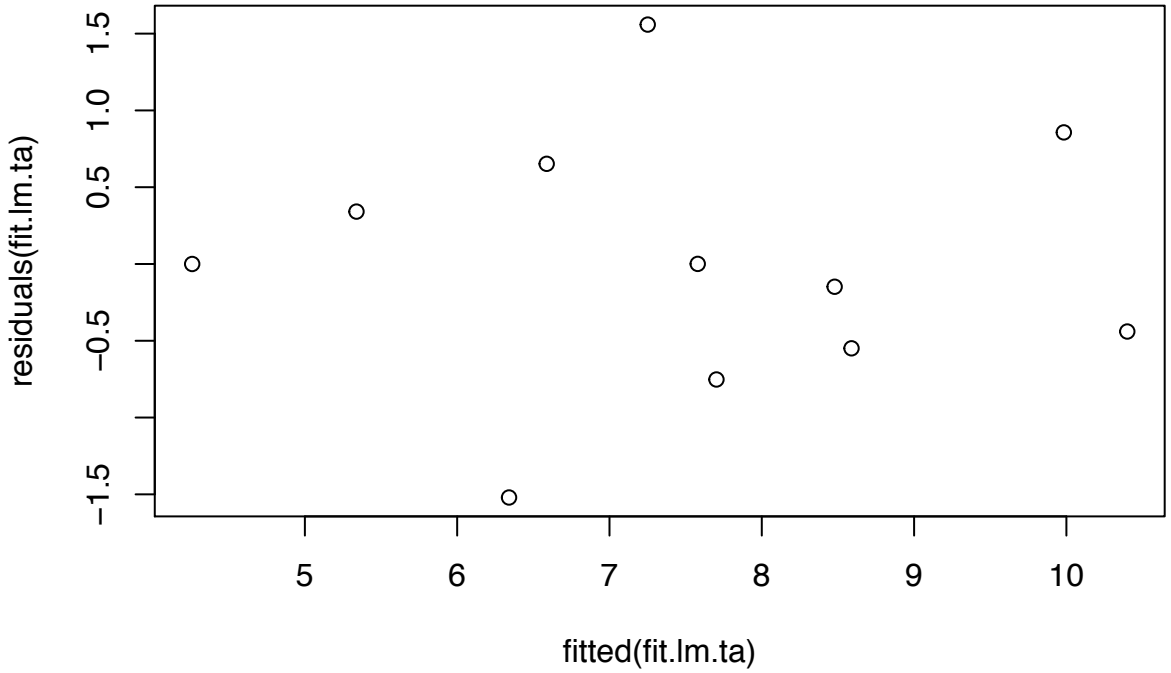
2.6 Analyse der Residuen und Überprüfung der Modellannahmen

Die Residuen $r_i = y_i - \hat{y}_i$ dienen als Annäherungen an die unbekannten Resteffekte ϵ_i und zur Überprüfung der Modellannahmen.

2.6.1 Tukey-Anscombe Plot

Der Tukey-Anscombe Plot ist ein graphisches Tool zur Feststellung von Abhängigkeiten zwischen den Residuen r_i und den gefitteten Werten $\hat{y}_i$. Im Tukey-Anscombe Plot werden auf der x-Achse die gefitteten Werte und auf der y-Achse die Resiuden aufgetragen. Idealerweise sind die Resiuden zufällig verteilt und zeigen kein Muster. In R erzeugt man den Tukey-Anscombe Plot über die Hilfsfunktionen `fitted()` und `residuals()`. Die Ergebnisse der beiden Funktionen werden einfach an die `plot()`-Funktion übergeben.

```
data("anscombe")
fit.lm.ta <- lm(y1 ~ ., data = anscombe)
plot(fitted(fit.lm.ta), residuals(fit.lm.ta))
```



Der obige Plot zeigt eine ideale Situation, wo keine systematischen Muster zu erkennen sind. Die folgenden vier Plots sind (Bühlmann and Mächler, 2014) entnommen und zeigen Probleme bei der Anpassung von linearen Modellen auf.

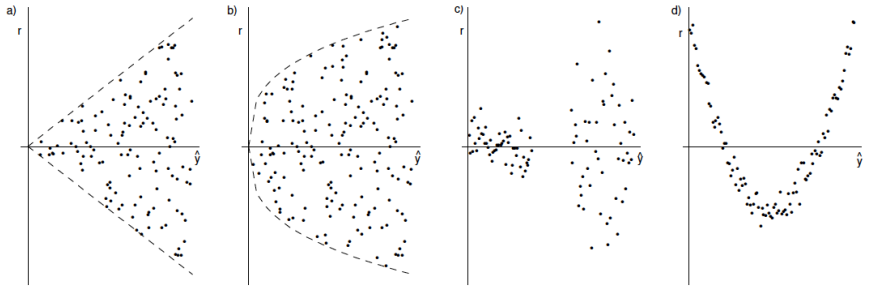


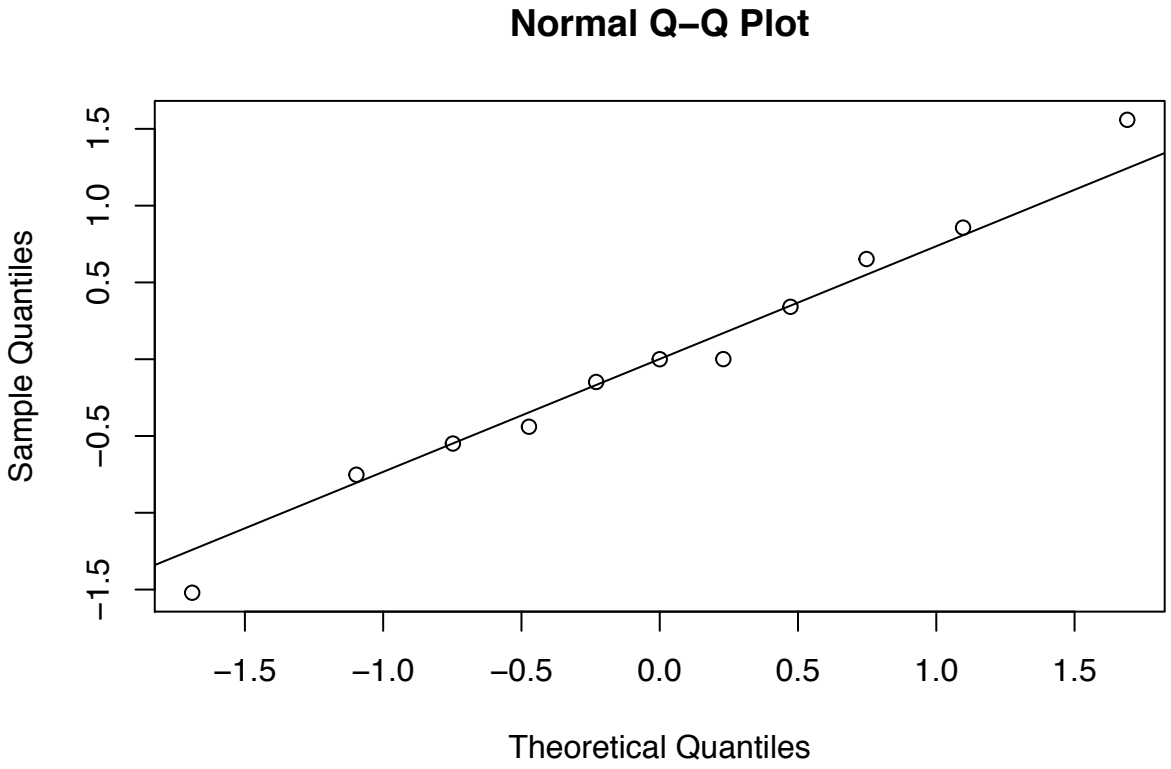
Figure 1.5: a) linear increase of standard deviation, b) nonlinear increase of standard deviation, c) 2 groups with different variances, d) missing quadratic term in the model.

2.6.2 Der QQ-Plot

Annahmen zur Verteilung der zufälligen Grössen im linearen Modell können mit dem sogenannten QQ-Plot überprüft werden. Die Abkürzung "QQ" steht hier für Quantil-Quantil und meint, dass wir die empirischen Quantile den theoretischen Quantilen einer bestimmten Verteilung gegenüberstellen. Im Fall, dass wir gegen die theoretischen Quantile einer Normalverteilung testen, heisst der QQ-Plot auch Normal Plot.

In R können wir den QQ-Plot für die Residuen, wie folgt erzeugen.

```
qqnorm(residuals(fit.lm.ta))
qqline(residuals(fit.lm.ta))
```



Stimmen die Quantile der empirische Verteilung der Residuen gut mit den theoretischen Quantilen überein, dann liegen die Punkte im QQ-Plot auf einer Geraden. Falls die empirische Verteilung bedeutende Abweichungen zeigt von der angenommenen Verteilung, so zeigt der Verlauf der Punkte systematische Abweichungen, wie das in den folgenden Graphiken der Fall ist.

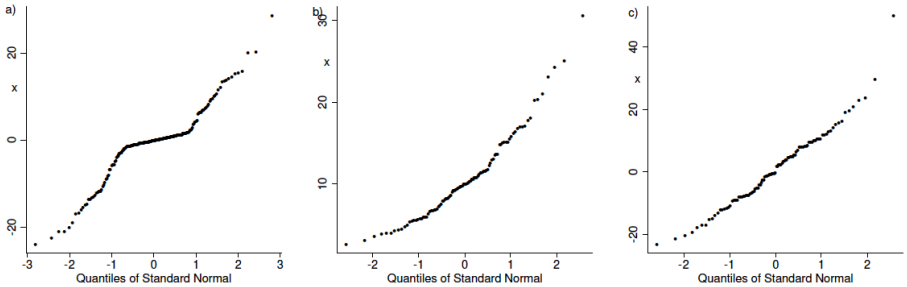


Figure 1.7: QQ-plots for a) long-tailed distribution, b) skewed distribution, c) dataset with outlier.

2.7 Selektion eines Modells

Gegeben sei das lineare Modell

$$y_i = \sum_{i=1}^{n} \beta_j x_{ij} + \epsilon_i \quad (i = 1, \ldots, n) \tag{2.25}$$

mit $\epsilon_1, \ldots, \epsilon_n$ i.i.d. , $E[\epsilon_i] = 0$ und $Var(\epsilon_i) = \sigma^2$.

Bis anhin hatten wir immer alle erklärenden Variablen $x_1, \ldots, x_p$ im Modell berücksichtigt. Wir können uns aber auch fragen, ob dies Sinn macht, wenn gewisse erklärende Variablen nicht relevant sind für die Modellierung der Zielgrössen. Hinzu kommt noch, dass für jede erklärende Variablen ein unbekannter Parameter β_j geschätzt werden muss. Jeder geschätzte Parameter trägt zur Variabilität der gefitteten Werte bei, ob die erklärende Variable relevant ist oder nicht. Somit wird oft nach dem **optimalen** Modell und nicht nach dem wahren Modell gesucht. Das optimale Modell ist definiert als das Modell mit dem minimalen Subset an erklärenden Variablen, welche alle relevant sind für die Modellierung der Zielgrösse.

Formell können wir das soeben Erklärte wie folgt zusammenfassen. Angenommen, wir wollen die folgende Vorhersage, nennen wir sie $\mathcal{M}$, optimieren.

$$\sum_{r=1}^{q} \hat{\beta}_{j_r} x_{ij_r} \tag{2.26}$$

welche die q erklärenden Variablen mit den Indices $j_1, \ldots, j_q \in \{1, \ldots, p\}$ enthält. Wir brauchen ein Entscheidungskriterium um die Vorhersage $\mathcal{M}$ mit den q Parametern mit dem vollen Modell, welches alle erklärenden Variablen enthält vergleichen zu können.

2.7.1 Mallows C_p-Statistik

Die Summenquadrate der Residuen $SSE(\mathcal{M})$ für die Vorhersage $\mathcal{M}$ können wir nicht als Kriterium verwenden, denn $SSE(\mathcal{M})$ nimmt ab mit zunehmener Anzahl q an Parametern. Der mittlere quadrierte Fehler bei der Verwendung von $\mathcal{M}$ anstelle vom vollen Modell, kann als

$$n^{-1} SSE(\mathcal{M}) - \hat{\sigma}^2 + 2\hat{\sigma}^2 q/n \tag{2.27}$$

geschätzt werden, wobei $\hat{\sigma}^2$ der geschätzte Restvarianz aus dem vollen Modell entspricht. Da n und $\hat{\sigma}^2$ für alle Submodelle $\mathcal{M}$ konstant sind, können wir als Kriterium für den Modellvergleich die Statistik

$$C_p(\mathcal{M}) = \frac{SSE(\mathcal{M})}{\hat{\sigma}^2} - n + 2q \tag{2.28}$$

verwenden.

2.7.2 Modellwahl mit dem C_p-Kriterium

Für das volle Modell mit p erklärenden Variablen gibt es $2^p - 1$ Submodelle oder Vorhersagen $\mathcal{M}$. Somit ist ein Vergleich der C_p-Statistik aller Submodelle nur machbar, wenn p nicht zu gross, d.h. kleiner als 16 ist. Für $p \geq 16$ werden die folgenden zwei schrittweisen Algorithmen vorgeschlagen.

2.7.2.1 Vorwärts-Integration ("Forward Selection) 1. Starte mit dem minimalen Modell $\mathcal{M}_0$, welches nur ein globales Mittel enthält 2. Wähle die erklärende Variable, welche die Summe der quadrierten Residuen am meisten reduziert und nimm diese ins Modell auf 3. Wiederhole Schritt 2 bis alle erklärenden Variablen im Modell aufgenommen wurden. Das produziert eine Sequenz von Submodellen $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \ldots$. 4. Wähle aus der Sequenz der Submodelle dasjenige mit dem kleinsten Wert der C_p-Statistik

2.7.2.2 Rückwärts-Elimination (Backward Selection)

1. Wir starten mit dem vollen Modell, welches alle erklärenden Variablen enthält
2. Entferne die erklärende Variable vom vollen Modell, welche die Summe der quadrierten Residuen am wenigsten reduziert.
3. Wiederhole Schritt 2 bis alle erklärenden Variablen entfernt wurden. Das führt zu einer Sequenz von Submodellen.
4. Wähle dasjenige Submodell aus der Sequenz an Submodellen mit minimaler C_p-Statistik

2.7.3 Bemerkungen

Rückwärts-Elimination von erklärenden Variablen funktioniert im allgemeinen besser als Vorwärts-Integration, aber ist auch teuerer im Bezug auf Rechenleistung. In (Bühlmann and Mächler, 2014) wird die Vorwärts-Integration für den Fall dass $p \geq n$ als taugliche Methode bezeichnet. Erfahrungen im Bereich der Effektschätzung in der genomischen Selektion haben aber gezeigt, dass Vorwärts-Integration zu keiner stabilen Prozedur für die Selektion eines guten Modells führt.

Schon die Autoren in (Meuwissen et al., 2001) haben für simulierte Daten gezeigt, dass die Vorwärts-Integration von SNP-Effekten als erklärende Variablen bei der Identifikation der wichtigen SNP-Effekte versagte. Offenbar gibt es bie einer sehr grossen Anzahl von erklärenden Variablen p im Vergleich zur Anzahl der verfügbaren Beobachtungen n das Problem, dass im Schritt 2 der Vorwärts-Integration viele erklärende Variablen die Summe der quadrierten Residuen um mehr oder weniger den gleich Betrag reduzieren. Dann haben wir das Problem, dass wir eine Auswahl zwischen fast gleichwertigen erklärenden Variablen treffen müssen. Diese Auswahl ist offensichtlich kritisch und kann zu sehr verschiedenen Endergebnissen in der Modellwahl führen.

Chapter 3

Genomic Best Linear Unbiased Prediction

Chapter 4

Least Absolute Shrinkage And Selection Operator (LASSO)

Das lineare Modell

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \tag{4.1}$$

für eine Beobachtung i $(i = 1, \ldots, n)$ wird zur Modellierung von Zusammenhängen zwischen den erklärenden Variablen $x_{i1}, \ldots, x_{ip}$ und der Zielgrösse y_i verwendet. In einem Regressionsmodell werden die unbekannten Parameter β_j $(j = 0, \ldots, p)$ mit Least Squares geschätzt.

Die $(p+1)$ Werte $\beta_0, ..., \beta_p$ und die Resteffekte ϵ_i sind unbekannt. Es wird angenommen, dass die Werte der erklärenden Variablen $(x_{i1}, x_{i2}, ..., x_{ip})$ exakt, d.h. ohne Messfehler oder andere Ungenauigkeiten, bekannt sind. Für einen Datensatz mit n Beobachtungen werden die resultierenden n Gleichungen vorzugsweise in Matrix-Vektor-Schreibweise notiert.

$$y = X\beta + \epsilon \tag{4.2}$$

4.1 Stochastische Restkomponente

Die n unbekannten Resteffekte im Vektor ϵ werden als zufällige Effekte modelliert, wobei angenommen wird, dass sich diese Resteffekte im Mittel aufheben, d.h., dass deren Erwartungswert $E(\epsilon) = 0$ ist. Die Streuung der Resteffekte wird im Standardmodell als konstant angenommen. Für die Covarianz des Vektors der Resteffekte bedeutet das, dass $var(\epsilon) = I * \sigma^2$ ist. Die Varianzkomponente σ^2 ist neben den Koeffizienten im Vektor β ein weiterer unbekannter Parameter, welcher von den Daten geschätzt werden muss.

4.2 Parameterschätzung

Unter der Annahme, dass die Matrix X vollen Kolonnenrang hat, d.h. die Anzahl Beobachtungen n grösser ist als die Anzahl Parameter (hier $p + 1$) lassen sich die unbekannten Parameter β mit **Least Squares** schätzen. Der Least Squares Schätzer $\hat{\beta}$ für β wird berechnet aus

$$\hat{\beta} = argmin_\beta ||y - X\beta||^2 \tag{4.3}$$

wobei $||.||$ für die Euklidsche Norm (Länge) im n-dimensionalen Raum steht. Wird das Minimierungsproblem in Gleichung (4.3) aufgelöst, dann resultiert der folgende Ausdruck für $\hat{\beta}$

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{4.4}$$

Betrachten wir den Ausdruck in Gleichung (4.4) wird klar, weshalb die Matrix X vollen Kolonnenrang haben muss, da nur so die Inverse $(X^T X)^{-1}$ berechnet werden kann.

4.3 Alternativen zu Least Squares

Das lineare Modell (4.1) erweist sich in der Praxis als sehr brauchbar. Mit der Least Squares-Technik besteht auch eine einfache und sehr gut etablierte Methode zur Parameterschätzung. In kürzerer Vergangenheit auch mit dem Aufkommen des Phänomes von "Big Data", welches das systematische Sammeln von grossen Datenmengen ermöglicht, treten häufiger Probleme auf, bei welchen die im einleitenden Abschnitt aufgestellte Bedingung an Least Squres ($n > p$) nicht zutrifft.

Da wir die positiven Eigenschaften des linearen Modells gerne beibehalten möchten, wurde nach Alternativen zu Least Squres gesucht. Diese möglichen Alternativen können in drei Kategorien eingeteilt werden.

1. **Subset Selektion**: Aus den p erklärenden Variablen wird ein Subset von "relevanten" Variablen ausgewählt. Alle anderen Variablen werden ignoriert. Die relevanten Variablen werden oft aufgrund der Signifikanz des geschätzten Regressionskoeffizienten β_j identifiziert.
2. **Regularisierung (Shrinkage)**: Alle p erklärenden Variablen werden verwendet. Die geschätzten Regressionskoeffizienten werden durch bestimmte Techniken gegen den Nullpunkt "gedrückt". Dieser Prozess wird als Schrumpfung (Shrinkage) bezeichnet. Die so erzeugte Reduktion der Variabilität der Schätzwerte wird als Regularisierung bezeichnet.
3. **Dimensionsreduktion**: Die p erklärenden Variablen werden zu m Linearkombinationen reduziert. Diese Reduktion kann mit Techniken, wie Principal Components Analysis oder Faktoranalyse gemacht werden.

4.4 Lasso

Es gibt Schätzverfahren, welche mehrere der oben genannten Alternativen zu Least Squares kombinieren. Ein Beispiel dafür ist LASSO. LASSO steht für Least Absolute Shrinkage and Selection Operation und kombiniert "Subset Selection" und Regularisierung. Die Regularisierung wird durch das Hinzufügen eines Terms zu den Rest-Summenquadraten (RSS), welche bei Least Squares minimiert werden. In Gleichung (4.3) haben wir gesehen, wie RSS verwendet werden zur Berechnung der Least Squares Schätzer

$$\begin{aligned}
\hat{\beta}_{LS} &= argmin_\beta ||y - X\beta||^2 \\
&= argmin_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \\
&= argmin_\beta RSS
\end{aligned} \tag{4.5}$$

4.4.1 Regularisierung bei LASSO

Bei LASSO wird nun zu RSS ein sogenannter Strafterm (penalty term) hinzugefügt. Dieser Strafterm beträgt $\lambda \sum_{j=1}^{p} |\beta_j|$. Der Term wird deshalb als Strafterm bezeichnet, weil er mit steigender Summe der Absolutbeträge aller β_j immer grösser wird. Diese führt zum gewünschten Effekt der Regularisierung. Das

heisst durch das Hinzufügen dieses Strafterms werden die Absolutbeträge und somit die Variabilität der Koeffizientenschätzungen begrenzt, was der eigentliche Sinn und Zweck der Regularisierung ist.

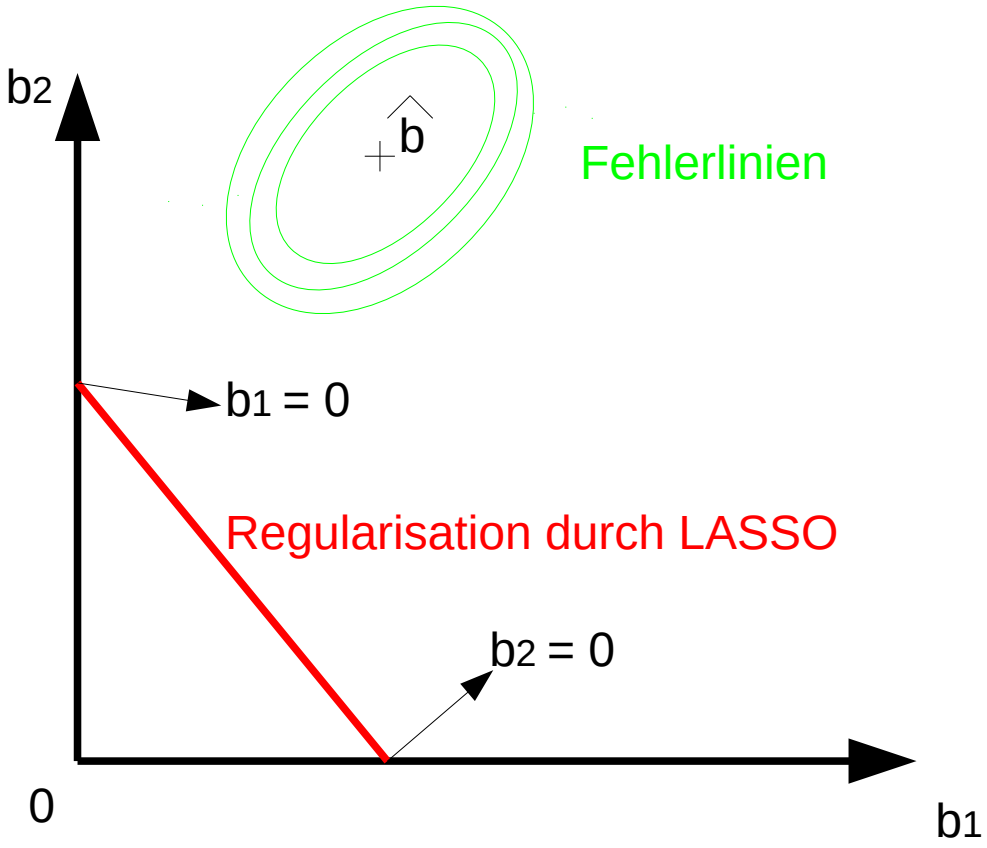
In Formeln ausgedrückt, lauten die geschätzten Regressionskoeffizienten für LASSO, wie folgt:

$$\begin{aligned} \hat{\beta}_{LASSO} &= argmin_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \\ &= argmin_{\beta} \left\{ RSS + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \end{aligned} \tag{4.6}$$

4.4.2 Subset Selection bei LASSO

Wie schon im vorangegangenen Abschnitt beschrieben, dient der Strafterm $\lambda \sum_{j=1}^{p} |\beta_j|$ zur Regularisierung der geschätzten Koeffizienten β_j im linearen Modell. Der Strafterm spielt auch eine entscheidene Rolle bei der Subset Selection. Dadurch, dass der Strafterm die Absolutbeträge der Koeffizienten β_j summiert, werden die Schätzungen von gewissen Koeffizienten explizit auf Null gesetzt. Weshalb dieser Effekt der Subset Selection bei LASSO eintritt kann mit folgender Abbildung (siehe nächste Seite) erklärt werden.

In dieser Abbildung sind nur zwei erklärende Variablen gezeigt und somit ist $p = 2$. Die Koeffizienten zu den erklärenden Variablen werden in der Abbildung mit b und nicht mit β bezeichnet. Unter der Annahme, dass wir unendlich viele Daten hätten, wäre der Schätzer der Koeffizienten b_j mit minimalem Fehler am Punkt, welcher in der Abbildung mit $\hat{b}$ bezeichnet ist. Die grünen Ellipsen um diesen Punkt $\hat{b}$ sind die Linien mit konstantem Fehler. Die rote Linie steht für die Grenze, welche durch den Strafterm aus LASSO entsteht. Das heisst geschätzte Koeffizienten können nur links dieser roten Linie liegen. Da wir den geschätzten Koeffizienten $\hat{b}_j$ einerseits minimalen Fehler erreichen wollen und auf der anderen Seite innerhalb der Regularisierungsgrenzen sein müssen, liegen die besten Schätzer für b_j am Schnittpunkt zwischen den grünen Ellipsen und der roten Linie. Durch den Verlauf der roten Linie ist die Wahrscheinlichkeit, dass sich die grünen Ellipsen und die rote Linie auf einer Koordinatenachse schneiden sehr hoch. Schneiden sich die grünen Ellipsen und die rote Linie auf einer Koordinatenachse, dann wurde ein Schätzer für einen Koeffizienten b_j auf Null gesetzt und somit haben wir den gewünschten Effekt der Subset Selection erreicht.



4.5 Bestimmung von λ

Der Strafterm, welcher in Gleichung (4.6) eingefügt wurde und für die Regularisierung bei LASSO verantwortlich ist, enhält eine Variable λ. Diese Variable bestimmt das Ausmass der Regularisierung und muss als zusätzlicher Parameter aus den Daten bestimmt werden. Für die Bestimmung von λ wird eine sogenannte Kreuzvalidierungsprozedur (cross validation) verwendet. Bei einer Kreuzvalidiuerng werden die Beobachtungen zufällig in ein sogenanntes Trainings-Set und in ein Test-Set unterteilt, wobei das Test-Set meist weniger Beobachtungen enthält als das Trainings-Set. Mit dem Trainings-Set werden dann die Koeffizienten β_j geschätzt. Dann werden für vorher bestimmte Werte von λ die Beobachtungen im Test-Set vorhergesagt. Der Wert von λ, welcher die tiefsten Vorhersagefehler liefert, wird als optimaler Schätzwert von λ betrachtet.

4.6 Analyse mit LASSO in R

In diesem Abschnitt wird gezeigt, wie ein Datensatz mit LASSO in R analysiert werden kann. Wir verwenden dazu den `Hitters`- Datensatz aus dem Buch von James et al. (2013). Dieser Datensatz enthält als Zielgrösse das Einkommen von Baseballspielern und zu diesen Spielern noch weitere erklärende Variablen. Der Datensatz ist im R-Package `ISLR` integriert. Für die Analyse werden wir die Funktion `glmnet()` aus dem gleichnamigen R-Package verwenden. Als erstes installieren wir die beiden Packages und ignorieren alle Records, welche fehlende Daten aufweisen.

```
if (!require(ISLR)) {
  install.packages("ISLR")
```

```
  require(ISLR)
}
```

```
## Loading required package: ISLR
```

```
if (!require(glmnet)){
  install.packages("glmnet")
  require(glmnet)
}
```

```
## Loading required package: glmnet
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-5
```

```
### # records mit fehlenden Daten ignorieren
data(Hitters)
Hitters <- na.omit(Hitters)
dim(Hitters)
```

```
## [1] 263  20
```

Da wir für die Bestimmung von λ mit Kreuzvalidierung ein Trainings- und ein Test-Set benötigen, bestimmen wir diese durch den Zufallszahlengenerator und der Funktion `sample()`

```
set.seed (1)
train <- sample (c(TRUE ,FALSE), nrow(Hitters), rep=TRUE)
test  <- (! train )
```

Wir verwenden die Funktion `glmnet()` zur Modellierung mit LASSO. Für diese Funktion muss das Modell anders spezifiziert werden als für die Funktion `lm()`. Wir brauchen dazu die Objekte `x` und `y`.

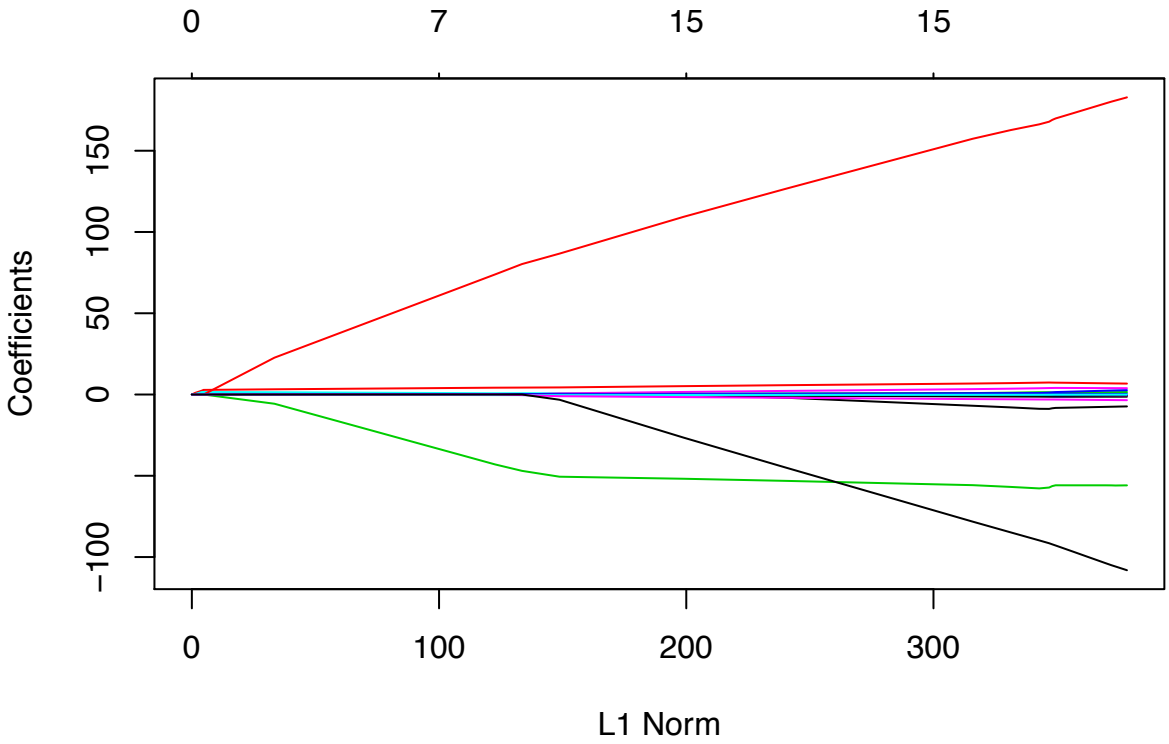
```
x <- model.matrix (Salary ~ ., Hitters)[,-1]
y <- Hitters$Salary
```

Die vorgegebenen Werte für λ werden in der Variablen `grid` abgelegt. Es handelt sich um 100 Werte zwischen 10^10 und 10^{-2}.

```
grid <- 10^ seq (10,-2, length =100)
```

The following statements fits a LASSO model.

```
lasso.mod <- glmnet (x[train ,],y[train],alpha =1, lambda = grid)
plot(lasso.mod)
```



Der Plot zeigt, wie sich der Strafterm für verschiedene Werte (durch Farben codiert) verhält. Nun wollen wir den besten Wert für λ bestimmen. Dies wird durch Kreuzvalidierung gemacht.

```
set.seed (1)
cv.out <- cv.glmnet (x[train ,],y[train],alpha =1)
bestlam <- cv.out$lambda.min
```

Der Anteil an Koeffizienten, welcher durch LASSO null gesetzt wird kann mit folgenden Statements überprüft werden.

```
out <- glmnet(x, y, alpha = 1, lambda = grid)
lasso.coef <- predict(out, type = "coefficients", s=bestlam )[1:20,]
lasso.coef
```

```
##   (Intercept)         AtBat          Hits         HmRun          Runs
##  8.898370e-01 -5.575622e-03  2.007078e+00  0.000000e+00  0.000000e+00
##           RBI         Walks         Years        CAtBat         CHits
##  0.000000e+00  2.268641e+00 -3.428874e-02  0.000000e+00  0.000000e+00
##        CHmRun         CRuns          CRBI        CWalks       LeagueN
##  8.315024e-03  2.102106e-01  4.211554e-01  0.000000e+00  1.695962e+01
##     DivisionW       PutOuts       Assists        Errors    NewLeagueN
## -1.143553e+02  2.343374e-01  0.000000e+00 -6.607899e-01  0.000000e+00
```

Abkürzungen

Abbreviation	Meaning
QTL	Quatitative Trait Loci
GWAS	Genome Wide Association Study
GLMM	Generalized Linear Mixed Models
i.i.d.	independent, identically distributed
QQ	Quantil Quantil
LASSO	Least Absolute Shrinkage And Selection Operator

Bibliography

Balding, D. J., Bishop, M., and Cannings, C., editors (2009). *Handbook of Statistical Genetics*. Wiley.

Bühlmann, P. and Mächler, M. (2014). Computational statistics. Vorlesungsunterlagen ETHZ, 2014.

Garrick, D., Taylor, J., and Fernando, R. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*, (41(1)):55.

Hofer, A. (1990). *Schätzung von Zuchtwerten feldgeprüfter Schweine mit einem Mehrmerkmals-Tiermodell*. PhD thesis, ETH Zürich.

James, G., Witten, D., Hastie, T., and Tibshirani, R., editors (2013). *An Introduction to Statistical Learning*. Springer.

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, (157):1819–1829.

von Rohr, P. (2016). Züchtungslehre. Vorlesungsunterlagen ETHZ, HS2016.