

Fachhochschule Regensburg
Hochschule für Technik, Wirtschaft und Sozialwesen
Fachbereich Informatik / Mathematik

Diplomarbeit

Thema: **Multivariate statistische Auswertung der 18 jährigen Versuchsreihe zur Eignung von Sorten Deutschen Weidelgrases (*Lolium perenne* L.) für weidelgrasunsichere Lagen (Ausdauerprüfung in Bayern) der LBP**

Verfasser: Irmgard Hollweck
Studiengang Mathematik, Semester MA8 VB W
Matrikelnummer 1013513

Betreuer: Landesanstalt für Bodenkultur und Pflanzenbau (LBP):
Dr. S. Hartmann
Fachhochschule Regensburg:
Prof. Dr. P. Wirtz

Abgabetermin: 26.05.2000

Inhaltsverzeichnis

Einleitung	4
Problemstellung	6
1 Datenerfassung, Datenbeschreibung	8
1.1 Versuchsaufbau.....	8
1.2 Merkmale.....	12
1.2.1 Beschreibung	12
1.2.2 Zusammensetzung in die Rechnung eingehender Noten.....	15
1.3 Datensatz	23
2 Statistische Methoden	25
2.1 Hauptkomponentenanalyse	26
2.1.1 Grundprinzip.....	26
2.1.2 Rechnerische Vorgehensweise	28
2.2 Multiple Lineare Regression	29
2.2.1 Grundprinzip.....	29
2.2.2 Gütekriterien.....	30
2.2.3 Vorgehensweise.....	34
2.3 Logitanalyse	36
2.3.1 Grundprinzip.....	36
2.3.2 Gütekriterien.....	39
2.3.3 Vorgehensweise.....	42
3 Praktische Ausführung	44
3.1 Ausgangsmaterial	44
3.2 Problemstellung.....	45
3.3 Multiple Regression	47

3.3.1	Regression mit Mittelung über alle Orte und nachfolgender Bestandserzeugung	50
3.3.2	Regression mit Mittelung über alle Orte und nachfolgender Bestandserzeugung und Zusammenfassung nach Zeitpunkten..	54
3.3.3	Zusammenfassung.....	58
3.4	Logitanalyse	59
3.4.1	Untersuchung von Nutzungsjahr, Winter, Fusarium-Resistenz (nur Haupteffekte)	61
3.4.2	Untersuchung von Nutzungsjahr, Winter, Fusarium-Resistenz (mit Interaktionen)	64
3.4.3	Zusammenfassende Tendenz.....	66
4	Resümee / Ausblick	67
Anhang A: SAS-Programme / SAS-Ergebnisse		69
A 1	Einführung	69
A 2	Programme.....	71
A 3	Ausdrucke	76
Anhang B: Mathematische Grundlagen		96
B 1	Definitionen / Sätze	96
B 2	GLM (Logitanalyse)	100
B 2.1	Charakterisierung	100
B 2.2	Parameterschätzung: Maximum-Likelihood-Schätzung.....	101
Anhang C: Bezeichnungen Notationen		104
C 1	Notationen.....	104
C 2	Bezeichnungen	105
Literaturverzeichnis		106

Einleitung

„Ein fachkundiger Problemlöser muss zwei unvereinbare Qualitäten besitzen – rastlose Phantasie und geduldige Beharrlichkeit.“ (Howard W. Eves)

Beide Eigenschaften des Zitates bewogen mich dazu, mich für das Thema „Multivariate statistische Auswertung der 18 jährigen Versuchsreihe zur Eignung von Sorten Deutschen Weidelgrases (*Lolium perenne* L.) für weidelgrasunsichere Lagen (Ausdauerprüfung in Bayern) der LBP“ als Diplomarbeit zu entscheiden. Letztgenannte Beharrlichkeit war für die Wahl der Statistik, erstgenannte Phantasie für den landwirtschaftlichen Bezug verantwortlich.

Den Einstieg in die Materie erleichtert folgende kurze Einführung:

Das Deutsche Weidelgras oder Englische Raygras (*Lolium perenne* L.) gehört wegen seiner vielseitigen Nutzbarkeit und hohen Ertragsfähigkeit zu den wichtigsten und wertvollsten Futter- und Begrünungsgräsern der gemäßigten Zonen überhaupt.

Es handelt sich um ein Untergras mit niedrig bis aufrecht wachsenden, dichten Horsten. Der Fruchtstand ist eine locker besetzte, zweizeilige Ähre mit mehrblütigen Ährchen, die der Ährenachse mit der schmalen Seite anliegen. Die Deckspelze ist nicht begrannt, ein wichtiges Unterscheidungsmerkmal zum Einjährigen und Welschen Weidelgras.

Das Deutsche Weidelgras gehört zu den mehrjährigen bis ausdauernden Gräserarten. Es zeichnet sich durch dichte Narben, gute Trittfestigkeit, gutes Nachwuchsvermögen und qualitativ hohen Futterwert (Note 8) aus.

Weltweit zählt es zu den am intensivsten züchterisch bearbeiteten Futtergräsern, was eine große Sortenvielfalt schuf. So werden jährlich ca. 30 neue Stämme beim Bundessortenamt zur Wertprüfung angemeldet. Die deutsche Sortenliste umfasst zur Zeit 203 Arten, aufgeteilt in 101 Rasen- und 102 Futtergräser. Bei dem für Grassamen typischen Weltmarkt und der im Vergleich zu Getreide sehr geringen Zahl in diesem Bereich tätiger Züchter in Deutschland,

wird deutlich, dass nicht alle Sorten an die besonderen klimatischen Eigenschaften und Böden Bayerns gleich gut angepasst sein können.

Die Bayerische Landesanstalt für Bodenkultur und Pflanzenbau (LBP) führt schon seit 18 Jahren Sortenversuche bei Deutschem Weidelgras an Standorten durch, an denen erfahrungsgemäß nach dem Winter größere Schäden zu erwarten sind.

Von besonderer Bedeutung ist die Standorteignung des Saatgutes (Ansaat- und Nachsaatmischungen sowie der Mischungsbestandteile). Gerade bei Saatgut für Dauergrünland, dessen positive oder negative Auswirkungen über Jahre wirken. Äußerst wichtige Teilaspekte dieser Standorteignung sind für das bayerische Dauergrünland jedoch Winterfestigkeit und Ausdauer der mehrjährigen Gräserarten unter hiesigen regionalen Bedingungen.

Das Thema dieser Diplomarbeit: „Multivariate statistische Auswertung der 18 jährigen Versuchsreihe zur Eignung von Sorten Deutschen Weidelgrases (*Lolium perenne* L.) für weidelgrasunsichere Lagen (Ausdauerprüfung in Bayern) der LBP“ umfasst die Wahl, Erstellung und Überprüfung einer adäquaten Schätzmethode für die Zielgröße Ausdauernde aus den vorhandenen Daten der LBP.

Die Programme wurden mit SAS 6.12 (Statistical Analysis System) ausgeführt, die Diplomarbeit mit Word 2000 geschrieben.

Ein herzlicher Dank gebührt noch all jenen, welche mich während dieses Mathematikstudiums begleiteten und mich sowohl tatkräftig als auch moralisch darin unterstützten. Es fühle sich jeder angesprochen, der irgendwelche mathematischen Begriffe während dieser Zeit von mir zu hören oder zu lesen bekam.

Problemstellung

Wie bereits in der Einleitung deutlich wurde, hat das Deutsche Weidelgras eine herausragende Stellung unter den Gräsern.

Da Bayern verglichen mit Restdeutschland besondere klimatische Bedingungen besitzt z. B. teilweise Schneelagen bis zu vier Monaten oder Spätfröste bis Ende April; ist das Deutsche Weidelgras stärker durch Auswinterung, sei es durch Krankheitsbefall oder Kälteeinwirkung gefährdet als z. B. in Norddeutschland, dem angrenzenden Frankreich oder Niederlande.

Speziell Fusariumbefall (*Fusarium nivale*), eine Pilzkrankheit, welche durch lange Schneebedeckung bei relativ milder Witterung gefördert wird, schädigt Deutsches Weidelgras.

Solche Ausfälle bedeuten immer wirtschaftliche Nachteile in der Futtergewinnung und Bestandverschlechterungen, die im Extremfall sogar die Neuansaat von Dauergrünland notwendig machen.

Wegen den eklatant vorhandenen Sortenunterschieden in der Standortanpassung prüft die Landesanstalt für Bodenkultur und Pflanzenbau die Ausdauer der verschiedenen Sorten an Orten, an denen innerhalb möglichst kurzer Zeit ein aussagekräftiges Urteil möglich ist.

Am Ende einer Versuchsreihe erhält jede geprüfte Sorte eine Endbeurteilung, die „Beurteilung der Anbaueignung von Sorten des Deutschen Weidelgrases in Grenzlagen“ (Ausdauerbeurteilung). Sie gibt Auskunft über das Ausdauervermögen unter den spezifischen Gegebenheiten Bayerns. Die schlechteste Beurteilung wird mit der Note eins bewertet, die beste mit neun.

Grundlage für die Bayerische Ausdauer bildet eine regelmäßige Beobachtung der Versuchspartellen über die Versuchsdauer, mit gleichzeitiger Bewertung der Vegetation des Deutschen Weidelgrases. Dazu vergibt man sogenannte Boniturnoten zu verschiedenen Merkmalen, welche den Zustand der Parzelle klassifizieren.

Die Aufgabe besteht darin, diese Daten zur Ausdauerprüfung des Deutschen Weidelgrases in Bayern im Hinblick auf die Bayerische Winterhärte multivariat auszuwerten, um anhand dieser Beurteilung, falls notwendig, pflanzenbauliche Schlussfolgerungen ziehen zu können.

Alle erhobenen Merkmale sind ordinale Daten (Boniturnoten von eins bis neun) bzw. gar nominal (Nutzungsjahr). Außerdem gibt es mehrere Einflussfaktoren zur Ausdauer. Beides hat man bei der Wahl der Methoden zu berücksichtigen.

Einen weiteren Einblick in die Ausgangssituation erläutert das erste Kapitel.

1 Datenerfassung, Datenbeschreibung

1.1 Versuchsaufbau

Die Prüfung der Ausdauer des Deutschen Weidelgrases (WD) wird seit 1981 durchgeführt. Seitdem liefen 7 Versuchsreihen mit unterschiedlichen Sorten, an verschiedenen Orten und mit variabler Dauer.

Tabelle 1: *Versuchsreihen*

Versuch400	angesät 1981	Sortenanzahl: 30		Nutzungsjahre: 4 - 5	
Orte:	Strassmoos			Samerberg	
Versuch401	angesät 1985	Sortenanzahl: 44		Nutzungsjahre: 3 - 5	
Orte:	Baernzell	Plenting	Unterthingau	Buchau	
Versuch402	angesät 1988	Sortenanzahl: 23		Nutzungsjahre: 2 – 3	
Orte:	Osterseeon	Rettenbach		Riedern	
Versuch403	angesät 1990	Sortenanzahl: 15		Nutzungsjahre: 1 - 4	
Orte:	Osterseeon	Rettenbach	Auerberg		
Versuch404	angesät 1992	Sortenanzahl: 18		Nutzungsjahre: 4	
Orte:	Osterseeon	Hoetzelsdorf	Auerberg	Buchau	Riedern
Versuch405	angesät 1994	Sortenanzahl: 11		Nutzungsjahre: 2 –3	
Orte:	Osterseeon	Hoetzelsdorf	Auerberg	Buchau	Riedern
Versuch496	angesät 1996	Sortenanzahl: 23		Nutzungsjahre: 3	
Orte:	Osterseeon	Hoetzelsdorf	Auerberg	Buchau	Riedern

Es wurden im Laufe der Jahre Standorte aus verschiedenen Gründen aufgegeben. Bis man wieder vergleichbare Orte (ähnliche Lage, Boden, usw.) dafür fand, entstand teilweise eine reduzierte Datenerhebung. So wurde Samerberg durch Buchau; Unterthingau durch Auerberg und Plenting, Baernzell, Rettenbach durch Hoetzelsdorf ersetzt.

In der Einführung zum Deutschen Weidelgras des Versuchsergebnisheftes 1998 der LPZ wird auch der Versuchsaufbau einer Versuchsreihe erklärt. Dort heißt es:

„Zur Überprüfung der Anbaueignung von Sorten des Deutschen Weidelgrases in auswinterungsgefährdeten Lagen werden auf typischen Grünlandstandorten (mit Auswinterungsneigung) seit dem Jahr 1981 an verschiedenen Standorten in Bayern Beobachtungspartellen angelegt. Bis 1995 besaß ein Versuchsglied dieser Versuchsreihe eine Wiederholung, ab 1996 zwei. Die Parzellengröße

beträgt etwa 15 m². Eine Ertragsermittlung findet nicht statt. Die Standorte wurden so ausgewählt, dass ein möglichst breiter Umweltbereich in den Gebieten mit regelmäßigen Auswinterungsschäden abgetestet werden kann. Die Versuchsstandorte lassen im Merkmal „Ausdauer in Grenzlagen“ eine Differenzierung erwarten, da dort schon in „durchschnittlichen“ Wintern Schäden am Deutschen Weidelgras auftraten. Unter Berücksichtigung der Standortvoraussetzungen wurden Nutzung und Düngung für die Grasart WD weitgehend optimiert:

- Angestrebte Nutzung 4 bis 5 pro Jahr
- etwa 70 kg/ha Rein-Stickstoff zu jedem Aufwuchs
- Datum der ersten Nutzung früh

Die Einstufung aller Beobachtungen erfolgt in Noten von 1 (sehr gering) bis 9 (sehr stark) und damit in der Laufrichtung entsprechend den Anweisungen des Bundessortenamtes für Wertprüfungen.“

(Bemerkung: Boniturnoten sind Noten, welche zur Beurteilung der Versuchsflächen hinsichtlich der ausschlaggebenden Merkmale¹ taxiert werden.)

Um die Vergleichbarkeit der Versuchsergebnisse über die Jahre zu ermöglichen, wurden einzelne Sorten über mehrere Versuche immer wieder angebaut. Sie geben eine Orientierungshilfe hinsichtlich der einzelnen Bonituren einer Parzelle. Solche Vergleichssorten waren:

Tabelle 2: Vergleichs-, Referenzsorten

Anbaujahre	Sorten
1981 - 1987	Cropper Donata Lilope Moretti Wendy
1981 - 1988	Gremie Limedia Barlenna Parcour
1981 - 1994	Lihersa
1981 - 1996	Vigor Citadel
1985 - 1992	Bastion
1985 - 1996	Bardonna
1988 - 1996	Fennema
1990 - 1992	Phoenix

¹ siehe auch Merkmalsbeschreibung S. 12

Schwierigkeiten:

Für eine mathematisch-statistische Auswertung benötigt man vollständige und vergleichbare Daten.

Nun wurden für das Merkmal² *Weidelgrasanteil*

bis 1993 *Verunkrautung* in %

ab 1993 *Verunkrautung* in Noten

ab 1997 *Weidelgrasanteil* in Noten

für das Merkmal *Dichtigkeit*

bis 1993 *Narbenbild* in Noten

ab 1993 *Lückigkeit* in Noten

ab 1997 *Dichtigkeit* in Noten

und für *Dichtigkeit nach Winter*

bis 1997 *Austriebsdichte* in % der Fläche“ erhoben.

Die direkte Vergleichbarkeit dieser Erhebungen in dieser Arbeit kann man in Zweifel ziehen. Allerdings resultieren aus der Umwandlung der Prozentangaben in Noten - nach den Anweisungen des Bundessortenamtes für Wertprüfungen (siehe S. 22) mit gleichzeitiger Gegenläufigkeit der Noten zum interessierenden Merkmal - der Praxis entsprechende vernünftige Bonituren. Diese sind die Grundlage dieser Diplomarbeit.

Einflussgrößen für die Bonitur sind:

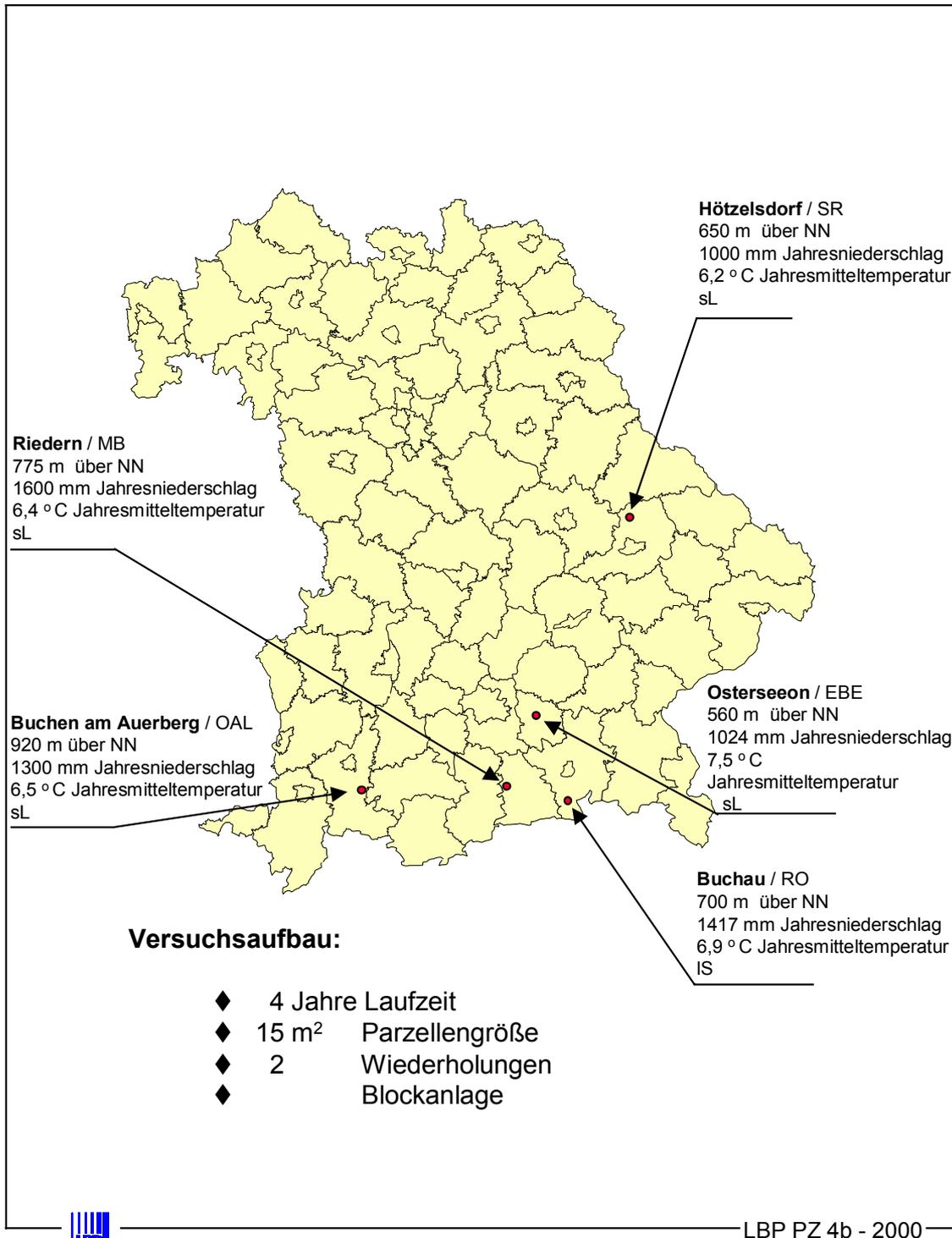
- Nutzungsjahr (= Jahr seitdem angesät wurde)
- Ort
- Versuch (unterschiedliche Sorten)
- Regenerationsfähigkeit der Sorte
- Jahr (entspricht Wetter)

Entsprechend unterschiedlich ist die Notenvergabe bei der Einzelerfassung.

² siehe auch Merkmalsbeschreibung S. 12

Eine Ortsbeschreibung der Versuchsstellen liefert die folgende Abbildung 1:

Abbildung 1: Beschreibung der Orte



1.2 Merkmale

Um sich eine Vorstellung von der Datengrundlage machen zu können, folgt auf den nächsten Seiten eine Beschreibung der verwendeten Merkmale und ein Beispieldatensatz für eine beliebige Sorte eines beliebigen Versuches. Der zweite Abschnitt beschreibt die durchgeführten Veränderungen der Originaldaten.

1.2.1 Beschreibung

Grundsätzlich entspricht die Bezeichnung *vor Winter* und *am Vegetationsende* dem Herbst, während *nach Winter* dem Frühjahr des jeweiligen Nutzungsjahres gleichkommt. *Vor* bzw. *nach Schnitt* kennzeichnet den Zeitpunkt der Boniturerhebung vor oder nach dem Mähen des Grases. Hohe Noten drücken eine positive Eigenschaft aus.

Dichtigkeit (Boniturnoten 1,...,9):

Darin werden Deckungsgrad, Lückigkeit, Größe der Horste und Vielzahl der Halme des WD an der Fläche einer Versuchsparzelle ausgedrückt.

Beispiel: Wenn ein WD-Bestand vollständig vorhanden ist, dabei aber eine sehr schütterere Pflanzendichte besitzt, so entspricht dies einer mittelmäßigen *Dichtigkeit* (Boniturnote 5).

Weidelgrasanteil (Boniturnoten 1,...,9)

Beschreibt den Gewichtsanteil des Deutschen Weidelgrases am Aufwuchs einer Versuchsparzelle hinsichtlich der Pflanzenzusammensetzung, z. B. Note 1 bedeutet: Nur Fremdgräser und Kräuter sind vorhanden.

Bestand (Noten 1,...,9)

Zusammenfassung von *Dichtigkeit* und *Weidelgrasanteil*.

Nutzungsjahr (1,...,4)

Das Ansaatjahr des WD wird als nulltes Nutzungsjahr angesehen. Versuche an unterschiedlichen Orten während einer Versuchsreihe können am Ende auch eine unterschiedliche Anzahl der Nutzungsjahre aufweisen. Es hängt vom Gesamtzustand des Versuches ab, wann dieser jeweils abgebrochen wird.

Winter (Einteilung 1,..,3)

Mit diesem Merkmal werden die Wetterverhältnisse in der vegetationsarmen Zeit vom 01.11. - 10.04. ausgedrückt

Krankheitsresistenz (Noten 1,..,9)

Krankheitsresistenzen werden bei Auftreten von *Fusarium* (eine Pilzerkrankung während des Winters), und *Rost* (eine Blatterkrankung im Sommer) erfasst.

Stand (Noten 1,..,9)

Der *Stand vor bzw. nach Winter* beschreibt den Gesamteindruck der Versuchsparzelle. Das Gegenteil davon wird als Mangel bezeichnet.

Lückigkeit

Entspricht dem Ausmaß der Fehlstellen einer Versuchsparzelle. Geht als *Dichtigkeit* in die Berechnungen mit ein.

Regenerationsvermögen

Das Regenerationsvermögen bedeutet die Fähigkeit der Pflanze sich während der Vegetationszeit wieder zu erholen. Das Merkmal wird nicht extra erhoben. Es wird indirekt durch die Schnitte erfasst bzw. es kann auch mit der Differenz vom *Bestand vor Winter* (des aktuellen Jahres) zum *Bestand nach Winter* (des aktuellen Jahres) ausgedrückt werden.

Tabelle 3 : Beispieldatensatz für die Sorte xyz vom Versuch 4711

ORT	vor Winter			nach Winter			Schnitt 1			Schnitt 2			Schnitt 3			Schnitt 4										
	NU	fr	rr	dw	svw	ww	dnw	snw	wnw	dvs	dns	wvs	wns	dvs	dns	wvs	wns	dvs	dns	wvs	wns	dvs	dns	wvs	wns	
A	0	8		8																						
A	1	6		8			9			8	8			8	8							9	8			
A	2	7		4		5	4			5	4			7	6							8				
A	3			6	4		5			5	6	4		5	6							6		7		
B	0	8					8	5		7																
B	1				5		7	8		8	8			8	7										8	8
B	2					1																				
H	0	9		9			9	5		8				9												
H	1	9					8	9		9	8			9	8							4				
H	2			9			1	1		1				6	3							4				
H	3			4		4	7	5	6		7	5														
O	0	9		9																						
O	1	7		9			9	9		9	9			9	9							9	9		7	
O	2	8		7			8	8		8	9			9	9							9	9		6	
O	3	7		7			8	9	9	9	9			9	9							9	9		9	
R	0	7							5	8																
R	1	9			7		7	7		7	8			8	8										9	8
R	2	5		7			2	2		5	7			8	7							8	7		8	8
R	3			4			4	3	5	6	6	7		6	7							7		6	7	7

Die Abkürzungen bedeuten :Nu: Nutzungsjahr, fr: Fusariumresistenz, rr: Rostresistenz, ff: Frostresistenz, Anfangsbuchstaben d: Dichtigkeit, s: Stand, w: Weidelgrasanteil
 Nachfolgende Buchstaben: vw: vor Winter, nw: nach Winter, vs: vor Schnitt n, ns: nach Schnitt n, ve: bei Vegetationsende. Für die Orte: A: Auerberg,
 B: Buchau, H: Hoetzelsdorf, O: Osterseen, R: Riedern

Wie man sieht sind die Boniturnoten teilweise sehr unterschiedlich. Auch fehlen manche Bonituren oder sollen ähnliches aussagen (wie dvs,dns bei Osterseen im 3 Nutzungsjahr). Da im 0-ten Jahr nur angesät wurde, sind außer vor Winter keinerlei Bonituren erhoben. Das Krankheitsvorkommen war sehr differenziert, so kam Rost offensichtlich nur in einem Ort in einem Jahr vor. Der Winter vom 1-ten auf das 2-te Nutzungsjahr war sehr stark, so dass der Versuch in Buchau nach dem 2-ten Jahr abgebrochen wurde.

1.2.2 Zusammensetzung in die Rechnung eingehender Noten

An dieser Stelle wird auf die Charakterisierung des Winters und die mögliche Zusammenfassung von Merkmalen aufgrund des Zeitpunktes und des Boniturwechsels / Umrechnung der Originalbonituren eingegangen.

Winter

Die Klimadaten wurden größtenteils vom Deutschen Wetterdienst München von den zu den Versuchsstandorten nächstgelegenen Wetterstationen erfasst. Diese waren:

- Bernbeuren-Prachtsried (950 Meereshöhe) für Buchen am Auerberg,
- Aschau-Sachrang (740 M) für Buchau,
- Wiesenfelden-Kesselboden (605 M) für Hötzelndorf.

Von 1991 bis 1997 wurden die Temperaturdaten des Agrarmeteorologischen Messnetzes hinzugezogen.

Ausgang für den Zeitraum des Winters ist die vegetationsarme Zeit. Da an den Versuchsorten die Vegetation um den 05.04. beginnt, wurde der 10.04. als Winterende festgesetzt.

Mit der Durchschnittstemperatur, der Anzahl der Frosttage und der Schneedeckung erfolgt die Einteilung des Winters über die Jahre hinweg in mild, normal und hart.

Tabelle 4: Einteilung der Hauptmerkmale zur Beurteilung des Winters

EINTEILUNG	Temperatur (°C)		Frosttage (Anzahl)		Schneetage (Anzahl)	
	von	bis	von	bis	von	bis
mild	...	> -2,52	...	< 321	...	< 271
normal	-2,52	-3,89	321	368	271	351
hart	< -3,89	...	> 368	...	> 351	...

Diese Einordnung der kalten Jahreszeit sollte bezüglich der Reaktionen der Pflanze erfolgen, deshalb wurden Orte mit allgemein hohem Krankheitsdruck unberücksichtigt gelassen.

Übrig blieben Hötzelndorf, Auerberg und Buchau.

Des Weiteren ist die Kenntnis über späte Fröste/Schnee bei der Beurteilung des Winters nützlich.

Fasst man alle Daten in einer Tabelle zusammen und ordnet sie nach ihren Rang, so ergibt sich die Beurteilung (Tab. 5) der einzelnen Jahre / Merkmal (beim Winter wird nur das Jahr des Beginns genannt, d. h. 1982 entspricht dem Winter 82/83.) Die darin eingegangenen Werte sind in der Tabelle 6, S. 18 aufgeführt; woraus die Einteilung des Winters folgt:

mild :	1982 1987 1988 1989 1992 1993
normal:	1984 1985 1986 1990 1994 1996 1997 1998
hart:	1981 1983 1991 1995

Begriffserklärung zu den nachfolgenden Tabellen:

Alle Werte wurden aus den 3 Orten zusammen ermittelt.

Temperatur: arithmetisches Mittel des Tagesminimums der Lufttemperatur in 20 cm Höhe.
01.11. – 10.04

Frosttage: Anzahl der Tage, in denen das Tagesminimum der Lufttemperatur in 20 cm Höhe $< 0\text{ °C}$ beträgt
01.11. – 10.04

Schneetage: Anzahl der Tage, in denen eine Gesamtschneehöhe $\geq 0\text{ cm}$ beträgt.
01.11. – 10.04

Frost_april_t: Summe der Frosttage $< 0\text{ °C}$
01.04. – 30.04

Schnee_mai_t: Summe der Tage mit Gesamtschneehöhe $\geq 0\text{ cm}$
01.05.– 31.05.

Tabelle 5: Einteilung Winter

endgültige Beurteilung WINTER	Hauptmerkmale				Hilfsgröße Wetter								
	Temperatur	Frosttage	Schneetage	Schnee	Frost	april t	Schnee	april t	Frost	mai t	Schnee	mai t	
mild													
1982	mild	mild	mild	normal	mild	normal	normal	normal	mild	mild	keine	keine	
1987	mild	mild	normal	normal	normal	normal	normal	normal	normal	normal	keine	keine	
1988	mild	mild	mild	mild	mild	mild	mild	normal	normal	normal	keine	keine	
1989	mild	mild	mild	mild	mild	mild	normal	normal	mild	mild	keine	keine	
1993	mild	mild	normal	normal	normal	normal	hart	hart	hart	hart	keine	keine	
normal													
1984	hart	normal	normal	normal	mild	normal	normal	normal	normal	normal	hart	hart	
1985	normal	normal	hart	hart	normal	normal	normal	normal	keine	keine	keine	keine	
1986	hart	normal	normal	normal	mild	normal	normal	normal	normal	normal	hart	hart	
1990	normal	mild	normal	normal	mild	normal	normal	normal	normal	normal	hart	hart	
1992	normal	normal	normal	normal	normal	normal	mild	mild	keine	keine	keine	keine	
1994	mild	mild	normal	normal	normal	normal	normal	normal	hart	hart	hart	hart	
1996	hart	normal	normal	normal	hart	normal	normal	normal	hart	hart	keine	keine	
1997	normal	normal	mild	mild	normal	mild	mild	mild	normal	normal	keine	keine	
1998	normal	hart	normal	normal	normal	normal	mild	mild	normal	normal	keine	keine	
hart													
1981	normal	normal	hart	hart	hart	hart	hart	hart	normal	normal	normal	normal	
1983	hart	hart	hart	hart	hart	hart	hart	hart	normal	normal	keine	keine	
1991	normal	hart	normal	normal	hart	normal	normal	normal	normal	normal	keine	keine	
1995	hart	hart	hart	hart	hart	hart	normal	normal	normal	normal	keine	keine	

1981 und 1991 war die lange Frostdauer im April ausschlaggebend für die Einordnung als harten Winter
1994 war die lange Frostdauer im Mai für die Einordnung als normalen Winter verantwortlich

Tabelle 6: Eingegangene Werte Winter

WINTER	Temperatur	Frosttage	Schneetage	Frost_april_t	Schnee_april_t	Frost_mai_t	Schnee_mai_t
	Mittelwert °C	N	N	N	N	N	N
1981	-3,7255	356	458	49	38	6	2
1982	-1,8263	293	262	6	15	0	
1983	-4,2918	378	373	45	28	5	
1984	-4,1973	341	300	15	14	3	6
1985	-3,8679	353	372	24	23	0	
1986	-4,7455	357	328	18	13	2	7
1987	-1,3547	296	308	25	22	2	
1988	-1,4965	306	219	7	10	5	
1989	-2,3959	305	217	14	18	1	
1990	-3,9143	104	303	19	15	6	7
1991	-3,9218	392	344	43	18	4	
1992	-2,9787	343	280	23	7	0	
1993	-2,1365	309	300	39	27	10	
1994	-1,9303	308	282	22	21	16	3
1995	-5,4519	433	369	46	22	4	
1996	-4,0099	368	312	64	19	20	
1997	-3,0444	345	220	38	6	4	
1998	-3,1418	379	346	31	11	3	

1990: nur Auerberg, nicht bei Mittelwertberechnung dabei

Einteilung des Bereichs: Normal (entspricht mittleres Drittel der Spannweite eines Merkmals)

	Bereich oberhalb der Obergrenze	Bereich unterhalb der Untergrenze
Obergrenze	-2,52	368
Mittelwert	-3,51	345
Untergrenze	-3,89	321

Beim Merkmal Temperatur hat der Bereich oberhalb der Obergrenze positive Bedeutung (= milder Winter), im Gegensatz zu den restlichen Merkmalen. Bei diesen gehört das obere Drittel zu einem harten Winter.

Zusammenfassung von Merkmalen

Wie man unter 1.2 Merkmale Tabelle 3, S. 14 sehen kann, sind in einem kurzen Zeitraum ähnliche Bonituren erhoben worden.

Folgende Merkmale drücken Verwandtes aus:

- *Weidelgrasanteil vor Winter / Weidelgrasanteil bei Vegetationsende*
- *Dichtigkeit vor Winter / Stand vor Winter / Dichtigkeit bei Vegetationsende*

- *Weidelgrasanteil vor erstem Schnitt / nach erstem Schnitt*
- *Weidelgrasanteil vor zweitem Schnitt / nach zweitem Schnitt*
- *Weidelgrasanteil vor drittem Schnitt / alle Schnitte danach*

- *Dichtigkeit vor erstem Schnitt / nach erstem Schnitt*
- *Dichtigkeit vor zweitem Schnitt / nach zweitem Schnitt*
- *Dichtigkeit vor drittem Schnitt / alle Schnitte danach*

Da bei 3 Merkmalen sowohl der Median als auch das arithmetische Mittel ordinale Daten unter- bzw. überschätzt, kann man die Streubreite berücksichtigend (je näher am Winter umso größer die Streuung) die oben genannten Merkmale mit dem arithmetischen Mittel zusammenfassen.

Eventuell kann man sich in Zukunft auch auf eine Erhebung pro Zusammenfassung beschränken.

Begründung für eine Erhebung pro Zusammenfassung:

In Tabelle 7 sind prozentuale Anteile der Abstände und deren Maxima für die in Frage kommenden Merkmale aufgeführt: Ein Abstand ist dabei immer ein Absolutbetrag.

Tabelle 7: Streubreite der Merkmale mit ähnlichem Zeitpunkt

Schnitte	Abstand der Angegebenen Merkmale					
	Dns - Dvs		Wns - Wvs		Wvw - Wve	
	kleiner 1 %	Max	kleiner 1 %	Max	kleiner 1 %	Max
1	68	7	82	4	72	3
2	76	4	91	2.5	Dvw - Dve	
3	88	2	94	2	72	4
4	93	2.5	97	2	Dve - Svw	
5	71	2	97	2	58	4

Obwohl der maximale Abstand beim Weidelgrasanteil vor / nach dem Schnitt i. d. Regel zwei Noten beträgt, liegt der Abstand über 75% der Bonituren unter einer Note. Stand vor Winter ist nur vorhanden, wenn keine Dichtigkeit vor Winter bonitiert wurde. Der Abstand der Noten zur Dichtigkeit bei Vegetationsende, der größer eins ist beträgt 42%, man sollte hier also keine Bonitur weglassen.

Nachfolgende Tabelle zeigt zwei Wachstumsverläufe in einem Jahr, welche die extremen Ausreißer etwas erklären sollen.

Tabelle 8: Boniturerhebung über ein Jahr für zwei Sorten

nach Winter		Schnitt 1				Schnitt 2				Schnitt 3				Schnitt 4					
dnw	snw	dvs	dns	wvs	wns	dve	wve												
	4.5			4.5	5	5.5	3	6	3.5	3.5	3.5	4.5	3.5	3	3.	3.5	3.5	3.5	
3		8	2	3		5	3			4	3					3			

Die unterlegten Zellen offenbaren deutlich die Schwierigkeiten einen Bestand bei den ersten zwei Schnitten richtig zu schätzen, was sich aber im Laufe der weiteren Schnitte korrigiert. Auch zeigt sich, dass die Vegetation ausreichend von der Bonitur des nächsten Schnittes erfasst wird.

Bestand

Um den Gesamtzustand einer Versuchsparzelle zu beschreiben, wurden die Merkmale *Dichtigkeit* und *Weidelgrasanteil* zum Merkmal *Bestand* folgendermaßen zusammengefasst:

Tabelle 9: Merkmal *Bestand*

		Dichtigkeit Boniturnoten								
		1	2	3	4	5	6	7	8	9
Weidel- gras- anteil Boniturno- ten	1	1	1	2	2	2	2	3	3	3
	2	1	2	2	3	3	3	4	4	4
	3	2	2	3	3	4	4	5	5	5
	4	2	3	3	4	4	5	5	6	6
	5	2	3	4	4	5	5	6	6	7
	6	2	3	4	5	5	6	6	7	7
	7	3	4	5	5	6	6	7	7	8
	8	3	4	5	6	6	7	7	8	8
	9	3	4	5	6	7	7	8	8	9

Die inneren hell- und dunkelgrau hinterlegten Zellen beschreiben die Note des Merkmals *Bestand*, welche man der Versuchsparzelle aufgrund des *Weidelgrasanteils* und der *Dichtigkeit* insgesamt geben kann. Die Einteilung geschah durch Multiplikation der Noten der beiden Merkmale mit anschließender Einteilung der Ergebnisse in 9 Abstände (Mittelwertbildung würde eine verfälschende Einteilung ergeben.)

Boniturwechsel / Umrechnung nach dem BSA

Im Laufe der Jahre gab es drei Boniturnotenwechsel. Um diese vergleichen zu können, musste man Umrechnungsarten festlegen.

Auch sollten laut Bundessortenamt (BSA) günstige Eigenschaften mit höheren Noten und ungünstige Eigenschaften mit niedrigen Noten dargestellt werden. Die Regeln zu beiden Bedingungen sind in Tabelle 10 aufgeführt.

Tabelle 10: Umrechnung der Boniturnoten

Merkmal	Erhebung	Umgerechnetes Merkmal	Umrechnung
Krankheit- / Pilzbefall	Boniturnote 1..9	Krankh. /Pilzresistenz	10 – Boniturnote
Mängel v. / n. Winter	Boniturnote 1..9	Stand v. / n. Winter	10 – Boniturnote
Narbenbild / Lückigkeit	Boniturnote 1..9	Dichtigkeit	10 – Boniturnote
Lückigkeit / Verunkrautung	Prozentualer Anteil der Fläche	Lückigkeit/ Verunkrautung als Boniturnoten	unter 1% = 1 1 - 5% = 2 6 - 10% = 3 11 - 15% = 4 16 - 20% = 5 21 - 40% = 6 41 - 60% = 7 61 - 80% = 8 über 80% = 9
Verunkrautung	Boniturnote	Weidelgrasanteil	10 – Boniturnote
Dichtigkeit / Weidelgrasanteil	Prozentualer Anteil der Fläche	Dichtigkeit / Weidelgrasanteil als Boniturnoten	unter 10% = 1 11 - 20% = 2 21 - 30% = 3 31 - 40% = 4 41 - 50% = 5 51 - 60% = 6 61 - 70% = 7 71 - 80% = 8 über 81% = 9

1.3 Datensatz

In diesem Abschnitt wird ein Überblick zu Umfang und Differenzierung der Boniturnoten gegeben. Es wurden dafür das Ansaatjahr und das vierte Erhebungsjahr ausgeschlossen. (Vom Ansaatjahr sind alleine die Bonituren *vor Winter* interessant; das vierte Jahr wird nur in Ausnahmefällen erhoben.)

Leere Zellen:

Bei den Versuchen 404, 405, 496 wurden insgesamt Bonituren an 35 Merkmalen in 5 Orten an 52 Sorten für 3 Jahren erhoben. Woraus eine Matrix von $3 \times 52 \times 5 = 780$ bzw. 769 Zeilen³ und 35 Spalten entsteht. Damit sind $769 \times 35 = 26915$ Zellen vorhanden, wovon 8787 einen Wert enthalten. Das extreme Missverhältnis relativiert sich, wenn man bedenkt, dass die 35 Merkmale oftmals zu ähnlichen Zeitpunkten Verwandtes aussagen. Es wäre wünschenswert, wenn man sie auf 7 Merkmale zusammenfassen könnte (*vor Winter, nach Winter, drei Schnitte, Fusariumresistenz, Rostresistenz*). Dann gäbe es nämlich nur noch $7 \times 769 = 5383$ Zellen, mit 4088 besetzten. Im unteren Teil der Tabelle 11 wurde dies berücksichtigt.

Tabelle 11: Anteil der besetzten Zellen pro Nutzungsjahr und Ort

Nutzungsjahr	Auerberg %	Buchau %	Hoetzelndorf %	Osterseeon %	Riedern %
gesamten Zellenanzahl : 1820 := 100%					
1	22.5	23.5	25.3	36.5	25.6
2	40.2	25.4	27.5	54.8	32.0
3	34.6	30.0 ⁴	30.6	54.7	28.2
zusammengefasste Zellenanzahl: 364 := 100%					
1	65.1	79.4	75.0	84.1	85.4
2	71.4	70.6	70.1	87.1	85.7
3	62.1	51.7 ⁵	76.4	85.7	70.1

Die gesamte Zellenanzahl berechnet sich zu 52×35 , analog die $364 = 52 \times 7$

Osterseeon ist am besten besetzt während Buchau im dritten Jahr aufgrund des schlechten Zustandes oftmals nur noch wenige Bonituren besitzt

³ beim Versuch 405 wurde Buchau nach dem zweiten Jahr abgebrochen, woraus nur 145 Zeilen statt 156 resultieren.

⁴ von 1435 = 100% berechnet

⁵ von 287 = 100% berechnet

Differenzierung der Boniturnoten:

Tabelle 12 zeigt das Verhältnis hoher zu niedrigen Noten nach Art der Merkmale. Die ersteren sind am Ende der Vegetationszeit bevorzugt vergeben worden, die letzteren im Frühjahr.

Tabelle 12: Zusammensetzung der Boniturnoten nach Merkmal

Note	Gesamt N	Vor Winter N	Nach Winter N	Resistenzen N	Schnitt 3 N
Hohe >6	4794	589	392	325	1820
Niedrige < 4	993	29	202	57	338

Gesamt beinhaltet noch den ersten und zweiten Schnitt. Der normale Bereich der Noten ergibt sich aus dem Rest⁶: $8787 - 4794 - 993 = 3000$. Da es sich hier nur um das Verhältnis der Boniturnoten handelt, wurde bei den Resistenzen auf eine Unterscheidung in *Fusarium* und *Rost* verzichtet.

Nachfolgend wurde die Notenaufteilung nach örtlichen Unterschieden aufgeschlüsselt.

Tabelle 13: Anteil der Boniturnoten am Gesamtumfang nach Ort

Nutzungsjahr	Auerberg %	Buchau %	Hoetzelsdorf %	Osterseeon %	Riedern %	
		Hohe Note >6			100% = 4794	
1	5.1	4.7	7.6	12.8	7.1	
2	0.8	2.9	7.2	17.7	5.8	
3	0.6	1.4	5.8	16.9	3.6	
		Niedrige Note < 4			100% = 993	
1	1.6	6.7	0.7	0	0.8	
2	14.8	8.1	4.2	0.4	2.2	
3	27.0	10.2	7.9	0.1	15.3	

Hier favorisieren hohe Noten das erste Nutzungsjahr und den Ort Osterseeon während der untere Bereich der Noten, im dritten Nutzungsjahr und an den drei Standorten Auerberg, Buchau, Riedern gehäuft auftritt.

⁶ siehe S. 23

2 Statistische Methoden

Zum Einstieg in die statistische Untersuchung eines Datenmaterials eignet sich die deskriptive (beschreibende) Statistik. Sie vermittelt einen raschen Überblick über Verteilungen und Zusammenhänge, mit Hilfe von Grafiken, Histogrammen, Punktwolken, Kontingenztabelle.

Bei vielen Einflussfaktoren zum Hauptuntersuchungsmerkmal wird jedoch eine gezielte Einzelbetrachtung benötigt, da die grafischen Darstellungen bezüglich ihrer Ausdrucksfähigkeit eingeschränkt sind.

Zum Verständnis des Ausgangsmaterials können weiterhin Werte wie Maximum Minimum, Mittelwert und Varianz / Standardabweichung, Pearsonscher Korrelationskoeffizient beitragen. Bei ordinalen Daten verwendet man für letztere besser Median, Quantile und den Spearman'schen Korrelationskoeffizienten .

Hat man einen Einblick in den Zusammenhang der Merkmale gewonnen, so werden nachfolgend erläuterte statistische Verfahren angewendet. Ausgang ist immer ein linearer Zusammenhang der Merkmale.

Die Beschreibung der ersten beiden Methoden stützten sich auf die Bücher BACKHAUS [3], BORTZ [4], FAHRMEIR [6], HARTUNG [7], MYERS [9]; Abschnitt 3 auf ANDREß [2], FAHRMEIR [5].

2.1 Hauptkomponentenanalyse

Bei einer Fülle von zu untersuchenden Komponenten bietet sich die Hauptkomponentenanalyse (principal components analysis) zur Datenreduktion an.

2.1.1 Grundprinzip

Das Ziel der Hauptkomponentenanalyse liegt in der möglichst umfassenden Reduktion der Datenstruktur auf möglichst wenige Faktoren, welche den größten Teil der Gesamtvarianz der beobachtbaren Variablen auf sich vereinigen sollen.

Dabei werden die beobachtbaren Variablen linear zu neuen Variablen, den Hauptkomponenten, transformiert. Diese sind untereinander unkorreliert und nach fallender Varianz geordnet. Die ersten Hauptkomponenten vereinigen demnach ein Maximum der Gesamtvarianz auf sich.

Voraussetzung für die Durchführung der Analyse ist, dass die Beobachtungsvektoren der Variablen linear unabhängig sind.

Zur Vergleichbarkeit der Variablen und zur Interpretationserleichterung geht man in der Regel von einer standardisierten Ausgangsdatenmatrix Z aus.

Diese soll in die Form:

$$Z = FL' + E$$

gebracht werden. Hierbei ist F die Matrix mit Faktorwerten (entspricht der Designmatrix bei der multiplen Regression) deren Spalten orthonormiert sind, L die Koeffizientenmatrix (Ladung) und E die Matrix der Einzelrestwerte.

Die Hauptkomponentenmethode liefert die Kleinste-Quadrat-Lösung⁷ für oben genannte Gleichung.⁸

⁷ siehe Anhang B, S. 98

⁸ Fahrmeir [6], S. 668

Ausgehend von einem orthogonalen Faktorenmodell stellt die Ladungsmatrix L die Kovarianzmatrix von Z dar⁹, und die resultierenden Faktorvariablen sind untereinander unkorreliert.

Korrelationsmatrix R

Die Eignung der Eingangsvariablen überprüft man anhand der Korrelationsmatrix R . Hierbei gibt es verschiedene Testverfahren, z. B. Signifikanzniveau, Inverse, AIC-Matrix und Kaiser-Mayer-Olkin-Kriterium¹⁰. All diese Kriterien zeigen an, in welchem Umfang die Ausgangsvariablen „zusammengehören“. Je stärker die Variablen miteinander korrelieren, desto weniger Faktoren benötigt man zur Aufklärung der Gesamtvarianz – was eine Hauptkomponentenanalyse sinnvoll erscheinen lässt.

Das letztgenannte Kriterium berechnet die sogenannte MSA (measure of sampling adequacy). Diese wird im allgemeinen als adäquate Prüfgröße angesehen (Werte unter 0.5 sind ein Anzeichen, die Variablen nicht zu berücksichtigen).

Bestimmung der Anzahl der zu verwendenden Faktoren

Die Bestimmung der zu verwendenden Faktoren liegt im subjektiven Ermessen. Üblicherweise werden die folgenden Kriterien zur Entscheidungsfindung mit herangezogen:

Scree-Test: Typischerweise zeigen Eigenwerte von R einen gleichmäßigen abfallenden Verlauf. Wird dieser unterbrochen, kann man annehmen, dass Eigenwerte nach diesem Knick von zufälliger Natur sind. So wählt man also die Anzahl der Hauptkomponenten, welche zu den Eigenwerten vor diesem Knick gehören. Bsp. siehe FAHRMEIR [6], S. 669 bzw. SCHUEMER [15], S. 278.

Kaiser-Kriterium: Berücksichtigt werden alle Hauptkomponenten mit Eigenwert größer eins. Die Eigenwerte beschreiben den Erklärungsanteil eines Faktors in Hinblick auf die Varianz aller Variablen. Man sondert also die Komponenten aus, welche weniger als eine Variable ($\sigma^2 = 1$) zur Varianzerklärung beitragen.

⁹ Fahrmeir [6], S. 643

¹⁰ Backhaus [3], S. 203-206

Mindesterklärung der Varianz: Ein willkürlich festgesetzter Anteil der Gesamtvarianz muss durch die Hauptkomponenten erklärt sein.

2.1.2 Rechnerische Vorgehensweise

Multipliziert man Z mit den normierten Eigenvektoren der Korrelationsmatrix R von Z so erhält man die Hauptkomponenten. Da diese die Ausgangsmatrix repräsentieren sollen, müssen sie noch standardisiert werden, was durch die Division ihrer Streuung geschieht (Der Mittelwert entspricht null¹¹.) Damit folgt:

$$F = YA^{-1/2},$$

mit F den Faktorwerten,
 Y den Hauptkomponenten,
 $A = \text{diag}\{\lambda_1, \dots, \lambda_p\}$ den Eigenwerten von R

Als Kriterium der Modellanpassung eignet sich sowohl die Residualkorrelationsmatrix¹² als auch die Kommunalitäten-Schätzung. Die Kommunalitäten-Schätzung stellt die Stärke der Beziehungen zwischen jeder der Variablen und den Faktoren dar. Sie wird aus dem quadrierten multiplen Korrelationskoeffizienten berechnet.

Der Zusammenhang zwischen den Ausgangsvariablen und den neuen Faktoren wird als Faktorladung bezeichnet. Sie ist mit der Korrelation der ursprünglichen Variablen zum neuen Faktor identisch. Die Faktorladung ist erklärt, durch:

$$L = TA^{1/2}$$

für T : Matrix der Eigenvektoren von R .

¹¹ Bortz [4], S. 499 bzw. S. 491

¹² Schuemer [15], S. 261

Die Faktorladung stellt eine Hilfe bei der Zuordnung der ursprünglichen Variablen zu den einzelnen Faktoren dar.

Für eine Auslegung und Benennung der neuen Faktoren ist eine Rotation (Varimax, Promax, Oblimax) nützlich. Hierbei wird versucht, Faktoren herzustellen, deren Ladungen möglichst kleine oder große Werte annehmen.

Man nimmt diejenige Methode, welche die beste Interpretation zulässt, wobei orthogonale Rotationen wegen der Beibehaltung ihrer Unabhängigkeit der Faktoren bevorzugt werden sollten. Ladungen größer 0.5 geben Hinweise zur Gruppenbildung¹³. Daneben dient die graphische Darstellung der Variablen im Raum der gemeinsamen Faktoren einer besseren Interpretation¹⁴.

2.2 Multiple Lineare Regression

Kodiert man ordinale Daten nach Fechner¹⁵, so ist eine multiple lineare Regression anwendbar. Besitzen diese Daten eine größere Bandbreite, so können sie andererseits auch als metrisch betrachtet werden, wodurch die Kodierung je nach Datenmaterial nicht mehr unbedingt notwendig ist.

Die multiple lineare Regression hat gegenüber der in Abschnitt 2.3 erklärten kategorialen Regression den Vorteil, dass sie weit geringere Anforderungen an die Vollständigkeit der Datenerhebung und den Stichprobenumfang stellt. (Fehlende Daten gehen mit Null in die Regression ein.)

2.2.1 Grundprinzip

Bei der multiplen linearen Regression wird eine lineare Beziehung zwischen einer metrischen Variablen y (Zielvariable, abhängige Variable) und p (Anzahl) weiteren Variablen x_1, \dots, x_p (unabhängige Variablen, Regressoren, Prädiktoren) angenommen.

¹³ zu Faustzahlen der Interpretation siehe Bortz [4], S. 509

¹⁴ Bsp. siehe Fahrmeir [6], S. 683

¹⁵ Hartung [7], S. 277

Bei I Wiederholungen des x_j -ten Versuchswertes lautet der Ansatz für die multiple lineare Regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, I$$

bzw. mit $x_{i0} = 1 \Rightarrow \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}) = 0$

Dabei enthalten \mathbf{y} und \mathbf{X} die beobachteten Werte und $\boldsymbol{\beta}$ ist der unbekannte Parametervektor. $\boldsymbol{\varepsilon}$ stellt den Fehlervektor, welcher um Null streuen soll, dar. Kategoriale Regressoren kodiert man mit einer Dummy-Kodierung¹⁶

Die Regressionskoeffizienten $\boldsymbol{\beta}$ werden mit der Methode der kleinsten Quadrate¹⁷ geschätzt (Voraussetzung dafür: Die Fehler sind normalverteilt und unkorreliert und \mathbf{X} hat vollen Rang) :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

Stichprobengröße

Die optimale Stichprobengröße kann man nach BORTZ [4], S. 429 berechnen.

Eine grobe Faustregel lautet:

Die Zahl der Beobachtungen sollte wenigstens doppelt so groß sein, wie die Anzahl der Variablen in der Regressionsgleichung¹⁸.

2.2.2 Gütekriterien

Es gibt eine Vielzahl von anwendbaren Gütekriterien für die multiple lineare Regression. Anschließend werden die wichtigsten davon erklärt.

¹⁶ siehe S. 38, Designmatrix

¹⁷ siehe Anhang B, S. 98

¹⁸ Backhaus [3], S. 49

Bestimmtheitsmaß

Das Bestimmtheitsmaß wird durch das Verhältnis von erklärter Streuung zur Gesamtstreuung definiert. Je mehr sich das Bestimmtheitsmaß R^2 seinem Maximum 1 nähert, desto besser wird die abhängige Variable durch die unabhängigen Variablen bestimmt.

Sei $\hat{y} = X\hat{\beta}$; \bar{y} das arithmetische Mittel der y_n , $n = 1, \dots, N$, so ist:

$$R^2 = \frac{\sum_{n=1}^N (\hat{y}_n - \bar{y})^2}{\sum_{n=1}^N (y_n - \bar{y})^2} \quad \text{bzw. mit } x_{n0} = 1 \text{ gilt:} \quad R^2 = 1 - \frac{\sum_{n=1}^N (y_n - \hat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2}$$

R^2 wird durch die Zahl der Regressoren und die Größe der Stichprobe beeinflusst. Tendenziell gilt: Je mehr Regressoren hinzukommen, und je kleiner die Stichprobengröße ist, um so größer das Bestimmtheitsmaß. Deshalb wird oftmals auch das folgende parameterbereinigte Bestimmtheitsmaß angegeben:

$$R^2 = 1 - \frac{N-i}{N-p} \cdot \frac{\sum_{n=1}^N (y_n - \hat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2}$$

wobei n Anzahl der Beobachtungen
 p Anzahl der Parameter
 i konstant 1 falls $x_{n0} = 1$ sonst konstant 0

R wird auch als empirischer Korrelationskoeffizient zwischen y und \hat{y} aufgefasst.

Signifikanztest¹⁹

- 1) Für die Regression als Ganzes:

Die Prüfung der Regression als Ganzes wird durch den F-Test durchgeführt.

$$F = \frac{R^2}{1-R^2} \cdot \frac{N-p-1}{p}$$

Die Nullhypothese besagt, dass kein Zusammenhang zwischen den abhängigen und den unabhängigen Variablen in der Grundgesamtheit besteht.

- 2) Für die j-te unabhängige Variable gilt:

Es wird der statistisch signifikante Erklärungswert für die Zielvariable mit dem F-Test überprüft:

$$F = t^2 = \frac{\hat{\beta}^2}{\hat{\sigma}^2 a_{jj}} \quad , \text{ mit } a_{jj} \text{ j-tes Diagonalelement von } (\mathbf{X}'\mathbf{X})^{-1}$$

Die Nullhypothese lautet: $\beta_j = 0$

Kann man die Nullhypothese nicht verwerfen, so kann man die Regressoren eliminieren; allerdings heißt das nur, dass sie keinen zusätzlichen Beitrag bei gleichzeitiger Berücksichtigung der verwendeten Variablen zur Erklärung der Zielvariable bringen.

Konfidenzintervalle

Konfidenzintervalle beschreiben den Schwankungsbereich einer möglichen Schätzung. Man kann sie für jeden Regressionskoeffizienten bestimmen und damit auch für die Schätzung der Zielvariable.

¹⁹ zu Voraussetzung und Verteilungsannahmen siehe Schuemer [15], S. 92 und Bortz [4], S. 417 bzw. Fahrmeir [6], S. 110

Residuenanalyse

Die einfache Abweichung der beobachteten Zielvariable von den geschätzten Werten bezeichnet man als Residuen:

$$e = y - \hat{y}$$

Um Heteroskedastizität (Streuung der Residuen ist nicht konstant) oder Autokorrelation (Residuen sind korreliert, Durbin / Watson-Test²⁰) zu entdecken, verwendet man die standardisierten Residuen:

$$\hat{e}_i = \frac{e_i}{\sigma_e}$$

mit σ_e der Standardabweichung der Residuen.

Normalerweise wird die Residuenanalyse als grafische Darstellung ausgeführt.²¹

Linearität

Falls Nichtlinearität zwischen den Beziehungen der unabhängigen Variablen und der abhängigen Variable vorliegt, können Schätzwerte verzerrt werden. Hinweise dafür können Tests auf Autokorrelation und Heteroskedastizität geben.

Multikollinearität

Darunter versteht man die wechselseitige Abhängigkeit der Variablen. Bei großen Korrelationen zwischen den Variablen werden die Schätzwerte unzuverlässig und die zugehörigen Standardabweichungen sehr hoch.

Ein Test dazu ist die Toleranz, die als $1-R^2$ für eine Variable hinsichtlich aller anderen Regressoren des Modells definiert ist. Je kleiner dieser Wert, desto kritischer das Model.

²⁰ siehe SAS User's Guide [12], S. 1434 bzw. Myers [9], S. 288

²¹ Bsp. siehe S. 78

Ebenfalls deckt eine Kollinearitätsanalyse²² die Kollinearität auf. Zunächst wird die $X'X$ -Matrix so normiert, dass die Diagonale Einsen enthält. Danach wird ein Index aus dem Verhältnis von größtem Eigenwert zu jedem einzelnen Eigenwert dieser Matrix berechnet. Der „condition indices“ von SAS zieht hieraus noch die Quadratwurzel.

Falls eine Komponente mit einem hohen Index bedeutsam zu zwei oder mehr der Prädiktorvariablen beiträgt - ausgedrückt durch einen hohen kollinearitäts-abhängigen Varianzanteil von β_i - liegt ein Kollinearitätsproblem vor²³.

2.2.3 Vorgehensweise

Die multiple lineare Regression ist in der Anwendung ein iterativer Prozess und gliedert sich in folgende Schritte:

- Bestimmung der abhängigen und unabhängigen Variablen aufgrund von sachlogischen Zusammenhängen
- Variablenselektion durch Vorwärts- bzw. Rückwärtsselektion
- Durchführung und Überprüfung der Regression anhand der Gütekriterien
- Interpretation der Regressionskoeffizienten

Bis jetzt wurde ein Hauptpunkt der multiplen linearen Regression nämlich die Überprüfung erklärt.

Was noch fehlt ist die Variablenselektion und die Interpretation. Sie folgen nun im Anschluss.

Variablenselektion

SAS erlaubt eine schrittweise Regression. Allerdings sollte man bei Anwendung darauf achten, dass sachlich wichtige Variablen nicht eliminiert werden²⁴.

²² Schuemer [15], S. 109

²³ Myers [9], S. 368 – S. 373

²⁴ Bortz [4], S. 427 unterster Abschnitt

Backward-Elimination: Man geht vom vollständigen Modell aus, in dem alle Variablen enthalten sind. Dann werden sukzessiv Variablen, aufgrund ihrer Signifikanz eliminiert. (voreingestelltes Signifikanzniveau bei SAS: 0.10)

Forward-Selektion: Entspricht der entgegengesetzten Vorgehensweise zur Backward-Selektion, d. h. die unabhängigen Variablen werden sukzessive in das Modell anhand ihrer Signifikanz (SAS: 0.5) aufgenommen.

Stepwise: Ist eine Kombination von Forward-Selektion und Backward-Elimination (Signifikanzniveau: 0.5).

Interpretation der Regressionskoeffizienten

Der einzelne Parameter β_i wird als die erwartete Veränderung pro Einheit von x_i an der Zielvariable bei konstant gehaltenen x_j (entspricht den restlichen unabhängigen Variablen) verstanden. Je größer der absolute Betrag von β_i ist, umso größer ist der vermutete Einfluss auf die abhängige Variable.

Der standardisierte Regressionskoeffizient erzielt eine direkte Vergleichbarkeit der Prädiktoren bezüglich der Zielvariablen, unabhängig von ihren möglicherweise unterschiedlichen Messdimensionen. Für die Berechnung gilt:

$$\hat{b} = \beta_i \cdot \frac{\sigma_{X_i}}{\sigma_Y}$$

Bei der Auslegung sind Suppressionseffekte²⁵, (Unterdrückung von irrelevanten Varianzen einer Prädiktorvariablen durch eine andere Variable) sowie Korrelationen untereinander zu beachten. Man kann somit die Regressionskoeffizienten nur im Zusammenhang mit allen eingehenden Variablen beurteilen.

Der Strukturkoeffizient (Einzelkorrelation dividiert durch die multiple Korrelation) drückt den Anteil einer Prädiktorvariablen an der geschätzten Kriteriumsvarianz ohne die Berücksichtigung der übrigen unabhängigen Variablen aus. Er kann eine Hilfe bei der Interpretation der einzelnen Komponenten sein.

²⁵ Bortz [4], S. 425

2.3 Logitanalyse

Speziell für diskret ordinal / nominal skalierte Daten sind Verfahren des generalisierten linearen Modells entwickelt worden.

Sie modellieren entweder die Häufigkeit oder den Anteil, mit dem bestimmte Ausprägungen oder Kombinationen von Ausprägungen der kategorialen Variablen vorkommen.

Hat man am Schluss dann die Wahrscheinlichkeit für eine bestimmte Komponente geschätzt, so kann man daraus eine Ausprägung der Zielvariablen prognostizieren²⁶.

Für geringe Zellbesetzungen und einen relativ geringen Stichprobenumfang ist die Logitanalyse allen anderen Verfahren (Grizzle-Starmer-Koch-, Probitverfahren) im Bereich der kategorialen Regression überlegen.

2.3.1 Grundprinzip

Sei für folgende Erklärungen:

- $i = 1, \dots, l$ Anzahl der Merkmalskombinationen
 $\Rightarrow x_i$ ist die i -te Merkmalskombination
- $q = R - 1$ wobei R die Kategorienanzahl der Zielvariablen darstellt
- $r = 1, \dots, q$
- p die Anzahl der Haupt- und Interaktionseffekte = Anzahl der Zufallsvektoren

Im einfachsten Fall (binäres Logitmodell) modelliert man eine nichtlineare Beziehung zwischen der dichotomen abhängigen Variablen y und der erklärenden, unabhängigen Variablen x in der Form

$$\pi_1 = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

mit $E(y_n | x_n) = \boldsymbol{\pi} = (\pi_{n1}, \dots, \pi_{nq})'$ der Auftrittswahrscheinlichkeit der Zielvariablen.

Linearisiert man dieses Modell, so kommt man zum **Binären Logitmodell**:

²⁶ Hartung [7], S. 140

$$g_r(\hat{\pi}_i) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_{ip}$$

wobei das Wahrscheinlichkeitsverhältnis der Kategorien $\pi/(1-\pi)$ als **Odds** und der Logarithmus davon als **Logit** bezeichnet wird.

Bei mehreren Ausprägungen der Zielvariable verwendet man das **Multinominale Logitmodell**:

$$g_r(\hat{\pi}_i) = \ln\left(\frac{\hat{\pi}_{ir}}{1 - \sum_{j=1}^q \hat{\pi}_{ij}}\right) = \beta_{0r} + \beta_{1r} x_{i1} + \dots + \beta_{pr} x_{ip}$$

Falls die Zielvariable ordinale Daten aufweist so kann man das **Kumulative Logitmodell (proportional-odds model)** einsetzen:

$$g_r(\hat{\pi}_i) = \ln\left(\frac{\sum_{j=1}^r \hat{\pi}_{ij}}{1 - \sum_{j=1}^r \hat{\pi}_{ij}}\right) = \mathbf{x}_i' \boldsymbol{\beta}_r$$

Bei dem kumulativen Logitmodell ist zu beachten, dass es die eingegebenen Zielvariablen als Erfassung einer metrischen Größe bewertet. Die Regressionskoeffizienten sind für jede Ausprägung der Zielvariablen gleich, was einer groben Bewertung entspricht. Den Intercept kann man zum Schwellenwert der einzelnen Kategorien umrechnen.

Die Parameterschätzung von $\boldsymbol{\beta}$ wird mit der Maximum-Likelihood-Schätzung bewerkstelligt. Sie hat den Vorteil, dass nur der Gesamtstichprobenumfang $N \rightarrow \infty$ gehen muss, damit der ML-Schätzer asymptotisch effizient ist.

Für natürliche Linkfunktionen²⁷, welche in diesem Modell Verwendung finden, ist die ML-Theorie gut entwickelt, während für andere Linkfunktionen die asymptotischen Optimalitätseigenschaften nur unter strengeren Annahmen gesichert sind.

Designmatrix²⁸

Die Koeffizienten der Schätzgleichung: $\mu = X'\beta$ werden durch die Designmatrix X beschrieben. In unserem Fall hat sie immer vollen Rang. Da kategoriale Daten einfließen ist es sinnvoll diese zu kodieren.

Die Kodierung der einzelnen Ausprägungen lassen sich für $i = 1, \dots, I-1$ wie folgt darstellen,

Dummykodierung

$$x_i^A = \begin{cases} 1 & \text{falls Kategorie } i \text{ der Variablen } A \text{ vorliegt} \\ 0 & \text{sonst} \end{cases}$$

Effekt-Kodierung

$$x_i^A = \begin{cases} 1 & \text{falls Kategorie } i \text{ der Variablen } A \text{ vorliegt} \\ -1 & \text{falls Kategorie } I \text{ der Variablen } A \text{ vorliegt} \\ 0 & \text{sonst} \end{cases}$$

Als redundante Referenzkategorie wählt man gerne die I-te Kategorie (Faktorstufen, Ausprägungen) eines Merkmals. Man könnte sich aber auch für jede andere Ausprägung entscheiden. Deshalb wird diese Kategorie dann in der Matrix nicht mehr berücksichtigt.

Die Parameter der Dummy-Kodierung führen inhaltlich zu den gleichen Schlussfolgerungen über die Beziehungen von Häufigkeiten wie die Effektkodierung. Man benötigt die Unterscheidung zur Interpretation der Effekte.

²⁷ siehe Anhang B, S. 97

²⁸ Andreß [3], S. 71 / 149 / 278

Im Gegensatz zur Effekt-Kodierung, bei der die einzelnen Effekte die Abweichung vom Gesamtdurchschnitt ausdrücken, beziehen sich die Effekte der Dummykodierung auf die durchschnittliche Ausprägung der Referenzkategorie. Im allgemeinen wird die Effektkodierung bevorzugt.

Zu den erklärenden Variablen fügt man noch einen sogenannten Intercept hinzu. Bei der klassischen Regression entspricht dies der Regressionskonstanten x_0 . In der kategorialen Regression bewirkt die Hinzunahmen des Intercepts nicht nur eine Verschiebung des Modells in Richtung der Zielvariablen und somit eine bessere Anpassung, sondern er repräsentiert den Bezugspunkt der einzelnen Effekte des Modells. Dementsprechend ist er immer vorhanden und damit konstant eins.

Beispiele für eine Designmatrix findet man auf S. 61 oder in FAHRMEIR [5], S. 226.

2.3.2 Gütekriterien

Die Gütekriterien sind ähnlich denen der multiplen linearen Regression.

Hohe Werte der Standardabweichung von $\hat{\beta}$ drücken sowohl eine Problematik mit den statistischen Tests als auch die große Unsicherheit einer Schätzung aus. Einen besseren Einblick in die Größe der Bandbreite einer Schätzung geben dann die Konfidenzintervalle von $\hat{\beta}$. Falls diese verschiedene Vorzeichen besitzen, wird eine Interpretation sehr schwierig.

Für einen Modellvergleich wird der AIC²⁹ (Akaikes Informationskriterium) gerne herangezogen. (Das Modell mit dem kleineren Wert ist zu bevorzugen).

²⁹ Andreß [2], S. 287

Wald-Test

Er dient:

- 1) Zur Überprüfung der Homogenität der unabhängigen Variablen:
 Homogenität der unabhängigen Variablen heißt, dass die Zielvariablenausprägungen für jeden Prädiktor gleichwahrscheinlich sind, z. B. bei der Auswahl von Farben (Zielvariable) nach Geschlecht.
 Sie wird durch den Wald-Test (gewichtete Quadratsumme der Residuen zwischen vollem und restringiertem Modell) oder dem Score-Test³⁰ getestet.
 Die Nullhypothese lautet: Das restringierte Modell trifft zu bzw. die Prädiktoren sind homogen.
 Im Gegensatz zu den meisten Tests erwartet man hier eine Annahme der Nullhypothese.

- 2) Zur Überprüfung der Regressionskoeffizienten auf Signifikanz:
 Die Prüfung der Regressionskoeffizienten wird durch den Chi-Quadrat-Test durchgeführt,

$$\chi^2 = \left(\frac{\hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2$$

mit $\sigma_{\hat{\beta}}$ dem Standardfehler des geschätzten Regressionskoeffizienten.

Likelihood-Ratio-Test

Hier wird die geschätzte Likelihood-Funktion $L(\hat{\beta})$ mit der größtmöglichen Likelihood-Funktion $L(\hat{P})$ in Beziehung gesetzt. Durch die Verwendung der log-Likelihood-Funktion wird sie dann Chi-quadrat-verteilt.

$$-2 \ln \lambda(\beta) = -2 [l(\hat{\beta}) - l(\hat{P})]$$

³⁰ Fahrmeir [6], S. 88

Die Nullhypothese lautet: Alle Regressionskoeffizienten sind null.

Falls man für das größtmögliche Modell, das saturierte Modell (enthält alle möglichen Kombinationen und Interaktionen) nimmt, so nennt man diesen Wert auch Devianz.

Für kleine Häufigkeiten, insbesondere $N_i=1$ sind die Tests unzuverlässig.

Pseudo- R^2

Pseudo- R^2 misst analog zum Bestimmtheitsmaß die Erklärungskraft des Modells. Der maximale erreichbare Wert ist hierbei eins. Trifft dieser Wert zu, so stellt dies einen Hinweis auf Separabilität (bedeutet Gruppenbildungen der unabhängigen Variablen) dar. Werte über 0.2 deuten auf einen starken Zusammenhang zwischen den Prädiktoren und der Zielvariablen hin.

Die Berechnung erfolgt nach:

$$R^2 = 1 - \left[\frac{L(0)}{L(\hat{\beta})} \right]^{\frac{2}{n}}$$

mit $L(0)$ der Likelihood-Funktionen der Regressionskonstanten

$L(\hat{\beta})$ der Likelihood-Funktion des geschätzten Modells

n der Stichprobenumfang.

Wie bei der multiplen Regression gibt es auch ein „adjusted“ R^2 , welches die gleiche Bedeutung hat, wie in Abschnitt 2.2.2, S. 31.

Residuenanalyse

Da die normalen Residuen je nach Verteilungsfunktion der Zielvariable sehr groß werden können³¹, werden für die Kategoriale Regression die Pearson-Residuen bevorzugt angegeben. Falls sie einen Betrag größer zwei besitzen, ist es ein Hinweis auf einen Anpassungsmangel. Sie werden folgendermaßen berechnet:

³¹ Agresti [1], S. 90

$$e_i = \frac{y_i - n_i \hat{\pi}_i}{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}$$

Eine Alternative dazu sind die Devianz-Residuen³². Sie werden bei besonders schiefverteilten Residuen angewandt. Grundsätzlich gilt für die Anwendung, dass die lokalen Stichprobengrößen größer eins sein sollten.

2.3.3 Vorgehensweise

- Bestimmung der abhängigen und unabhängigen Variablen aufgrund von sachlogischen Zusammenhängen
- Merkmalszusammenfassungen, damit eine ausreichende Zellbesetzung erfolgt
- Durchführung und Überprüfung der Analyse
- Interpretation der Regression

Durchführung

- Man hat eine Verteilungsfunktion für die Zielvariable und eine Linkfunktion für \hat{y} zu bestimmen.
- Bei der Logitanalyse ist die Verteilungsfunktion für die Zielvariable multinomial bzw. binomial verteilt und die Linkfunktion ist das Logitmodell.
- Nach Kodierung der Rohdaten (Erstellung der Designmatrix, erstellt üblicherweise die Software automatisch), führt man eine Maximum-Likelihood-Schätzung durch (näheres siehe Anhang B 2.2).
Sie wird iterativ bestimmt. Als Abbruchkriterium für die Iteration sind bei SAS die Schrittweite als auch die maximalen Iterationsschritte einstellbar. Standardmäßig gilt Schrittweite 0.5 und max. Iterationsschritte = 20.
- Die ermittelten Parameter hat man anschließend in Wahrscheinlichkeiten umzurechnen:

³² Formel siehe Fahrmeir [6], S. 280

binäre Logitmodel:

$$\pi_i = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})}$$

multinomiale Logitmodel:

$$\pi_{ir} = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta}_r)}{1 + \sum_{j=1}^q \exp(\mathbf{x}'_i \boldsymbol{\beta}_j)}$$

kumulative Logitmodel:

$$P(y \leq r | \mathbf{z}_i) = \sum_{j=1}^r \pi_{ij} = \frac{\exp(\mathbf{z}'_i \boldsymbol{\beta}_r)}{1 + \exp(\mathbf{z}'_i \boldsymbol{\beta}_r)}$$

mit der modifizierten Designmatrix \mathbf{z}_i .³³

Interpretation des Modells

Die Regressionskoeffizienten kann man ähnlich wie bei der Multiplen Regression S. 35 in Abschnitt 2.2.3 auslegen. Nur dass es sich hier beim Anstieg der Zielvariable um die Logits handelt. Auch muss man sich überlegen, welche Ausprägungen der unabhängigen Variablen man in das Modell einfließen lassen will (Skalierung, Kodierung).

Der Exponent des Regressionskoeffizienten e^{β} gibt den Faktor an um den sich die Odds (Verhältnis von einer Wahrscheinlichkeit zur Gegenwahrscheinlichkeit der Zielvariable: $\pi/(1-\pi)$) bei einem Anstieg der erklärenden Variablen um eine Einheit ändern, oder anders ausgedrückt, er ist das Odds-Ratio (das Verhältnis zweier Odds) von einer Gruppe (bei binären Daten: z. B. $x_j=1$) zu einer Vergleichsgruppe (bei binären Daten: z. B. $x_j=0$).

Bezogen auf die Odds handelt es sich um ein multiplikatives Modell. Dies ist bei der Interpretation zu beachten (1 bedeutet keinen Effekt).

³³ siehe Fahrmeir [5], S. 222

3 Praktische Ausführung

3.1 Ausgangsmaterial

Im Folgenden wird die „Beurteilung der Anbaueignung von Sorten des WD in Grenzlagen“ umgangssprachlich mit „Bayerischer Winterhärte / Winterhärtenote“ bezeichnet.

Es stehen 7 Versuchsreihen mit insgesamt 128 Sorten an 6 verschiedenen Orten zur Verfügung (Nähere Erläuterung siehe Kapitel 1). Aufgrund einiger Schwierigkeiten in der Datenaufbereitung beschränke ich mich in diesen Ausführungen auf die letzten drei Versuchsreihen: 404, 405, 496. Sie umfassen 1101 Zeilen; das heißt, Beobachtungen mit unterschiedlichen Merkmalen und 52 Sorten, dies entspricht der gleichen Zahl von Bayerischen Winterhärtenoten. Auf einen Beispieldatensatz S. 14 sei hingewiesen.

Um die Ergebnisse der nachfolgenden Untersuchungen richtig verstehen zu können, ist es sinnvoll, die bisherige Indexberechnung der Bayerischen Winterhärte (Baywihae) zu kennen:

$$Index^* = \frac{\sum (\text{arithmetisches Mittel aller Merkmale über alle Orte / Nutzungsjahr})}{\text{Anzahl der Nutzungsjahre}}$$

Für ein Nutzungsjahr³⁴ flossen demnach, bei einer vollständigen Erhebung und Teilnahme von 5 Orten z. B.

- 2 x 5 Noten nach Winter
- 4 x 2 x 5 Noten für die Schnitte
- 2 x 5 Noten vor Winter
- 2 x 5 Noten für Krankheitsresistenzen

in die Berechnung mit ein – woraus bereits die Fülle der Einflussfaktoren ersichtlich wird (es gingen bisher immer ca. 150 Einzelwerte in die endgültige Berechnung der Bayerischen Winterhärtenoten mit ein).

³⁴ zur Erklärung des Begriffs Nutzungsjahr siehe S. 12

Auch ist der Anteil der Schnitte am Index auffallend.

Dieser Index wurde dann anhand der Referenzsorten in die entsprechenden Bayerischen Winterhärtenoten eingeteilt.

Wie man auf S. 23, Abschnitt 1.3 sehen kann, liegen die besonderen Schwierigkeiten bei den vorhandenen Datensätzen:

- in der Menge der Merkmale,
- den örtlichen Unterschieden,
- in den unbesetzten Zellen.

Aufgrund der vielen eingehenden Komponenten brachte die deskriptive Statistik keinerlei neue Erkenntnisse.

3.2 Problemstellung

Die Aufgabe, eine Schätzgleichung für die Bayerische Winterhärtenote zu erstellen, bedeutet, dass man Noten als Eingangsgrößen verwenden und Noten als Ergebnis erhalten möchte.

Da keine Sicherheit für die Beibehaltung der Versuchsorte besteht, sollte die Gleichung auch die Orte entsprechend berücksichtigen. Damit ist bei Ausfall eines Ortes die Gleichung immer noch verwendbar. Dies ist auf zwei verschiedene Arten denkbar:

1. Es wird für jeden Ort eine Gleichung aufgestellt, die dann addiert und anschließend durch die Anzahl der Orte geteilt wird.
2. Man vollzieht eine Mittelwertbildung über alle Orte pro Versuch, Sorte und Nutzungsjahr, so dass der einzelne Ort nicht mehr das Endergebnis beeinflusst

Grundvoraussetzung für die weitere Überprüfung einer solchen Formel in der Praxis wäre eine möglichst fehlerfreie Berechnung der vorhandenen Zielgröße. Die Fehleinschätzungen von mehr als 5 von 52 Endnoten lassen die Vermutung

zu, dass sich bei anderen Bonitureingaben die falschen Bewertungen häufen werden, was in jedem Fall zu vermeiden ist. Damit wurde die zulässige Fehlerquote von null bis fünf festgelegt.

Die Gleichung darf der Praxis nicht widersprechen z. B. muss eine positive Bewertung eines Merkmals, eine positive Auswirkung auf die Zielvariable besitzen.

Zur Durchführung der mathematischen Methoden ist zuerst eine Zusammenfassung der Datenbasis (35 Merkmale, welche gering besetzt sind) absolut notwendig. Denkbare Möglichkeiten für eine Zusammenfassung sind:

- Nach Zeitpunkt / Verwandtschaft der Merkmale, sachlogisch begründbar
- Nach Sorte, Nutzungsjahr und Versuch; besitzt den Vorteil, dass die extremen Ortsunterschiede geglättet werden.
- Nach Korreliertheit untereinander (Weidelgrasanteil³⁵ und Dichtigkeit hängen zusammen)
- Statistische Verfahren (Hauptkomponentenanalyse, Clusteranalyse)

Korrelationsuntersuchungen bei ordinalen Daten werden üblicherweise mit dem Spearmanschen³⁶ Korrelationskoeffizienten durchgeführt. Die multiple Regression geht jedoch von stetigen Merkmalen aus. Womit nur noch der Pearsonsche Korrelationskoeffizient³⁷ in Frage kommt. Für die *Dichtigkeit* und den *Weidelgrasanteil* wurden nach einer zeitpunktmäßigen und einer gemittelten Zusammenfassung folgende Korrelationen³⁸ ermittelt. (Tabelle 14)

³⁵ zur Erklärung von Dichtigkeit und Weidelgrasanteil siehe S. 12

³⁶ Für die Formel siehe Sachs [10], S. 309

³⁷ Für die Berechnung siehe Sachs [10], S. 315

³⁸ Programm siehe S. 71

Tabelle 14: Spearmannsche Korrelationskoeffizient zwischen Dichtigkeit und Weidelgrasanteil

Zusammengefasst	vor Winter	nach Winter	Schnitt 1	Schnitt 2	letzter Schnitt
zeitpunktmäßig	0.46	0.05	0.57	0.54	0.60
und gemittelt	0.86	0.64	0.36	0.38	0.45

Für die Korrelationskoeffizienten bei der zeitpunktmäßigen Zusammenfassung gehen die Ortsunterschiede mit ein. Die gemittelten Werte zeigen einen deutlichen bis mäßigen Zusammenhang bei allen Erhebungen.

Bemerkung zur Beurteilung der Koeffizienten: Je näher der Koeffizient an $|1|$ liegt, desto stärker der Zusammenhang.

3.3 Multiple Regression

Eine multiple Regression sollte ein geeignetes Instrument für die Erstellung solch einer Formel sein. Diese muss im Prinzip dann folgende Form aufweisen:

$$\begin{aligned}
 \text{Baywihae} = & \text{Merkmale des 0-ten Nutzungsjahres} \\
 & + \text{Merkmale des 1-ten Nutzungsjahres} \\
 & + \dots \\
 & + \text{Merkmale des 4-ten Nutzungsjahres}
 \end{aligned}$$

Für die Zielvariable y_i des allgemeinen Ansatzes der multiplen linearen Regression von S. 30 wird hier eine bestimmten Note der Bayerischen Winterhärte (*Baywihae*) eingesetzt. Der rechten Seite entsprechen die zugehörigen Kombinationen der Merkmale.

Andere Arten der Formel, wie z. B.: $\text{Baywihae} = \text{Nutzungsjahr} \cdot (\text{Merkmale})$, würden eine schlechte Annäherung bedeuten, da versucht würde, einen sehr großen Schwankungsbereich mit zu wenigen Parametern zu beschreiben.

Tabelle 15: Auflistung der unabhängigen und abhängigen Variablen

Art der Daten	unabhängige Variablen pro Nutzungsjahr		abhängige V. = Zielvariable
	Bezeichnungen	Abkürzungen	
ursprünglich	Weidelgrasanteil vor Winter	www, Wantvw	Bayerische Winterhärte (BwH, Baywihae)
	Dichtigkeit vor Winter	dvw, Dichvw	
	Stand vor Winter	scw, Standvw	
	Weidelgrasanteil nach Winter	wnw, Wantnw	
	Dichtigkeit nach Winter	dnw, Dichnw	
	Stand nach Winter	snw, Standnw	
	Weidelgrasanteil vor Schnitt n	wvs., Wantvs..	
	Weidelgrasanteil nach Schnitt n	wns., Wantns..	
	Dichtigkeit vor Schnitt n	dvs., Dichvs...	
	Dichtigkeit nach Schnitt n	dns., Dichns...	
	Weidelgrasanteil bei Vegetationsende	wve, Wantve	
	Dichtigkeit bei Vegetationsende	dve, Dichve	
	Fusariumresistenz	fr, Fusrmr	
Rostresistenz	rr, Rostr		
Zusammengefasst nach Bestand (Weidelgrasanteil und Dichtigkeit)	Bestand vor Winter	Bestvw	Bayerische Winterhärte (BwH, Baywihae)
	Stand vor Winter	Stvw	
	Bestand nach Winter	Bestnw	
	Stand nach Winter	Stnw	
	Bestand vor Schnitt n	Bestvs..	
	Bestand nach Schnitt n	Bestns..	
	Fusariumresistenz	Fusrmr	
	Rostresistenz	Rost1_..	
Zusammengefasst nach Bestand und Zeitpunkt	Zusammen Bestand vor Winter	zubvw	Bayerische Winterhärte (BwH, Baywihae)
	Zusammen Bestand nach Winter	zubnw	
	Bestand vor Schnitt n	bestvs..	
	Zusammen Bestand letzter Schnitt	zubvs3	
	Fusariumresistenz	Fusrmr	
	Rostresistenz	Rost1_	

Für jedes Nutzungsjahr wurde die Zahl an die Abkürzung angefügt (Bsp.: zubvw1, bestvs11, bedeuten Bestand vor Winter bzw. vor Schnitt 1 im ersten Nutzungsjahr)

Da es bei einer Variablenselektion darauf ankommt, welche Variablen bereits im Modell sind, um die Bedeutung eines Prädiktors für das Modell festzustellen, bietet das Verfahren der Backward-Elimination den Vorzug, dass man von einem vollen Modell ausgeht. Die unbedeutenden Variablen herauszufiltern,

ohne wichtige zu verlieren, wird vermutlich eher gelingen als bei anderen Variablenselektionen. Das vorgegebene Signifikanzniveau von 0.1 ist auch sehr gut geeignet, um die besten Schätzungen zu bekommen³⁹.

Deshalb wurde die Backward-Elimination bei dieser Diplomarbeit generell angewandt.

Die Behandlung der leeren Zellen war ein Problem, das mit Auffüllen von verwandten Merkmalen oder Verwendung der Mittelwerte angegangen wurde.

Allerdings verzerrte das die Schätzungen so sehr, dass es als vorteilhaft angesehen wurde, leere Zellen als eigene Ausprägung mit Null zu bewerten.

Untersuchung der Einzelorte

Ausgehend vom ersten Fall von Abschnitt 3.2, S. 45 wurde also für jeden Einzelort separat eine multiple Regression durchgeführt.

Verwendet man dazu die Originaldaten, so stellt man fest, dass die in die Regression eingehende Komponentenanzahl enorm hoch ist (27 bis 44 Komponenten) und die Standardabweichungen der einzelnen Parameter teilweise über 5 liegen. Beides spricht für die Zusammenfassung der Merkmale, die nach Zeitpunkt und Verwandtschaft erfolgte⁴⁰.

Mit der Merkmalszusammenfassung erreicht man durchaus akzeptable Regressionsergebnisse (Bestimmtheitsmaß > 0.75) und auch die Standardabweichungen liegen alle unter eins. Allerdings ist keinerlei Systematik bei der Wahl der Merkmale zu erkennen. Da die multiple Regression eine möglichst große Unkorreliertheit der Prädiktoren verlangt, wurden die Merkmale *Weidelgrasanteil* und *Dichtigkeit* zum Merkmal *Bestand*⁴¹ modifiziert.

Dadurch verschlechterten sich die Bestimmtheitsmaße im Durchschnitt auf 0.6 und die relevante Komponentenanzahl verkleinerte sich bis max. 12 in Hoetzeltdorf und setzte sich nun ortsspezifisch zusammen.

³⁹ SAS User's Guide [14], S. 765

⁴⁰ Nähere Beschreibung der Zusammenfassung S. 19

⁴¹ Nähere Beschreibung des Bestands S. 21

Aufgrund der ortsbedingten Signifikanz der Merkmale, und der damit auf lange Sicht unterliegenden Unsicherheit von diesen, sagt eine ortsspezifische Regression wenig über das Deutsche Weidelgras aus.

Um bessere Ergebnisse zu erhalten, beschäftigen wir uns nun mit der zweiten Möglichkeit, der Berücksichtigung der Orte von 3.2.

Mittelwertbildung über die Orte

Normalerweise ist die Bewertung eines Medians bei ordinalen Daten aussagekräftiger als der Mittelwert. Allerdings sind in der bisherigen Berechnung die Mittelwerte benutzt worden, weshalb dieses Maß beibehalten wurde.

Die Verwendung des arithmetischen Mittels über alle Orte pro Versuch, Nutzungsjahr und Sorte hat wie bereits erwähnt den Vorteil, ortsspezifische Schwankungen bezüglich Boniturnoten fehlender Zellen abzuschwächen.

Die Art und Reihenfolge der Zusammenfassungen trägt maßgeblich zur Signifikanz der einzelnen Merkmale bei. Es kommt also darauf an, die richtige Anordnung der Mittelungen zu wählen.

Die im Anschluss beschriebene Regression liefert das beste Ergebnis. Alle anderen Fälle arbeiten entweder mit einer zu großen Menge von einfließenden Komponenten, oder das Bestimmtheitsmaß liegt um 0.80, was eine Fehlerquote von mehr als 10 Noten bedeutet.

3.3.1 Regression mit Mittelung über alle Orte und nachfolgender Bestandserzeugung

Zuerst wurde eine Mittelung über alle Orte pro Nutzungsjahr, Sorte und Versuch, separat für jedes Merkmal vorgenommen. Danach wurde der *Bestand*⁴² erzeugt, ohne Berücksichtigung des vierten Jahres, da dies nur in Ausnahmefällen zur Verfügung steht und ohne die Bonituren *nach Schnitt*⁴³ n.

Mit diesen zwei Veränderungen der Originaldaten wurde dann eine Regressionsanalyse durchgeführt.

⁴² Erläuterung des Begriffs Bestand S. 21

⁴³ ein Schnitt bedeutet das Gras mähen, es gibt ca. 3 bis 6 Schnitte pro Jahr je nach Standort

Es ergab sich das Modell:

$$\begin{aligned} Baywihae = & -0.41 \cdot Stnw3 + 0.3 \cdot Bestnw3 + 0.31 \cdot Bestvs43 - 0.19 \cdot Rostr1_3 - 0.56 \cdot Bestvw3 \\ & + 0.38 \cdot Bestvs32 + 0.74 \cdot Bestvs42 + 0.46 \cdot Bestvs52 - 0.23 \cdot Bestvw2 \\ & + 0.41 \cdot Stnw1 + 0.41 \cdot Bestvs11 + 0.35 \cdot Rostr1_1 \\ & - 9.88 \end{aligned}$$

Hier ziehen bereits die negativen Vorzeichen einzelner Merkmale die Praxis-tauglichkeit der Formel in Zweifel. Wenn man nun die Kollinearitätsanalyse vom Ergebnisausdruck¹, S. 76 betrachtet, so stellt man eine Kollinearität (Erklärung siehe S. 33) zwischen der *Rostresistenz* des ersten Nutzungsjahres und dem *Stand nach Winter* des ersten Nutzungsjahres sowie dem *Bestand vor Winter* des dritten Nutzungsjahres fest. Begründung⁴⁴:

- Der Betrag der Pearsonschen Korrelationskoeffizienten der genannten Variablen gegenüber der *Rostresistenz* liegt bei allen genannten Variablen über 0.7.
- Die Toleranz liegt bei den genannten Merkmalen unter 0.1 – dies deckt eine enge Korrelation der Merkmale auf.
- Der „condition index“⁴⁵ bei der *Rostresistenz* ist höher 30, woran man die Auswirkungen einer Multikollinearität erkennt.
- Die kollinearitätsabhängigen Varianzen liegen bei der *Rostresistenz* und beim *Bestand vor Schnitt 4* des zweiten Jahres über 0.6 . Verbunden mit den kleinen Eigenwerten (unter 0.002) hat dies die Bedeutung, dass die Genauigkeit der Schätzungen beeinträchtigt ist.

Allerdings sind die Standardfehler $\sigma_{\hat{y}(x)}$ des individuellen vorhergesagten Wertes zwischen 0.42 und 0.54, was besagt, dass keine extrem starke Multikollinearität vorliegt.

⁴⁴ Myers [9], Kapitel 3 und Kapitel 8

⁴⁵ Faustzahl beträgt 30 siehe Myers [9], S. 370

Zur Vermeidung der vorhandenen Multikollinearität und Überprüfung der Werte schließt sich eine Hauptkomponentenanalyse an.

Hauptkomponentenanalyse

Ein Gleichungssystem ist dann eindeutig lösbar, wenn die Eingangsmatrix vollen Rang besitzt. Dies gilt insbesondere für die Hauptkomponentenanalyse. Sie ist deshalb nicht durchführbar, wenn man die Originaldaten ($35 \times 3 = 105$) für 52 Sorten verwenden möchte. Es wurde also mit den in die Regression, von Ergebnisausdruck1, eingegangenen Werten weitergerechnet.

Die MSA-Werte geben bei Bestvs13 einen Hinweis auf Aussonderung. Damit wurden Bestvs13 (MSA-Wert: 0.49) und alle Merkmale, die keine Werte enthielten (*Bestand vor Schnitt 6* im 2. Nutzungsjahr, *Stand vor Winter* im ersten Nutzungsjahr und *Bestand am Vegetationsende* im Ansaatjahr), eliminiert.

Die nächste Frage betrifft die Anzahl der zu verwendenden Faktoren. Zwar offenbart der Screenplot verschiedene Knicke (bei 4,7,10,15), auch ist eine möglichst hohe Erklärung der Gesamtvarianz wünschenswert. Allerdings wird die Korrekturfunktion der Faktoren immer größer und eine Interpretation immer schwieriger je mehr Faktoren man verwendet.

Hält man sich an die Faustregeln von BORTZ [4], S. 509, so muss man sich meist für das Kaiser-Kriterium entscheiden. In dem Ergebnisausdruck1.1, S. 80 sind dies vier Faktoren. Da man mit der Zielsetzung arbeitet, am Ende eine multiple Regressionsanalyse mit diesen Faktoren durchzuführen, wird grundsätzlich mit Varimax (Beibehaltung der Unkorreliertheit der Faktorvariablen) rotiert um die Ladungen zu erhalten (siehe S. 81).

Die Interpretation der Ladungen kann man folgendermaßen auslegen (Tab. 16):

Tabelle 16: Interpretation der Faktorladungen

	Interpretation
Factor1	Fusariumresistenz d. 1. Jahres und Rostresistenz aller drei Jahre Bonituren der ersten Hälfte des 2. Jahres Bonituren vor Winter aller drei Jahre
Factor2	Krankheiten im 2. Jahr Stand nach Winter des 1. und 3. Jahres Bonituren vor Winter des 0. bis 2. Jahres
Factor3	Schnitte des 3. Jahres Bonituren der zweiten Hälfte des 2. Jahres Bestand vor Winter des 1. Jahres
Factor4	Schnitte des 1. Nutzungsjahres

Eine mit den rotierten Faktorwerten durchgeführte Variablenselektion eliminiert zuerst den Factor2 als unbedeutend (ein Großteil davon wird durch Factor1 ausgedrückt). Die sich anschließende Regression erzeugt folgende Formel:

$$\begin{aligned}
 \text{Baywihae} = & 0.84 \cdot \text{factor3} (\text{Schnitte des 3.J./Bestvw des 1.J./ Bonituren 2. Hälfte v. 2.J.}) \\
 & + 0.33 \cdot \text{factor4} (\text{Schnitte des 1.J.}) \\
 & - 0.14 \cdot \text{factor1} (\text{Fusrmr 1.J./ Rostr aller J./ Bonituren 1. Hälfte v. 2.J / Bonituren vor Winter aller J.})
 \end{aligned}$$

Der Factor1 weist auf einen schwach negativen Einfluss der Krankheiten und der Bonituren *vor Winter* sowie der ersten Hälfte des zweiten Nutzungsjahres, hin. Er schwächt auch den sehr starken Einfluss des dritten Faktors bezüglich des *Bestandes vor Winter* im ersten Nutzungsjahr, während sowohl die Bonituren der zweiten Hälfte des zweiten Jahres als auch die *Schnitte* des dritten Jahres durchaus ausschlaggebend für die Notengebung sind. Die *Schnitte* des ersten Jahres (Factor4) tragen mittelmäßig zur Bayerischen Winterhärte bei.

Kombiniert man nun beide Regressionen, so führt dies zu folgenden Schlussfolgerungen (Tabelle 17):

Tabelle 17: Ergebnis zweier Regressionsanalysen (*Bestand, Hauptkomponentenanalyse*)

Einfluss	Merkmale
Positiv	Bestand vor Winter im ersten Nutzungsjahr
	Zweite Hälfte des zweiten Nutzungsjahres
	Schnitte des dritten Nutzungsjahres
Negativ	Krankheiten (schwacher Einfluss)
	Bonituren vor Winter
	Erste Hälfte des zweiten Nutzungsjahres
Mittelmäßig	Bonituren nach Winter im ersten Nutzungsjahr
	Schnitte des ersten Nutzungsjahres
	Bonituren nach Winter im dritten Nutzungsjahr
Unbedeutend	Bonituren des Ansaatjahres
	Fusariumresistenz im dritten Nutzungsjahr

3.3.2 Regression mit Mittelung über alle Orte und nachfolgender Bestandserzeugung und Zusammenfassung nach Zeitpunkten

Um eine klarere Vorstellung der Einflüsse zu bekommen, kann man die eingegangenen Werte der vorherigen Untersuchung noch nach Zeitpunkten mitteln und die ganze Analyse wiederholen. Im Ergebnisausdruck², S. 83 ist das Resultat dargestellt. Die erzeugte Gleichung wird hier nochmals aufgeführt:

$$\begin{aligned}
 Baywihae = & 0.59 \cdot Bestvs22 + 1.27 \cdot Zusbvs32 - 0.16 \cdot rostr1_2 - 0.33 \cdot zusbv2 \\
 & + 0.6 \cdot Bestvs11 + 0.21 \cdot Rostr1_1 \\
 & - 8.88
 \end{aligned}$$

Bei der Variablenselektion wurde das ganze dritte Jahr eliminiert, die eingegangenen Variablen zeigen einen negativen Einfluss der *Rostresistenz* und des Zustandes der Versuche *vor Winter* im zweiten Jahr. Dagegen besitzen die *Schnitte* von diesem Jahr ein sehr großes positives Gewicht. Das erste Jahr erhöht ebenfalls die Zielvariable mit dem ersten *Schnitt* und mit der *Rostresistenz*.

Bei dieser Regressionsanalyse werden 15 Noten falsch eingeschätzt.

Das Ergebnis wird relativiert, wenn man bei der Variablenselektion vom zweiten Jahr als Eingangsbasis nur die letzten Schnitte zulässt (Ergebnis 2 a):

Ergebnis 2.a: Regression (gemittelt, nach *Bestand* und Zeitpunkt zusammengefasst, ohne zweites Jahr, außer letztem *Schnitt*)

Variable ZUSBVW3 Removed R-square = 0.87653379 C(p) = 0.20854503

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-10.35230470	1.59553977	10.04656646	42.10	0.0001
ZUSBVS32	1.32151985	0.13820401	21.82051109	91.43	0.0001
BESTVS11	0.47909931	0.12784022	3.35177525	14.04	0.0005
ZUSBVS33	0.31568245	0.12618910	1.49353711	6.26	0.0161
FUSRMR1	0.31288693	0.11858616	1.66136744	6.96	0.0114
FUSRMR2	-0.19863424	0.11063287	0.76930528	3.22	0.0793
ROSTR1_1	0.18072679	0.03809247	5.37187092	22.51	0.0001

Man sieht hier sehr deutlich, dass die Regression sich fast genauso gut den beobachteten Werten annähert. Nimmt man noch den Zustand *vor Winter* des zweiten Nutzungsjahres dazu (Ergebnis 2 b), dann ist die Regressionsanalyse qualitativ gleichwertig, wobei die letzten *Schnitte* des dritten Jahres sehr wohl einen Einfluss zeigen.

Ergebnis 2.b: Regression (gemittelt, nach *Bestand* und Zeitpunkt zusammengefasst, ohne zweites Jahr, außer letztem *Schnitt* und *vor Winter*)

Variable ROSTR1_3 Removed R-square = 0.88956671 C(p) = 1.05425713

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-10.07582242	1.55245746	9.40971298	42.12	0.0001
ZUSBVW2	-0.23524526	0.12207102	0.82960507	3.71	0.0606
ZUSBVS32	1.23475037	0.14213395	16.85844227	75.47	0.0001
BESTVS11	0.42187983	0.16701838	1.42529181	6.38	0.0153
ZUSBVS33	0.44835973	0.14241816	2.21399796	9.91	0.0030
BESTVS21	0.20268488	0.14635072	0.42845755	1.92	0.1732
FUSRMR1	0.24622103	0.11863782	0.96218535	4.31	0.0440
FUSRMR2	-0.14752884	0.11312982	0.37988702	1.70	0.1991
ROSTR1_1	0.18103504	0.03836683	4.97357014	22.26	0.0001

Die nachfolgende Durchführung einer Hauptkomponentenanalyse bestätigt die beiden letzten Arten der Regression.

Die Interpretation der Ladungen erfolgt folgendermaßen:

Tabelle 18: Interpretation der Faktorladungen

	Interpretation
Factor1	Bonituren vor Winter des Ansaatjahres Bonituren des ersten Nutzungsjahres
Factor2	Rostresistenz des ersten Nutzungsjahres Krankheiten im zweiten Nutzungsjahr Bonituren der ersten Hälfte des zweiten Nutzungsjahres
Factor3	Fusariumresistenz im ersten Nutzungsjahr Bonituren des letzten Schnittes im zweiten Nutzungsjahr Rostresistenz im dritten Nutzungsjahr Zustand vor Winter des dritten Nutzungsjahres
Factor4	Korrekturfaktor für Bonituren nach Winter im dritten Nutzungsjahr und beim letzten Schnitt des dritten Nutzungsjahres

Der letzte Schnitt des 3. Jahres wurde sowohl im Factor1 als auch Factor4 berücksichtigt. Alle Bonituren des dritten Jahres kann man als große Einflussfaktoren ansehen, besitzen sie doch über alle Faktoren gesehen, Werte über 0.5

Deshalb lautet die sich ergebende Gleichung aus der Regressionsanalyse mit den rotierten Faktorwerten:

$$\begin{aligned}
 \text{Baywihae} = & 0.59 \cdot \text{factor1}_{(\text{Bonituren vor Winter 0.J. / Bonituren des 1.J.})} \\
 & + 0.45 \cdot \text{factor4}_{(\text{Korrekturfaktor für n. Winter 3.J. und letzter Schnitt 3.J.})} \\
 & + 0.43 \cdot \text{factor3}_{(\text{Fusmr 1.J. / letzter Schnitt 2.J. / Rostr 3.J. / vor Winter 3.J.})} \\
 & + 0.25 \cdot \text{factor2}_{(\text{Rostr 1.J. / Krankh 2.J. / Bonituren erste Hälfte 2.J.})}
 \end{aligned}$$

Den größten Einfluss auf die Endnote haben hier der Bewuchs der letzten *Schnitte* sowie die Bonituren *nach Winter* des dritten Jahres (diese beiden kompensieren sich in Factor1, Factor3 und Factor4) und am Anfang des Versuches (Factor1). Danach folgt die *Fusariumresistenz* des ersten, der letzte *Schnitt* des zweiten, die *Rostresistenz* und der *Bestand vor Winter* des dritten Jahres. Die Bonituren der ersten Hälfte und Krankheiten des zweiten Jahres sowie die *Rostresistenz* des ersten Jahres spielen ebenfalls eine Rolle (Factor2).

Es ist problematisch aus den Regressionsanalysen eine Übersicht zu ermitteln. Hält man sich an die Hauptkomponentenanalyse mit Ergänzung der beiden Resultate von S. 55 dann kann man folgende Einteilung vornehmen:

Tabelle 19: Ergebnis zweier Regressionen

(Zeitpunkt zusammengefasst, Hauptkomponentenanalyse)

Einfluss	Merkmale
Stark	Bonituren des ersten Jahres
	Bonitur des letzten Schnittes des zweiten Nutzungsjahres
	Vor / nach Winter des dritten Nutzungsjahres
Weniger stark	Rostresistenz des dritten Nutzungsjahres
Mittelmäßig	Krankheiten des zweiten und ersten Nutzungsjahres
	Erste Hälfte des zweiten Nutzungsjahres
Unbedeutend	Fusariumresistenz im dritten Jahr
	Ansaatjahr

3.3.3 Zusammenfassung

Aus den bisherigen Untersuchungen kann man nachfolgende Grunderkenntnisse gewinnen:

- Hohe Noten neigen dazu, mit negativem oder geringem Einfluss bewertet zu werden
- Das Ansaatjahr und die *Fusariumresistenz* im dritten Jahr besitzen einen unbedeutenden Faktorwert für die Regressionsmodelle.
- Die Krankheiten und die erste Hälfte des zweiten Nutzungsjahres beeinflussen die Endnoten weniger. Wobei die *Rostresistenz* im ersten Nutzungsjahr wegen ihrer Präsenz bei jedem untersuchten Modell auffällt.
- Haupteinflussfaktoren sind die Bonituren des ersten, der Zustand der Pflanze *vor Winter* im zweiten Nutzungsjahr und die ganzen Bonituren des dritten Nutzungsjahres
- Des weiteren ist aus den eingehenden Merkmalen ersichtlich, dass sie mehr die Regenerationsfähigkeit der Pflanze erfassen und eher indirekt deren Ausfall. Es ist also noch interessant, den Einfluss des Winters und des Fusariumsbefalls auf den Zustand der Versuche zu analysieren. Die im Anschluss aufgeführte Logitanalyse erforscht dies.

3.4 Logitanalyse

Möchte man die Auswirkungen von Kälte und gleichzeitig auftretenden Krankheiten auf die Pflanze überprüfen, so hat man mit dem Material zu arbeiten, das beides aufweist. Im Speziellen bedeutet dies hier, dass die zu verwendenden Merkmale die Bonituren *vor* und *nach Winter*, die Einteilung des *Winters* und die *Fusariumresistenz*, dort wo sie aufgetreten ist, sind. Gerade Letztere reduziert die zu verwendenden Beobachtungszeilen auf knapp die Hälfte der Originaldaten. Dabei entfallen der Ort Auerberg und das Ansaatjahr.

Da es sich bei der Logitanalyse um Häufigkeiten des Auftretens der einzelnen Bonituren handelt, ist es sinnvoll, so viele Einzelbonituren wie möglich einzubeziehen. Darum verbietet sich eine Mittelung über alle Orte von selbst. Allerdings sollte man darauf achten diese möglichst kompakt, für eine hohe Zellenbesetzung zu bekommen. Hierdurch ergibt sich eine Merkmalsreduzierung.

Deshalb wurden die Merkmale zeitpunktmäßig zusammengefasst (Mittelung: *dichve / dichvw / standvw, wantvs1 / wantns1* usw.). Auch rechtfertigt sich eine Auffüllung der leeren Zellen von *vor / nach Winter* mit dem letzten bzw. ersten *Schnitt*, da es auf den Zustand der Pflanze vor und nach der vegetationsarmen Zeit ankommt und sie dann nur einmal in die Berechnung eingehen, was somit keine Verzerrung der Datenbasis darstellen kann.

Schlussendlich wurde aus dem *Weidelgrasanteil* und der *Dichtigkeit* künstlich der *Bestand* gebildet. Anschließend teilt man die Merkmale in die interessanten Gruppen ein: schlecht (1), mittel (2), gut (3) bzw. hart (1), normal (2), mild (3) für den *Winter*⁴⁶.

Die *Differenz über bzw. nach Winter* wird mit *Bestand nach Winter* (aktuelles Jahr) minus *Bestand vor Winter* (Vorjahr) berechnet. Diese Variable ist zur weiteren Analyse die Aussagekräftigste.

In Tab. 20 ist die Datenbasis für die weitere Untersuchung aufgeführt:

⁴⁶ zur Einteilung siehe auch Programm S. 74

Tabelle 20: Relevante Merkmale für die Logitanalyse: *Differenz nach Winter.*

Beobach- tungen	Merkmale			Häufigkeiten			
				Differenz Winter			Gesamt
	N	NUTZGSJ	WINTER	FUSRMR	_1	_2	_3
1	1	2	1	16	2	0	18
2	1	2	2	18	35	8	61
3	1	2	3	24	61	14	99
4	2	1	1	1	2	0	3
5	2	1	2	5	7	0	12
6	2	1	3	11	7	0	18
7	2	2	1	19	9	0	28
8	2	2	2	18	2	0	20
9	2	2	3	0	21	0	21
10	2	3	1	1	6	0	7
11	2	3	2	8	13	0	21
12	2	3	3	1	7	0	8
13	3	2	1	17	1	0	18
14	3	2	2	41	36	0	77
15	3	2	3	15	41	0	56
16	4	1	1	10	0	0	10
17	4	1	2	9	0	0	9
18	4	1	3	17	0	0	17
			Gesamt	231	250	22	503
			Prozent	45.9	49.7	4.4	100

Die Abkürzungen bedeuten: *Nutzgsj*: Nutzungsjahr, *Diffwi*: Differenz über Winter, *Fusrmr*: Fusarium-resistenz. Die graue Unterlegung wurde nur zur besseren Lesbarkeit / Unterscheidung der Nutzungsjahre vorgenommen.

Es zeigt sich, dass ein positiver Notenunterschied bei insgesamt 22 Fällen eintritt, aber nur wenn die Kombination erstes Nutzungsjahr, normaler *Winter* und mindestens normale *Fusariumresistenz* besteht.

Versucht man mit den drei Kategorien ein kumulatives Logitmodell auszuführen, so kann man dies zwar berechnen, allerdings erhält man aufgrund der vielen Leerzellen bei der Ausprägung *_3* von *Differenz Winter* eine Inhomogenität der Parameter, was bedeutet, dass das Modell für diese Daten inadäquat ist.

Um ein binäres Logitsystem zu bekommen addiert man die dritte Kategorie zu den neutralen Notenunterschieden.

Zwar besitzt das vierte Nutzungsjahr eine Fusariumdifferenzierung, allerdings war das der Winter 1995 /1996, welcher der extremste Winter der letzten 18 Jahre gewesen ist. Deshalb gab es nur negative Differenzen im Frühjahr.

Kommt nun der Fall vor, dass für eine Häufigkeit die Zahl null besteht, so hat man zur Parameterschätzung entweder diesen Fall zu eliminieren oder ihm einen sehr kleinen Wert zuzuweisen (z. B. 1/ Kategorienanzahl der Zielvariablen). Eine Zuweisung stellt immer eine Verzerrung dar. Die Elimination ist, soweit man sie vertreten kann, dieser vorzuziehen.

Für das weitere Vorgehen wurde darum das vierte Jahr entfernt.

Damit veränderten sich die Anteile in Richtung neutral:

Differenz Winter	(1) schlecht	(2) neutral	Gesamt
Häufigkeiten	195	272	467
Prozent	41,8	58,2	100

3.4.1 Untersuchung von Nutzungsjahr, Winter, Fusarium-Resistenz (nur Haupteffekte)

Die Effektkodierung müsste man selbst programmieren. SAS stellt für die Dummykodierung eine Prozedur bereit, deshalb wurde diese verwendet.

Aus Tab.20 (ohne das 4. Jahr und `_2` und `_3` addiert) ergibt sich dann folgende Designmatrix. (Die Wahrscheinlichkeit $\pi = (_1 / \mathbf{ges})$ wurde noch daneben platziert. Sie wird bei der Berechnung dann zum Logit modifiziert.)

INTERCEPT	NUTZGS1 1.Jahr	NUTZGS2 2.Jahr	WINTER1 hart	WINTER2 normal	FUSRMR1 schlecht	FUSRMR2 normal	π negativ
1	1	0	0	1	1	0	0.88889
1	1	0	0	1	0	1	0.29508
1	1	0	0	1	0	0	0.24242
1	0	1	1	0	1	0	0.33333
1	0	1	1	0	0	1	0.41667
1	0	1	1	0	0	0	0.61111
1	0	1	0	1	1	0	0.67857
1	0	1	0	1	0	1	0.90000
1	0	1	0	1	0	0	0.00000
1	0	1	0	0	1	0	0.14286
1	0	1	0	0	0	1	0.38095
1	0	1	0	0	0	0	0.12500
1	0	0	0	1	1	0	0.94444
1	0	0	0	1	0	1	0.53247
1	0	0	0	1	0	0	0.26786

Die Referenzkategorien werden nicht aufgeführt (wegen linearer Abhängigkeit). SAS eliminiert den Fall $y=0$. Falls man die Null mit 0.3 ersetzt und durchrechnet, so ist keinerlei gravierende Änderung sichtbar, womit die Aussonderung durchaus akzeptabel ist.

Das daraus resultierende Ergebnis lautet (siehe auch S. 88):

$$\begin{aligned}\hat{\pi} = & -0.521 \cdot \text{Nutzungs}(1) - 0.162 \cdot \text{Nutzungs}(2) \\ & + 2.137 \cdot \text{Fusrmr}(1) + 0.996 \cdot \text{Fusrmr}(2) \\ & + 1.580 \cdot \text{Winter}(1) + 1.061 \cdot \text{Winter}(2) \\ & - 1.8875\end{aligned}$$

Die jeweilige Referenzkategorie ist hierbei nicht aufgeführt, weil die Regressionskonstante den Mittelwert der Logits von dieser darstellt und die Regressionskoeffizienten die Abweichungen von diesem. Jetzt hat man mehrere unabhängige Variablen miteinfließen lassen, somit ist der Intercept nur noch eine Verschiebung der Geraden. Obwohl also z. B. kein drittes Jahr aufgeführt ist - die Abweichung von dem Mittelwert des dritten Jahres ist ja null, beziehen sich alle Gewichte der Nutzungsjahre darauf.

Allein von den Regressionskoeffizienten lässt sich folgendes ableiten:

Eine schlechte *Fusariumresistenz* hat den größten Einfluss auf eine hohe Wahrscheinlichkeit, dass die Pflanze einen schlechten Zustand im Frühjahr besitzt. Zweitgrößter Einflussfaktor ist der harte *Winter*. Beim normalen *Winter* mit normaler *Fusariumresistenz* ist das Nutzungsjahr ausschlaggebend.

Eine stärkere Aussagekraft besitzen die Odds-Ratios.⁴⁷

Tabelle 21: Analyse des schlechten Zustandes der Pflanze im Frühjahr

Unabhängige Variablen	Schätzparameter ($\hat{\beta}$)	Odds-Ratios ($e^{\hat{\beta}}$)
INTERCPT	-1.8875	.
NUTZGS1	-0.5214	0.594
NUTZGS2	-0.1621	0.850
WINTER1	1.5806	4.858
WINTER2	1.0611	2.889
FUSRMR1	2.1376	8.479
FUSRMR2	0.9969	2.710

Im ersten Nutzungsjahr trifft ein negatives Aussehen der Pflanzen in der Versuchsparzelle um 68.4% ($1 / 0.594$) weniger ein als im dritten Jahr. Zum zweiten Nutzungsjahr ist keine Aussage möglich (siehe auch Konfidenzintervall S. 88). Diese Ausprägung würde bei einer Variablenselektion auch als unbedeutend eliminiert werden.

⁴⁷ zur Interpretation der Odds-Ratio siehe Andreß [2], S. 276-280

Der schlechte Zustand im Frühjahr wird von einem harten *Winter* 4.9 (390%) und von einem normalen *Winter* 2.9 (190%) mal stärker beeinflusst als von einem milden *Winter*.

Eine schlechte *Fusariumresistenz* steigert die Wahrscheinlichkeit eines schlechten Zustandes des WD nach dem Winter um den Faktor 8.5 als bei einer guten *Fusariumresistenz*. Falls sie mittelmäßig ist, ist die Gefahr des schlechten Zustandes der Pflanze immer noch 2.7 mal höher.

Nun weiß man aus Erfahrung, dass das Fusariumaufkommen mit dem Winter zusammenhängt. Man muss also noch Wechselwirkungen berücksichtigen. (Deshalb auch das mäßige Bestimmtheitsmaß von 0.14). Dazu vergleiche man nochmals Tabelle 20, S. 60.

Es gibt dort keinerlei Differenzierung des *Winters* zu den Nutzungsjahren (er hat immer die Note 2 [normal]). Eine Untersuchung der Wechselwirkungen *Fusarium / Winter* und *Fusarium / Nutzungsjahr* wird deshalb immer eine größere Bedeutung der Nutzungsjahre hervorbringen, was auch akzeptabel ist, da der normale Winter immer noch vorherrscht. Die Analyse beschränkt sich deshalb auf die Interaktion von *Fusariumresistenz / Winter*.

Zwischenbemerkung: Da im zweiten Nutzungsjahr alle Wetter- und Krankheitsverhältnisse vorkommen, könnte man dieses als repräsentatives Beispiel etwas genauer analysieren. Allerdings sind die Häufigkeiten der schlechten Pilzresistenz etwas anders als erwartet. Bei genauerer Betrachtung stellt man fest, dass die Häufigkeiten der Ausprägung $_2$ von *Differenz Winter* bei schlechter *Fusariumresistenz* ungewöhnlich hoch liegen. Diese auffälligen *Differenzen* (Anzahl: 2 2 21) entstanden in Osterseeon, wo der Winter nicht so extrem ist. Auch zeigt in diesem Nutzungsjahr ein normaler Fusariumbefall bei einem normalen Winter besonders negative Auswirkungen. Bei einer Durchführung der Regression würde deshalb die schlechte *Fusariumresistenz* durch eine Variablenselektion als unbedeutend ausgesondert werden. Somit ist die Einzelbetrachtung des zweiten Nutzungsjahres nicht sinnvoll.

3.4.2 Untersuchung von Nutzungsjahr, Winter, Fusarium-Resistenz (mit Interaktionen)

Eine Interaktion wird als Produkt der kodierten Haupteffekte gebildet. Haupteffekte sind Merkmale ohne Berücksichtigung der Wechselbeziehungen.

Zu nachfolgendem Ergebnisausdruck, (ausführlicher S. 90), der durchgeführten Regression noch einige Bemerkungen:

Der Einfluss des zweiten Nutzungsjahres ist hierbei wieder nicht signifikant, deshalb wird keine Beurteilung über dieses abgegeben. Für die schlechte *Fusariumresistenz* wurden die beiden Wechselwirkungen normaler *Winter*, schlechte *Fusariumresistenz* und beim letzten Regressionskoeffizienten noch das zweite Nutzungsjahr hinzugezogen.

Die Odds sind am aussagekräftigsten, daher werden nur sie, mit Einbezug der Wechselwirkungen beurteilt. (Die Haupteffekte alleine wurden bereits im letzten Modell betrachtet, von den Wichtigkeiten her ist keine Änderung erkennbar.) Hierbei werden die betroffenen Odds miteinander multipliziert, um die Interpretation der Wechselbeziehungen zu ermöglichen.

variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.8850	0.5693	10.9632	0.0009	.	.
NUTZGS1	1	-0.5380	0.2564	4.4033	0.0359	-0.144213	0.584
NUTZGS2	1	0.1842	0.3749	0.2415	0.6231	0.046395	1.202
WINTER1	1	1.9884	0.6076	10.7091	0.0011	0.281236	7.304
WINTER2	1	0.8742	0.5057	2.9880	0.0839	0.171209	2.397
FUSMR2	1	1.1510	0.2361	23.7719	0.0001	0.312330	3.161
WIFU12	1	-1.7752	0.7701	5.3138	0.0212	-0.155023	0.169
WIFU21	1	3.7077	0.6331	34.3022	0.0001	0.703722	40.758
WIFUN212	1	-2.1339	0.8091	6.9563	0.0084	-0.279602	0.118

Das Nutzungsjahr eins mindert - wie vorher - die Wahrscheinlichkeit, dass die Pflanze einen schlechten Start im Frühjahr hat, um 71% bezüglich des dritten Jahres.

Ein strenger *Winter* ohne *Fusarium* besitzt um 630% mehr Einfluss auf das WD als ein milder *Winter*. Ist ein normaler Pilzbefall vorhanden, so ist der Einfluss auf den Frühjahrsbestand nur noch 23% größer als bei einem milden *Winter*.

Bei einem normalen *Winter* und sehr wenig Krankheitsresistenz ist die Gefahr eines schlechten Vegetationsstarts sehr groß ($2.397 \times 40.758=97.7$) gegenüber einem milden *Winter* und bei gesunder Pflanze. Falls dies im zweiten Nutzungsjahr passiert, ist das Risiko eines schlechten Frühjahrsstarts nicht mehr so hoch.

Eine gute *Fusariumresistenz* verringert die Aussicht auf einen schlechten Bestand nach Winter um den Faktor 3.161 gegenüber einer mittelmäßigen *Fusariumresistenz*.

Anschließend wurde noch eine Variablenselektion durchgeführt (Ergebnis 4.1, S. 93) Damit erreichte man ein Bestimmtheitsmaß von 0.16. Dies bedeutet immer noch eine durchaus akzeptable Annäherung. Die Größen der daraus resultierenden Regressionsanalyse sind anschließend aufgeführt.

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.2971	0.1762	54.1708	0.0001		
WINTER1	1	1.5847	0.4749	11.1370	0.0008	0.224141	4.878
FUSMR2	1	1.1964	0.2312	26.7756	0.0001	0.324649	3.308
WIFU12	1	-1.8206	0.7686	5.6105	0.0179	-0.158988	0.162
WIFU21	1	2.7634	0.3655	57.1502	0.0001	0.524500	15.854

Ein harter *Winter* hat demnach ein um den Faktor 4.878 größeres Gewicht auf einen schlechteren Vegetationsstart des WD als ein milder *Winter*. Gesellt sich noch ein normaler *Fusarium*befall hinzu, schmälert sich das Risiko.

Eine normale *Fusariumresistenz* besitzt eine um den Faktor 3.3 höhere Wahrscheinlichkeit einen schlechten Pflanzenbestand zu erzeugen als eine gute *Fusariumresistenz*.

Falls ein normaler *Winter* mit einem sehr starken *Fusarium*befall auftritt, so ist die Wahrscheinlichkeit um den Faktor 15.854 größer, einen schlechten Bestand zu bekommen als bei einem milden *Winter* ohne Krankheitsbefall.

3.4.3 Zusammenfassende Tendenz

Insgesamt kann man aus den bisherigen Analysen die nachstehenden Tendenzen für den Zustand der Pflanze nach Winter erkennen:

- Die Etablierung des Pflanzenbestandes und seine Entwicklung im ersten Nutzungsjahr haben einen positiven Effekt auf den Gesamtzustand nach Winter.
- Die Auswirkungen des zweiten Nutzungsjahres zeigen sich uneinheitlich.
- Der Einfluss einer schwachen Fusariumresistenz ist bereits bei einem normalen Winter verheerend.
- Falls nach einem harten Winter durchschnittlicher Fusariumbefall auftritt, ist er mit einem besseren Frühjahrsstart in diesem Nutzungsjahr verbunden. Dies ist so zu verstehen, dass bei einem harten Winter die Witterungsverhältnisse größeres Gewicht besitzen als die Krankheitsresistenz. Zum Beispiel geschieht es sehr häufig, dass trotz geringen Fusariumbefall - durch Frost, ein lückiger Pflanzenbestand im Frühjahr entsteht. Demgegenüber bieten vergleichsweise bessere und damit dichter aus dem Winter kommende Versuchspartzellen, der Krankheit eine größere Angriffsfläche als stark zurückgefrorene Partzellen. Dies erklärt, warum einzelne Weidelgrassorten trotz einer ungünstigen Fusariumbonitur, nach einem starken Winter einen besseren Vegetationsstart besitzen können als andere Versuchsglieder, die bereits mit einer geringen Blattmasse aus dem Winter kamen.
- Gute Fusariumresistenz und milder Winter bedeuten beste Ausgangsbedingungen.

Diese Aussagen treffen allerdings nur zu, wenn Fusarium auftritt, was, wie am Anfang des Abschnitts erwähnt, nur bei der Hälfte der Bonituren der Fall war, womit die Merkmale Winter und Nutzungsjahr an Bedeutung gewinnen.

Die Stichprobengröße lässt ferner nur tendenzielle Aussagen zu, wobei Odds-Ratios mit Werten unter drei, aufgrund möglicher Schätzfehler, besonders vorsichtig zu betrachten sind.

4 Resümee / Ausblick

Obwohl die untersuchte Stichprobengröße nur einen relativ kleinen Umfang hatte und keinerlei Kreuzvalidierungen, d. h. Überprüfung der Ergebnisse durch eine andere Stichprobe, durchgeführt wurden, kann man als Fazit dieser Diplomarbeit folgendes feststellen:

- Eine klare Unterscheidung der Nutzungsjahre war erkennbar mit dem dritten Nutzungsjahr als größtem Einflussfaktor, und dem zweiten Nutzungsjahr als kleinstem Einflussfaktor.
- Die Krankheiten besitzen nur einen sehr geringen Anteil an der Gesamtnote Winterhärte, im Gegensatz zu der Regenerationsfähigkeit.
- Extreme Boniturunterschiede der Orte benötigen eine Glättung.
- Mit den verwendeten Methoden ist die Erstellung einer Schätzgleichung grundsätzlich möglich, allerdings ist die Fehlerquote der Schätzung zu hoch als dass sie praktisch anwendbar wäre. Auch war die untersuchte Stichprobe dafür zu klein.

Für eine weitere Vorgehensweise bezüglich des Datenmaterials stehen die anknüpfenden Betrachtungen:

Es ist überlegenswert, die Einteilung der Noten zu reduzieren. Es gibt zwar sehr wohl eine Differenzierung zwischen einer Fünf und einer Sechs. Allerdings ist es nur interessant, in welchem Feld (gut, mittel, schlecht) eine Sorte liegt (was in der kategorialen Regression ja zum Ausdruck kam).

Bei der bisherigen Berechnung der Bayerischen Winterhärte könnte man die extremen Ortsunterschiede dadurch ausgleichen, dass man entweder die Orte einzeln betrachtet und erst am Ende zusammenfasst oder sich auf eine bestimmte Anzahl von Bonituren einigt, z. B. Verwendung der Schnitte nur bis zum dritten Schnitt, *Stand vor Winter* prinzipiell Anfang November usw..

Aus den bisherigen Resultaten konnte ich feststellen, dass der gesunde Menschenverstand oftmals durch die Mathematik bestätigt wird. Deshalb ist es erwägenswert, das vorhandene Datenmaterial noch per kategorialer Regression hinsichtlich Winter / Nutzungsjahr oder Rostresistenz / Regenerationsfähigkeit zu untersuchen. Dies würde den Rahmen dieser Diplomarbeit jedoch übersteigen.

Zu guter Letzt schließe ich mit einem Zitat⁴⁸ von Bertrand Russell welches man bei der Verwendung von statistischen Methoden immer bedenken sollte:

„Wie können wir nur von den Gesetzen der Wahrscheinlichkeit sprechen? Ist Wahrscheinlichkeit nicht die Antithese zu jeglichem Gesetz?“

⁴⁸ Die Zitate entstammen dem Buch Simon Singh: Fermats Letzter Satz, 1997, dtv, München

Anhang A: SAS-Programme / SAS-Ergebnisse

A 1 Einführung

SAS ist ein umfangreiches Software-Paket, welches eine Vielzahl von statistischen Analysen ermöglicht. Um einen Überblick über die von SAS bereitgestellten Prozeduren zu bekommen sei der OVERVIEW : Chapter 1 Regression Procedures bis Chapter 9 The Four Types of Estimable Functions in SAS USER'S GUIDE [14] empfohlen.

Für kategoriale Daten stehen zur Zeit PROC GENMOD, PROC CATMOD, PROC LOGISTIC und PROC PHREG zur Verfügung.

Für eine Regressionsanalyse sind unter anderen PROC GLM, PROC REG, PROC RSREG und PROC NLIN vorhanden.

Für eine Hauptkomponentenanalyse kann man PROC PRINCOMP, PROC FACTOR oder PROC CALIS verwenden.

Folgende Syntax ist zu beachten:

- Bei SAS-Programmen schreibt man Kommentare mit `/* kommentar */`
- Variablendeklarationen als Makrovariable `%let eing`
- Befehle werden mit Strichpunkt abgeschlossen.
- SAS unterscheidet nicht zwischen Groß- und Kleinschreibung.
- Ausgabedatei ist `data outdatei;` und Eingabedatei `set eingabedatei;`
- Pfadangaben werden nach dem ersten Aufruf von SAS eingegeben und stehen während des Arbeitens mit `pfad.Datei` zum Abrufen bzw. Speichern zur Verfügung. Falls nur die Datei aufgerufen wird, so geschieht die Speicherung in einem SAS-eigenen Arbeitsverzeichnis WORK. Nach Beendigung von SAS wird das Arbeitsverzeichnis wieder gelöscht; die Dateien darin sind also für die nächste Arbeitssitzung verloren.
- Bei arithmetischen Operationen müssen leere Zellen mit Null aufgefüllt werden, da sonst eine leere Zelle als Ergebnis entsteht.
- Bei einfachen Prozeduren zur Mittelwert-, Summenbildung usw., welche PROC UNIVARIATE oder PROC MEANS durchführen, rechnet das Programm

eine Null als eigenständigen Wert; möchte man diese also nicht mitberechnet haben, so sollte dort eine leere Zelle stehen.

- Bei der Mittelwertbildung über mehrere Merkmale benötigt man ein selbst geschriebenes Programm, in dem man also für den Zähler leere Zellen mit Null besetzen muss, damit die Summe berechnet wird und für den Nenner diese wieder als leere Zellen zurücksetzen, um die Anzahl fehlerlos zu erhalten.

Zu Programme A 2

Die Programme sind nach der Reihenfolge ihrer Anwendung geordnet.

Es gehören somit

- zur multiplen Regression : Programm 1 bis Programm 5
- zur Logitanalyse : Programm 6 bis Programm 10

Zu Ergebnisausdrucken A 3

- Grau unterlegte Werte sind Größen, welche für die weiteren Vorgehensweisen zu beachten sind.
- Parameter Estimate gibt die Regressionskoeffizienten an.
- p-value oder Prob > F oder Prob > T : Bezeichnet die Irrtumswahrscheinlichkeit den Fehler 1. Art zu begehen (Ablehnung einer gültigen Nullhypothese). Normalerweise sollte der Wert unter 0.05 oder 0.1 liegen, je nach verwendetem Signifikanzniveau; da man dann die Nullhypothese, wegen einer sehr geringen Wahrscheinlichkeit eine falsche Entscheidung zu treffen, ablehnt.
- Mit RSquare wird das Bestimmtheitsmaß bezeichnet.
- Die Ergebnisse werden bei der ersten Anwendung einer Methode etwas ausführlicher diskutiert, während bei einem späteren Einsatz die Kommentare spärlicher ausfallen.
- Bei den Residuenplots werden, soweit nichts anderes angegeben, die empirischen Daten (Bayrische Winterhärtenote oder Wahrscheinlichkeit) zu den Residuen aufgezeigt.

A 2 Programme

Programm 1 : Korrelationsberechnung nach Spearman, Pfad: fin, Eingangsdatei: orale3mi enthält die nach Zeitpunkt und über alle Orte gemittelten Werte.

```
proc corr data=fin.orale3mi spearman;
  var dvorwin wvorwin;
run;
```

Programm 2 : Makro zum Erzeugen des künstlichen Merkmals Bestand, der Aufruf erfolgt z. B. mit %best (datei,dichvw,wantvw,vw);

```
%macro best (d,m1,m2,anf);
  data &d; set &d;
    if &m1 ^=0 and &m2 ^=0 then hilf&anf=&m1*&m2;
                                else if &m1 ^=0 then best&anf=&m1;
                                else best&anf=&m2;

    if (hilf&anf >=1 and hilf&anf <= 2) then best&anf=1;
    if (hilf&anf >2 and hilf&anf <= 6) then best&anf=2;
    if (hilf&anf >6 and hilf&anf <= 12) then best&anf=3;
    if (hilf&anf >12 and hilf&anf <= 20) then best&anf=4;
    if (hilf&anf >20 and hilf&anf <= 30) then best&anf=5;
    if (hilf&anf >30 and hilf&anf <= 42) then best&anf=6;
    if (hilf&anf >42 and hilf&anf <= 56) then best&anf=7;
    if (hilf&anf >56 and hilf&anf <= 72) then best&anf=8;
    if (hilf&anf >72) then best&anf=9;
%mend best;
```

Programm 3 : Hier werden die Merkmale zeitpunktmäßig zusammengefasst. wobei jedes Mal ein Makro aufgerufen wird.

```
data &aus; set &eing;
  %zus(&aus, stvw0, bestvw0, vw0);
  %zus(&aus, stnw1, bestnw1, nw1);
  %zus(&aus, stnw2, bestnw2, nw2);
  %zus(&aus, stnw3, bestnw3, nw3);
  %zus(&aus, stnw4, bestnw4, nw4);

  %zuslet2 (&aus, stvw1, bestve1, bestvw1, vw1);
  %zuslet2 (&aus, stvw2, bestve2, bestvw2, vw2);
  %zuslet2 (&aus, stvw3, bestve3, bestvw3, vw3);
  %zuslet2 (&aus, 0, bestve4, bestvw4, vw4);

  %zusletz(&aus, bestvs31, bestvs41, bestvs51, bestvs61, vs31); run;
  %zusletz(&aus, bestvs32, bestvs42, bestvs52, bestvs62, vs32); run;
  %zusletz(&aus, bestvs33, bestvs43, bestvs53, bestvs63, vs33); run;
  %zusletz(&aus, bestvs34, bestvs44, bestvs54, bestvs64, vs34); run;
run;
```

Programm 4: Ist das verkürzte Programm zu Ergebnisausdruck 1.2: Hauptkomponentenanalyse mit anschließender Regression

```

%let pfad = finrtst;
%let datei = bestand;      /* eingangsdatei */

option linesize=80;      /* für den Ausdruck seitenbreiteneinstellung */

/*****
/***** Standardisierung *****/

data arbeit; set &pfad.&datei; run;

proc standard data=arbeit mean=0 std=1 out= prg;
  VAR  stnw1 stnw2 stnw3
       stvw1 stvw2 stvw3 stvw0
       bestve3 bestve2 bestve1 bestve0
       bestvw3 bestvw2 bestvw1 bestvw0
       bestnw3 bestnw2 bestnw1
       bestvs12 bestvs11
       bestvs63 bestvs62 bestvs61
       bestvs53 bestvs52 bestvs51
       bestvs43 bestvs42 bestvs41
       bestvs33 bestvs32 bestvs31
       bestvs23 bestvs22 bestvs21
       fusrmr1 fusrmr2 fusrmr3
       rostr1_1 rostr1_2 rostr1_3
       Baywihae;
run;

/*****
/***** Hauptkomponentenanalyse *****/

proc factor data=prg method=prin nfactors=4 out=factorsc outstat=pattern msa scree
           residual rotate=varimax;
  VAR  stnw1 stnw2 stnw3
       stvw1 stvw2          stvw0
       bestve3 bestve2 bestve1
       bestvw3 bestvw2 bestvw1 bestvw0
       bestnw3 bestnw2 bestnw1
       bestvs12 bestvs11
       bestvs63          bestvs61
       bestvs53 bestvs52 bestvs51
       bestvs43 bestvs42 bestvs41
       bestvs33 bestvs32 bestvs31
       bestvs23 bestvs22 bestvs21
       fusrmr1 fusrmr2 fusrmr3
       rostr1_1 rostr1_2 rostr1_3;
run;

```

Programm 4: Fortsetzung

```

/*****
/***** Multiple lineare Regression *****/

proc sort data = factorsc out=h;
  by factor1 factor2 factor3 factor4
  baywihae;
run;

proc reg data= h;
  model baywihae= factor1 factor3 factor4 / DW collin influence ;
  output out=c P=pred r=resid student=sres;
run;

/*****
/**** Faktorladungen in eine übersichtliche Form bringen *****/

data ladg; set pattern; where _type_= 'PATTERN'; run;

proc transpose data=ladg out=tranpatt;
  VAR stnw1 stnw2 stnw3
      stvw1 stvw2 stvw0
      bestve3 bestve2 bestve1
      bestvw3 bestvw2 bestvw1 bestvw0
      bestnw3 bestnw2 bestnw1
      bestvs12 bestvs11
      bestvs63 bestvs61
      bestvs53 bestvs52 bestvs51
      bestvs43 bestvs42 bestvs41
      bestvs33 bestvs32 bestvs31
      bestvs23 bestvs22 bestvs21
      fusrmr1 fusrmr2 fusrmr3
      rostr1_1 rostr1_2 rostr1_3;
run;

/*****
/***** Residuenplot: beobachtet:geschätzte werten *****/

symbol1 v="" l=33 i=NONE c=blue;

goptions hsize=20 cm vsize=15 cm;
PROC GPLOT data= c;
  PLOT baywihae*pred / HAXIS=AXIS2 VAXIS=AXIS1 autovref lvref=46
                    autohref lhref=33 cvref=grey chref=grey;
  TITLE1 H=1.7 F=Simplex 'Beobachtete gegeneüber geschaezter Zielvariable';
  AXIS1 LABEL= (F=Simplex R=0 A=90 'beobachtete Bayer. Winterhaertenoten ');
  AXIS2 LABEL= (F=Simplex H=1 color= black 'geschaezte Bayer. Winterhaertenoten');
RUN;

```

Programm 5: Berechnet die Quantile der Normalverteilung für die Residuen einer multiplen linearen Regression, sogenannte Rankits (siehe MYERS [9], S. 62).

```

proc sort data=c; by sres; run;          /* sres sind die Residuen */
data c; set c;
  i=_N_;
  h = (i-0.375)/(52+0.25);             /* 52 entspricht der Anzahl */
  Qu = probit(h);                       /* der Beobachtungen */
run;

```

Programm 6: Programm zur Erzeugung von der Differenz vor / nach Winter

```

data hilf; set bestand; keep Ort Versuch Nutzgsj Sorte Baywihae bestvw bestnw;
%nu (hilf);                               /* Hier wird das Macro %nu: Aufteilung der*/
run;                                       /* Merkmale nach Nutzungsjahre aufgerufen */

data diffwi; set diffwi;                  /* Datei diffwi wurde im Macro erzeugt */
diff1 = bestnw1-bestvw0;
diff2 = bestnw2-bestvw1;
diff3 = bestnw3-bestvw2;
diff4 = bestnw4-bestvw3; run;

data diffwi2; set diffwi; keep Ort Versuch Sorte Baywihae diff1 diff2 diff3 diff4; run;

proc sort data= diffwi2; by Ort Versuch Sorte Baywihae; run;
proc sort data= bestand; by Ort Versuch Sorte Baywihae; run;

data difbest;
merge bestand diffwi2; by Ort Versuch Sorte Baywihae; run;

data difbest; set difbest;
if Nutzgsj=1 then diffwi = diff1;
if Nutzgsj=2 then diffwi = diff2;
if Nutzgsj=3 then diffwi = diff3;
if Nutzgsj=4 then diffwi = diff4; run;

data ausgang; set difbest; keep Ort Versuch Ort Nutzgsj Jahr Sorte Baywihae
Krankhf bestvw bestnw diffwi; run;

```

Programm 7: Makro zur Einteilung der Merkmale in drei Kategorien

```

%macro einteil (d1);
data &aus; set &d1; keep Versuch Ort Nutzgsj Jahr Sorte Baywihae diffwi
Krankhf bestvw bestnw ; run;

data eintlg; set &aus;
reg = Bestvw-Bestnw; /* Herbst - Frühjahr*/
winter = Jahr - 1;

if winter in (1982 1987 1988 1989 1993 ) then Eintlgw=3;
if winter in (1984 1985 1986 1990 1992 1994 1996 1997 1998) then Eintlgw=2;
if winter in (1981 1983 1991 1995) then Eintlgw=1;

if Bestnw <= 3.5 then EintlgBn = 1; /* schlecht */
if Bestnw > 3.5 and Bestnw < 7 then EintlgBn = 2;
if Bestnw >= 7 then EintlgBn = 3; /* gut */

if Bestvw <= 3.5 then EintlgBv = 1; /* schlecht */
if Bestvw > 3.5 and Bestvw < 7 then EintlgBv = 2;
if Bestvw >= 7 then EintlgBv = 3; /* gut */

if Diffwi < -1 then Eintlgd = 1; /* negativ */
if Diffwi >= -1 and diffwi <2 then Eintlgd = 2; /* neutral */
if Diffwi >= 2 then Eintlgd = 3; /* positiv */

if Reg < -1 then Eintlgr = 1;
if reg >= -1 and reg <2 then Eintlgr = 2;
if reg >= 2 then Eintlgr = 3;

if Krankhf <= 3.5 then EintlgF = 1; /* schlecht */
if Krankhf > 3.5 and Krankhf <7 then EintlgF = 2; /* normal */
if Krankhf >= 7 then EintlgF = 3; /* gut */

if Baywihae in (1 2 3) then BWH=1; /* schlecht */
if Baywihae in (4 5 6) then BWH=2; /* normal */
if Baywihae in (7 8 9) then BWH=3; /* gut */

run;
%mend einteil;

```

Programm 8: Zusammenfassung der Kombinationen von nutzgsj, winter, fusrmr, &y

```
data ein; set &ein; hilf=1; run;                               /* &ein ist eine beliebige Datei */
proc means data=ein sum nway noprint;                         /* welche ganz am Anfang festgelegt */
  class nutzgsj winter fusrmr &y;                             /* wird */
  var hilf;
  output out=s2
  sum = anz;
run;
```

Programm 9: Dummykodierung zur Erstellung der Designmatrix (nach Verschönerung der Form der Datei s2 entstand finkat.diffwi2)

```
proc transreg data=finkat.diffwi2;
  model monotone (_1) = class(nutzgsj winter fusrmr) / intercept;
  output out=dummy;
  id ges;
run;
```

Programm 10: Durchführung der Logitanalyse mit Ausdruck der vorhergesagten Wahrscheinlichkeiten.(hlog wurde aus der datei dummy gebildet; zur Entfernung alles Nutzlosen)

```
proc logistic data=hlog descending;
  model _1/ges = nutzgs1 nutzgs2 winter1 winter2
              fusrmr1 fusrmr2 / waldcl walldr1 rsq;
  output out=probs predicted=prob xbeta=logit;
run;
```

A 3 Ausdrücke

Ergebnis 1: Multiple lineare Regression mit Kollinearitätsanalyse: Jedes Merkmal gemittelt und zu Bestand zusammengefasst (ohne 4. Jahr, nur vorm Schnitt berücksichtigt)

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	12	80.42053	6.70171	39.841	0.0001
Error	39	6.56024	0.16821		
C Total	51	86.98077			
Root MSE	0.41014	R-square	0.9246		
Dep Mean	5.51923	Adj R-sq	0.9014		
C.V.	7.43103				

Root MSE ist die Standardabweichung des Fehlers ($\sqrt{0.16821}$),
 Dep Mean der Mittelwert der Zielvariable,
 C.V. der Variationskoeffizient (Root MSE *100 / Dep Mean).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	-9.883556	1.35850429	-7.275	0.0001
STNW1	1	0.409731	0.11760581	3.484	0.0012
STNW3	1	-0.405618	0.09686392	-4.188	0.0002
BESTVW3	1	-0.562288	0.12931717	-4.348	0.0001
BESTVW2	1	-0.228563	0.11482330	-1.991	0.0536
BESTNW3	1	0.299377	0.09416620	3.179	0.0029
BESTVS11	1	0.413307	0.12461656	3.317	0.0020
BESTVS52	1	0.456585	0.14272060	3.199	0.0027
BESTVS43	1	0.308645	0.11862634	2.602	0.0130
BESTVS42	1	0.735269	0.16495007	4.458	0.0001
BESTVS32	1	0.384583	0.13028477	2.952	0.0053
ROSTR1_1	1	0.356976	0.08194750	4.356	0.0001
ROSTR1_3	1	-0.187688	0.06621093	-2.835	0.0072

Variable	DF	Squared Partial Corr Type II	Tolerance	Abkürzungsbezeichnungen
INTERCEP	1	.	.	Regressionskonstante
STNW1	1	0.23735498	0.02385451	Stand nach Winter im 1. Jahr
STNW3	1	0.31016432	0.06631130	Stand nach Winter im 3. Jahr
BESTVW3	1	0.32649759	0.02287672	Bestand vor Winter im 3. Jahr
BESTVW2	1	0.09222818	0.23601383	Bestand vor Winter im 2. Jahr
BESTNW3	1	0.20582456	0.31265750	Bestand nach Winter im 3. Jahr
BESTVS11	1	0.22000086	0.38415546	Bestand vor Schnitt 1 im 1. Jahr
BESTVS52	1	0.20787345	0.12019389	Bestand vor Schnitt 5 im 2. Jahr
BESTVS43	1	0.14790417	0.18904460	Bestand vor Schnitt 4 im 3. Jahr
BESTVS42	1	0.33751798	0.17455705	Bestand vor Schnitt 4 im 2. Jahr
BESTVS32	1	0.18262161	0.28298348	Bestand vor Schnitt 3 im 2. Jahr
ROSTR1_1	1	0.32730884	0.04511894	Rostresistenz im 1. Jahr
ROSTR1_3	1	0.17083900	0.05544385	Rostresistenz im 3. Jahr

Ergebnis 1: Kollinearitätsanalyse

Collinearity Diagnostics							
Number	Eigenvalue	Condition Index	Var Prop INTERCEP	Var Prop STNW1	Var Prop STNW3	Var Prop BESTVW3	Var Prop BESTVW2
1	10.83496	1.00000	0.0000	0.0001	0.0002	0.0001	0.0000
2	1.42630	2.75618	0.0000	0.0012	0.0032	0.0033	0.0000
3	0.59744	4.25860	0.0000	0.0030	0.0127	0.0044	0.0000
4	0.06437	12.97361	0.0008	0.0000	0.0970	0.0082	0.0034
5	0.02210	22.14130	0.0003	0.0672	0.0483	0.0349	0.0075
6	0.02071	22.87521	0.0023	0.1092	0.3626	0.0247	0.0002
7	0.00923	34.26016	0.0043	0.0821	0.1542	0.4066	0.0444
8	0.00765	37.63128	0.0095	0.0056	0.0062	0.0116	0.0388
9	0.00569	43.61971	0.0129	0.1237	0.1201	0.1136	0.3775
10	0.00453	48.90094	0.0003	0.1075	0.0020	0.0582	0.4244
11	0.00405	51.70387	0.0489	0.0307	0.0000	0.0446	0.0002
12	0.00178	77.97487	0.0010	0.2088	0.1665	0.1629	0.0923
13	0.00117	96.03872	0.9197	0.2608	0.0269	0.1271	0.0113

Number	Var Prop BESTNW3	Var Prop BESTVS11	Var Prop BESTVS52	Var Prop BESTVS43	Var Prop BESTVS42	Var Prop BESTVS32	Var Prop ROSTR1_1
1	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001
2	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000
3	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0064
4	0.0204	0.0004	0.0011	0.0033	0.0018	0.0041	0.0658
5	0.1649	0.0112	0.0002	0.0027	0.0093	0.0208	0.0001
6	0.2433	0.0228	0.0004	0.0030	0.0001	0.0000	0.0011
7	0.1890	0.0310	0.0030	0.0017	0.0063	0.0817	0.0228
8	0.1321	0.0325	0.0037	0.5422	0.0442	0.0000	0.0112
9	0.0235	0.0153	0.0049	0.0126	0.0188	0.2902	0.0355
10	0.0247	0.2366	0.0001	0.2748	0.0057	0.1423	0.0420
11	0.0875	0.0243	0.4408	0.0001	0.0679	0.1317	0.0158
12	0.1073	0.4817	0.0107	0.0795	0.8453	0.3206	0.1676
13	0.0068	0.1442	0.5349	0.0799	0.0006	0.0086	0.6317

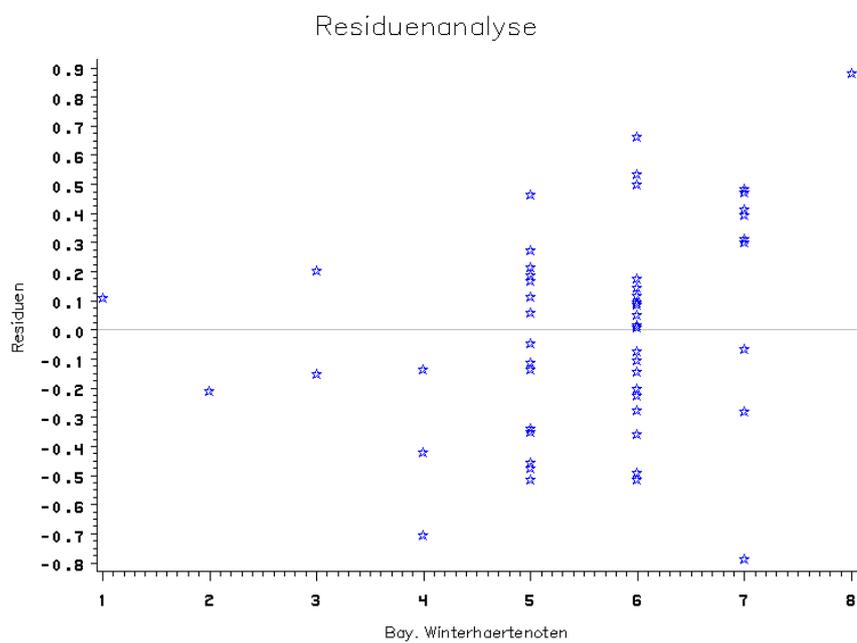
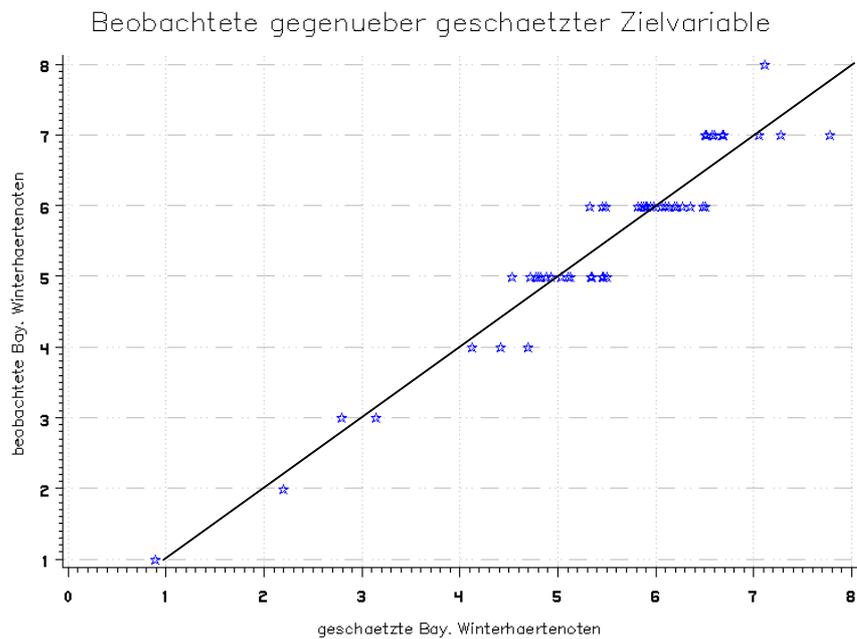
rostr1_3 wurde weggelassen, die werte davon liegen alle unter 0.5

Die unterlegten Werte bedeuten Multikollinearität, obwohl die Standardfehler der einzelnen Parameter noch im Toleranzbereich liegen.

Durbin-Watson D 2.051
 (For Number of Obs.) 52
 1st Order Autocorrelation -0.031

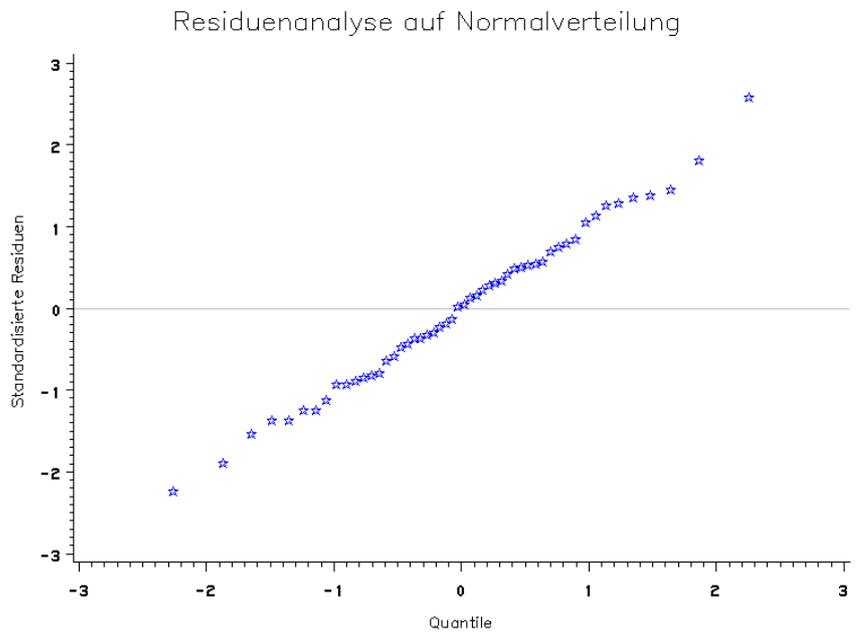
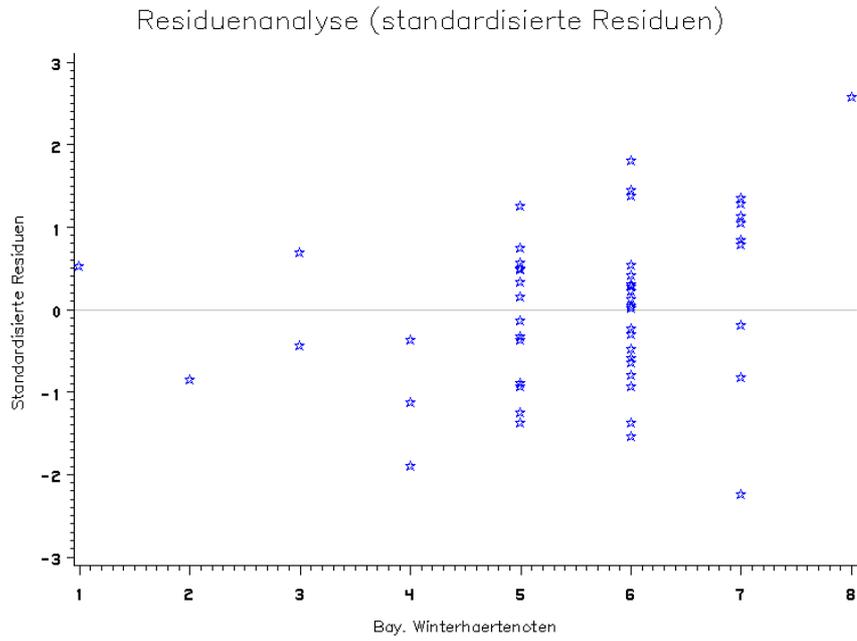
Der D-Wert von Durbin-Watson soll um 2 streuen (laut SCHUEMER [15], S. 101), sowohl zu niedrige als auch zu hohe Werte deuten auf Autokorrelation hin (besonders bei Zeitreihen ist dies zu beachten; PROC AUTOREG kann dies besser untersuchen). 1st order Autocorrelation gibt den Wert der Autokorrelation an.

Ergebnis 1: Residuenanalyse



Absoluter Fehler: 8 Schätzungen weichen von den beobachteten Noten mit mehr als 0.5 ab.

Ergebnis 1: Residuenanalyse



Die Residuen besitzen eine Normalverteilung.

Ergebnis 1.1: Hauptkomponentenanalyse: Jedes Merkmal gemittelt und zu Bestand zusammengefasst (ohne 4.Jahr , nur vorm Schnitt berücksichtigt)

Kaiser's Measure of Sampling Adequacy: Over-all MSA = 0.73853392

STNW1	STNW2	STNW3	STVW1	STVW2	STVW0	BESTVE3	BESTVE2
0.731633	0.845442	0.673894	0.895301	0.904513	0.782935	0.854411	0.767951
BESTVE1	BESTVW3	BESTVW2	BESTVW1	BESTVW0	BESTNW3	BESTNW2	BESTNW1
0.650694	0.700883	0.701193	0.772086	0.805684	0.766415	0.657670	0.555826
BESTVS13	BESTVS12	BESTVS11	BESTVS63	BESTVS61	BESTVS53	BESTVS52	BESTVS51
0.487395	0.816678	0.832117	0.649016	0.743530	0.567712	0.760601	0.770957
BESTVS43	BESTVS42	BESTVS41	BESTVS33	BESTVS32	BESTVS31	BESTVS23	BESTVS22
0.774308	0.639514	0.631741	0.806855	0.679451	0.706209	0.731811	0.734023
BESTVS21	FUSRMR1	FUSRMR2	FUSRMR3	ROSTR1_1	ROSTR1_2	ROSTR1_3	
0.713988	0.609103	0.591571	0.542325	0.821915	0.778975	0.896299	

Aufgrund der MSA wurde Bestvs13 eliminiert.

Nach der Varimaxrotation die erklärte Varianz
(Factor4 hatte vorher den Wert 1.4):

Variance explained by each factor			
FACTOR1	FACTOR2	FACTOR3	FACTOR4
13.186360	9.530567	6.602807	2.609601

Ergebnis 1.1: Nach der Varimaxrotation die Kommunalitäten-Schätzungen:

Final Commnality Estimates: Total = 31.929336

STNW1	STNW2	STNW3	STVW1	STVW2	STVW0	BESTVE3	BESTVE2
0.966621	0.982922	0.923411	0.985094	0.983522	0.965725	0.991646	0.978522
BESTVE1	BESTVW3	BESTVW2	BESTVW1	BESTVW0	BESTNW3	BESTNW2	BESTNW1
0.970691	0.976048	0.795018	0.839843	0.680053	0.709203	0.847585	0.750708
BESTVS12	BESTVS11	BESTVS63	BESTVS61	BESTVS53	BESTVS52	BESTVS51	BESTVS43
0.940584	0.798438	0.965149	0.980213	0.543937	0.887594	0.734927	0.865330
BESTVS42	BESTVS41	BESTVS33	BESTVS32	BESTVS31	BESTVS23	BESTVS22	BESTVS21
0.745875	0.812837	0.755085	0.757633	0.840792	0.771099	0.898031	0.827436
FUSRMR1	FUSRMR2	FUSRMR3	ROSTR1_1	ROSTR1_2	ROSTR1_3		
0.628117	0.533689	0.452175	0.955497	0.935856	0.952428		

Ergebnis 1.1: Die wichtigsten Faktorladungen (Rotated Factor Pattern)

OBS	BEZ1	FACTOR1	BEZ2	FACTOR2	BEZ3	FACTOR3	BEZ4	FACTOR4
1	ROSTR1_3	-0.88977	BESTVS61	-0.90049	FUSRMR2	-0.14740	FUSRMR1	-0.26982
2	BESTVS52	-0.78830	ROSTR1_2	-0.67242	ROSTR1_1	-0.11027	FUSRMR3	-0.20973
3	FUSRMR1	-0.70917	FUSRMR3	-0.59988	ROSTR1_2	-0.08321	BESTVS51	-0.11882
4	BESTVE1	-0.61095	ROSTR1_1	-0.57172	ROSTR1_3	-0.08181	BESTNW3	-0.07305
5	BESTVS63	-0.60881	BESTVS22	-0.41022	BESTVS61	-0.05231	ROSTR1_1	-0.06992
6	BESTVS31	-0.46004	ROSTR1_3	-0.39203	STNW3	0.01176	FUSRMR2	-0.04058
...								
24	BESTVS43	0.44369	BESTVE3	0.27937	BESTVW0	0.44492	BESTVS53	0.08548
25	BESTNW3	0.50502	STVW1	0.28113	BESTVS21	0.48620	BESTVS63	0.08778
26	BESTVW2	0.50865	STVW2	0.28113	BESTVS52	0.51265	BESTVS33	0.09433
27	BESTVW1	0.52780	BESTVS43	0.48253	BESTVS31	0.52998	BESTVS23	0.10521
28	BESTVS22	0.61649	BESTVS11	0.57346	BESTVS41	0.53134	BESTVS12	0.14712
29	ROSTR1_2	0.68998	BESTVW0	0.58693	BESTVS22	0.58538	BESTVS32	0.15550
30	BESTNW2	0.76973	BESTVS23	0.62298	BESTVS23	0.60807	BESTVW1	0.25058
31	ROSTR1_1	0.78204	FUSRMR2	0.62382	BESTVS43	0.65633	BESTVW2	0.26345
32	BESTVS51	0.80100	BESTNW1	0.67487	BESTVW2	0.65850	BESTVW0	0.37087
33	BESTVS12	0.87564	BESTVS63	0.76215	BESTNW3	0.66754	BESTVS42	0.41775
34	BESTVW3	0.93882	BESTVE1	0.76528	BESTVW1	0.70145	BESTNW1	0.47405
35	STNW2	0.94839	STVW0	0.87490	BESTVS42	0.70779	BESTVS31	0.56177
36	STVW2	0.94897	BESTVE2	0.89116	BESTVS53	0.71102	BESTVS41	0.60857
37	STVW1	0.94953	STNW1	0.90635	BESTVS33	0.71602	BESTVS11	0.65051
38	BESTVE3	0.95357	STNW3	0.91928	BESTVS32	0.80470	BESTVS21	0.76098

Interpretation der Ladungen:

Factor4: Vor Schnitt im ersten Nutzungsjahr

Factor3: Vor Schnitt (2..5) im 3.Nutzungsjahr + zweite Hälfte des 2.Nutzungsjahres
Ausreißer: Bestvw1

Factor2: Vor Winter 0..2.Jahr, Krankheiten im 2. Jahr,
nach Winter erstes und drittes Jahr, Ausreißer: bestvs63

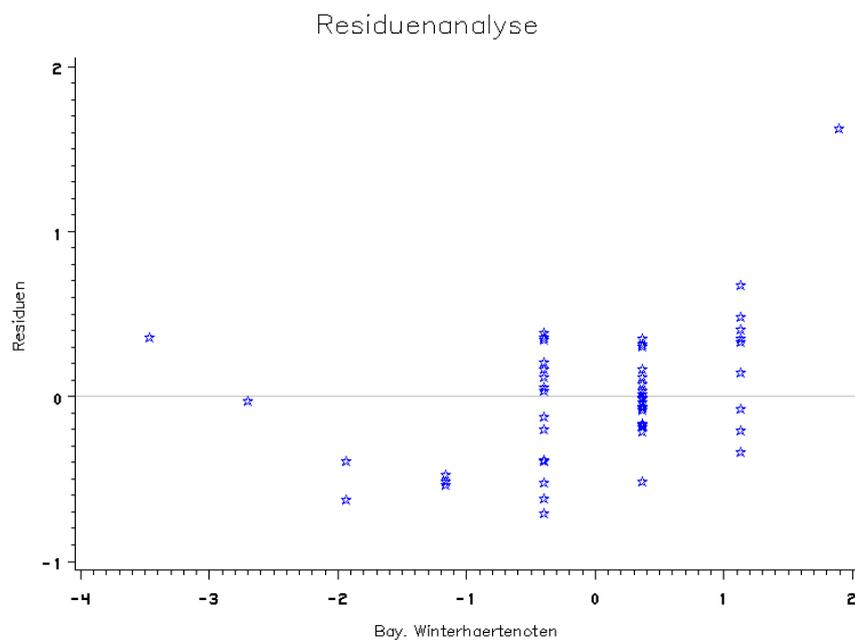
Factor1: Fusariumresistenz i. 1.Jahr, Rostresistenz alle drei Jahre, vor Winter alle drei Jahre
Erste Hälfte im 2.Jahr

Ergebnis 1.1: Nach der Varimaxrotation, Regression mit den rotierten Faktorwerten

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	42.85916	14.28639	84.235	0.0001
Error	48	8.14084	0.16960		
C Total	51	51.00000			
Root MSE		0.41183	R-square	0.8404	
Dep Mean		0.00000	Adj R-sq	0.8304	
C.V.		3.0138886E17			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	2.511012E-16	0.05711003	0.000	1.0000
FACTOR1	1	-0.142804	0.05766721	-2.476	0.0169
FACTOR3	1	0.844150	0.05766721	14.638	0.0001
FACTOR4	1	0.327709	0.05766721	5.683	0.0001



Absoluter Fehler: 9 Schätzungen weichen von den beobachteten Noten mit mehr als 0.5 ab.

Ergebnis 2: Multiple lineare Regression mit Kollinearitätsanalyse: Jedes Merkmal gemittelt und zu Bestand als auch nach Zeitpunkt zusammengefasst (ohne 4. Jahr, nur vorm Schnitt berücksichtigt)

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	6	77.18208	12.86368	59.076	0.0001
Error	45	9.79869	0.21775		
C Total	51	86.98077			
Root MSE		0.46664	R-square	0.8873	
Dep Mean		5.51923	Adj R-sq	0.8723	
C.V.		8.45472			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	-8.876904	1.03851116	-8.548	0.0001
ZUSBVW2	1	-0.330114	0.12286735	-2.687	0.0101
BESTVS11	1	0.604131	0.13118812	4.605	0.0001
ZUSBVS32	1	1.270591	0.14108137	9.006	0.0001
BESTVS22	1	0.590952	0.14026083	4.213	0.0001
ROSTR1_1	1	0.206800	0.08567517	2.414	0.0199
ROSTR1_2	1	-0.161493	0.08034289	-2.010	0.0504

Variable	DF	Squared Partial Corr Type II	Tolerance	Abkürzungsbezeichnungen
INTERCEP	1	.	.	Regressionskonstante
ZUSBVW2	1	0.13823879	0.39881713	Zusammenf. Bestand v. Winter j.2.J.
BESTVS11	1	0.32031039	0.44871424	Bestand vor Schnitt1 j.1.Jahr
ZUSBVS32	1	0.64316737	0.33103851	Zus. Bestand.vor Schnitt3 j.2.Jahr
BESTVS22	1	0.28288379	0.20488211	Bestand vor Schnitt2 j.2.Jahr
ROSTR1_1	1	0.11463160	0.05343437	Rostresistenz im 1.Jahr
ROSTR1_2	1	0.08238752	0.06519323	Rostresistenz im 2.Jahr

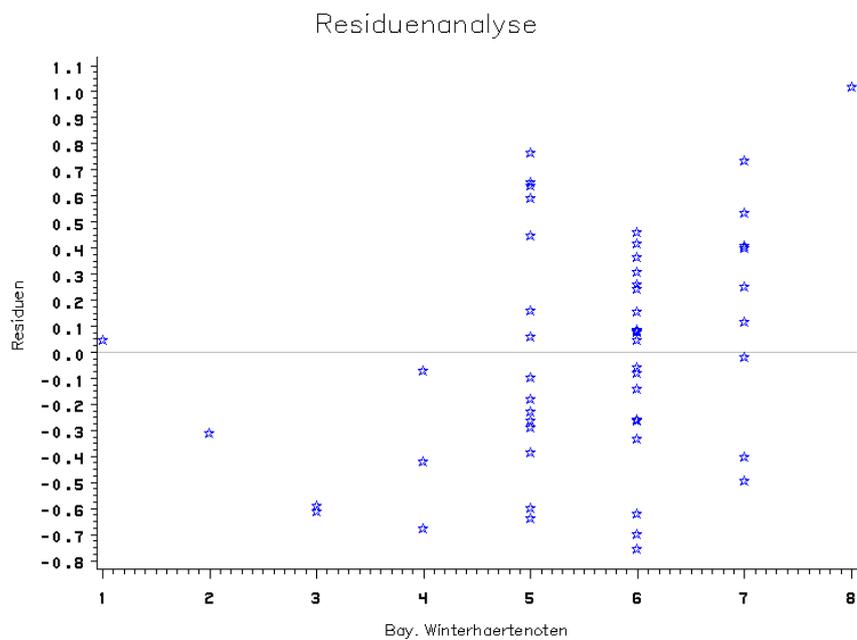
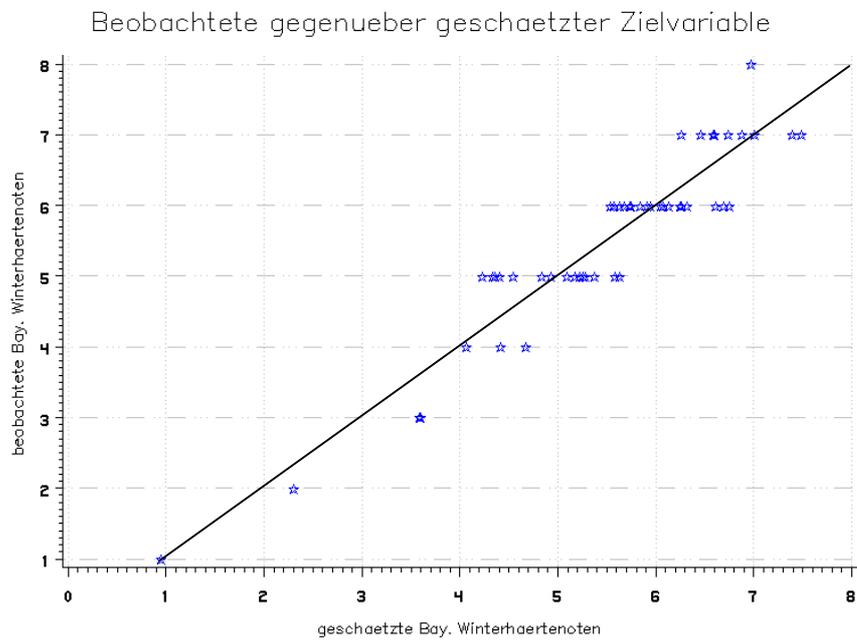
Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Var Prop INTERCEP	Var Prop ZUSBVW2	Var Prop BESTVS11	Var Prop ZUSBVS32	Var Prop BESTVS22
1	6.60665	1.00000	0.0001	0.0001	0.0001	0.0001	0.0001
2	0.35375	4.32157	0.0005	0.0008	0.0012	0.0016	0.0000
3	0.01664	19.92567	0.0001	0.1113	0.0070	0.0586	0.0203
4	0.01067	24.88652	0.0526	0.0341	0.1137	0.0153	0.1969
5	0.00768	29.32160	0.1116	0.1704	0.0412	0.0791	0.0145
6	0.00257	50.68818	0.0064	0.5599	0.3169	0.4117	0.6912
7	0.00203	56.98856	0.8287	0.1234	0.5199	0.4336	0.0770

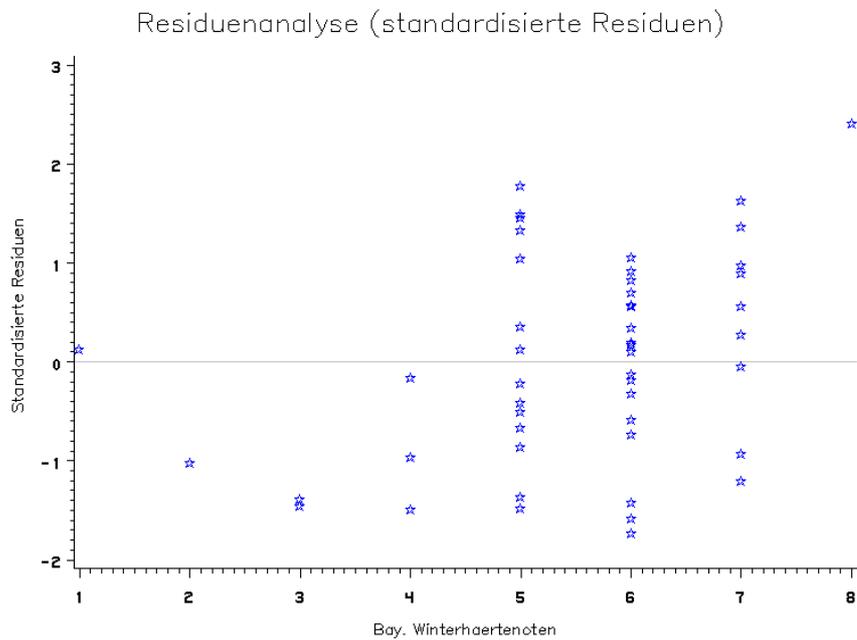
Rostr1_1 und Rostr1_2 wurden weggelassen, da sie werte unter 0.5 besitzen.

Durbin-watson D 1.767
 (For Number of Obs.) 52
 1st order Autocorrelation 0.100

Ergebnis 2: Residuenanalyse



Absoluter Fehler: 15 Schätzungen weichen von den beobachteten Noten mit mehr als 0.5 ab.

Ergebnis 2: Residuenanalyse

Ergebnis 2.1: Hauptkomponentenanalyse: Für jedes Merkmal gemittelt und zu Bestand als auch nach Zeitpunkt zusammengefasst (ohne 4.Jahr, nur vorm Schnitt berücksichtigt)

Kaiser's Measure of Sampling Adequacy: Over-all MSA = 0.80590809

ZUSBVW3	ZUSBVW2	ZUSBVW1	ZUSBVW0	ZUSBNW3	ZUSBNW2	ZUSBNW1	BESTVS13
0.791034	0.916276	0.887225	0.780603	0.703206	0.880228	0.691802	0.693073
BESTVS12	BESTVS11	ZUSBVS33	ZUSBVS32	ZUSBVS31	BESTVS23	BESTVS22	BESTVS21
0.811059	0.868956	0.826969	0.762555	0.810015	0.762066	0.858160	0.755295
FUSRMR1	FUSRMR2	FUSRMR3	ROSTR1_1	ROSTR1_2	ROSTR1_3		
0.874344	0.762002	0.708247	0.788704	0.792348	0.777242		

Nach der Varimaxrotation erklärte Varianz

(vorher waren Factor3: 2.757535 und Factor4: 1.152732):

variance explained by each factor

FACTOR1	FACTOR2	FACTOR3	FACTOR4
5.035128	4.858826	4.227517	2.912466

Ergebnis 2.1: Nach der Varimaxrotation die Kommunalitäten-Schätzungen und wichtigsten Faktorladungen

Final Commuality Estimates: Total = 17.033936

ZUSBVW3	ZUSBVW2	ZUSBVW1	ZUSBVW0	ZUSBNW3	ZUSBNW2	ZUSBNW1	BESTVS13
0.950483	0.817143	0.796402	0.466241	0.619979	0.843711	0.560314	0.681137
BESTVS12	BESTVS11	ZUSBVS33	ZUSBVS32	ZUSBVS31	BESTVS23	BESTVS22	BESTVS21
0.924947	0.832993	0.848138	0.825931	0.773973	0.787153	0.898923	0.846886
FUSRMR1	FUSRMR2	FUSRMR3	ROSTR1_1	ROSTR1_2	ROSTR1_3		
0.639907	0.624732	0.486060	0.954045	0.902699	0.952139		

Rotated Factor Pattern

OBS	BEZ1	FACTOR1	BEZ2	FACTOR2	BEZ3	FACTOR3	BEZ4	FACTOR4
1	ROSTR1_1	-0.24692	FUSRMR2	-0.78426	ZUSBVW3	-0.86654	ROSTR1_3	-0.31556
2	FUSRMR3	-0.24112	FUSRMR1	-0.39279	BESTVS12	-0.54963	ROSTR1_2	-0.09226
3	ROSTR1_3	-0.22303	BESTVS11	-0.39099	ZUSBVW2	-0.54640	BESTVS11	-0.06691
4	ROSTR1_2	-0.17860	BESTVS23	-0.33180	ROSTR1_1	-0.49169	ROSTR1_1	-0.01993
...								
13	ZUSBVS32	0.47796	ZUSBNW3	0.21909	BESTVS11	-0.05166	ZUSBNW2	0.33586
14	BESTVS23	0.58308	ZUSBVW3	0.25911	FUSRMR2	-0.03922	BESTVS12	0.33871
15	ZUSBVW2	0.59423	ZUSBVW1	0.32043	ZUSBNW3	0.07025	ZUSBVW1	0.37707
16	ZUSBVS33	0.60895	ZUSBVS31	0.37600	ZUSBVW1	0.07465	ZUSBVW2	0.39969
17	ZUSBVW0	0.61351	FUSRMR3	0.57460	BESTVS21	0.19787	ZUSBVS32	0.40058
18	ZUSBVS31	0.62440	BESTVS12	0.66498	FUSRMR3	0.28347	ZUSBVS31	0.44758
19	ZUSBNW1	0.71103	ZUSBNW2	0.77394	FUSRMR1	0.65014	BESTVS22	0.45141
20	ZUSBVW1	0.73890	BESTVS22	0.78130	ZUSBVS32	0.65568	BESTVS23	0.57783
21	BESTVS11	0.82035	ROSTR1_1	0.80680	BESTVS13	0.70335	ZUSBVS33	0.68201
22	BESTVS21	0.88469	ROSTR1_2	0.85074	ROSTR1_3	0.88766	ZUSBNW3	0.75168

Interpretation der Ladungen:

Factor4: Korrekturfaktoren für zusbvs33 und zusbnw3 (factor1, factor3)

Factor3: Fusariumresistenz im ersten Nutzungsjahr,
Rostresistenz im dritten Nutzungsjahr,
Letzte Schnitt im zweiten Nutzungsjahr,
Vor Winter im dritten Nutzungsjahr, bestvs13 siehe factor4 ...

Factor2: Rostresistenz im. ersten Nutzungsjahr,
Krankheiten im zweiten Nutzungsjahr,
Erste Hälfte des zweiten Nutzungsjahres

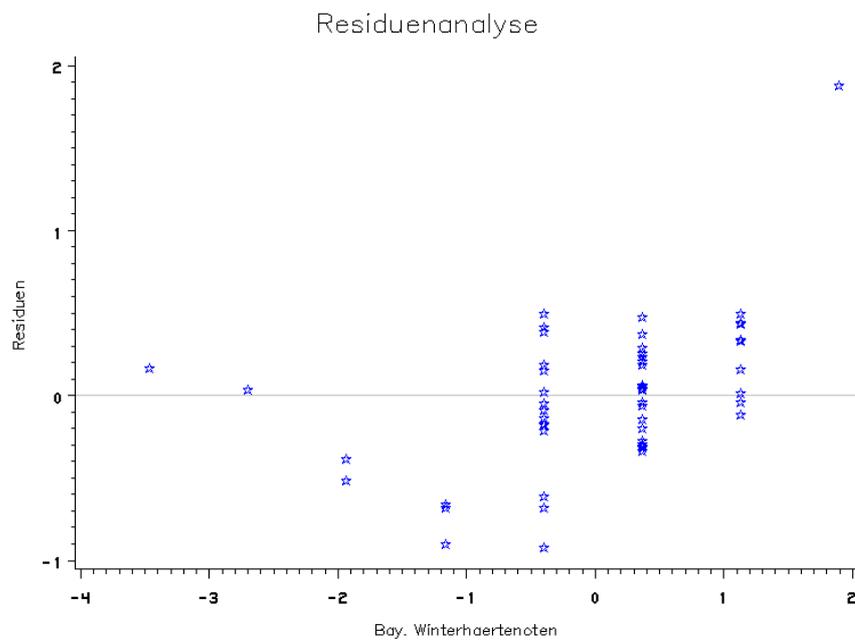
Factor1: vor Winter im nullten Nutzungsjahr,
erste Nutzungsjahr,
zusbnw3 siehe factor4

Ergebnis 2.1: Nach der Varimaxrotation, Regression mit den rotierten Faktorwerten

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	40.78358	10.19589	46.906	0.0001
Error	47	10.21642	0.21737		
C Total	51	51.00000			

Root MSE	0.46623	R-square	0.7997
Dep Mean	0.00000	Adj R-sq	0.7826
C.V.	1.2131683E17		

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	5.052034E-16	0.06465450	0.000	1.0000
FACTOR1	1	0.586461	0.06528529	8.983	0.0001
FACTOR2	1	0.245185	0.06528529	3.756	0.0005
FACTOR3	1	0.437671	0.06528529	6.704	0.0001
FACTOR4	1	0.451741	0.06528529	6.919	0.0001



Absoluter Fehler: 8 Schätzungen weichen von den beobachteten Noten mit mehr als 0.5 ab.

Ergebnis 3: Logitanalyse für Nutzungsjahr, Winter, Fusarium (Haupteffekte) mit Zielvariable Differenz über Winter

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept only	Intercept and Covariates	Chi-Square for Covariates
AIC	636.645	578.458	.
SC	640.792	607.482	.
-2 LOG L Score	634.645	564.458	70.187 with 6 DF (p=0.0001) 67.281 with 6 DF (p=0.0001)
RSquare = 0.1395		Max-rescaled RSquare = 0.1878	

Das RSquare mit 0.14 liegt unter zwei. Der -2 LOG L-Test lehnt mit einer Fehlerwahrscheinlichkeit von 0.0001 die Nullhypothese, dass alle Schätzparameter gleich Null sind, ab - es handelt sich hier deshalb um ein grundsätzlich annehmbares Modell.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.8875	0.5343	12.4800	0.0004	.	.
NUTZGS1	1	-0.5214	0.2428	4.6115	0.0318	-0.139765	0.594
NUTZGS2	1	-0.1621	0.3260	0.2472	0.6191	-0.040810	0.850
WINTER1	1	1.5806	0.5444	8.4290	0.0037	0.223555	4.858
WINTER2	1	1.0611	0.4743	5.0045	0.0253	0.207813	2.889
FUSRMR1	1	2.1376	0.3261	42.9813	0.0001	0.430831	8.479
FUSRMR2	1	0.9969	0.2246	19.6954	0.0001	0.270502	2.710

Das zweite Nutzungsjahr ist nicht signifikant! Bei einer Variablenselektion würde es entfernt werden.

Parameter Estimates and 95% Confidence Intervals

Variable	Parameter Estimate	wald Confidence Limits	
		Lower	Upper
INTERCPT	-1.8875	-2.9346	-0.8403
NUTZGS1	-0.5214	-0.9973	-0.0455
NUTZGS2	-0.1621	-0.8009	0.4768
WINTER1	1.5806	0.5136	2.6476
WINTER2	1.0611	0.1314	1.9907
FUSRMR1	2.1376	1.4986	2.7767
FUSRMR2	0.9969	0.5566	1.4371

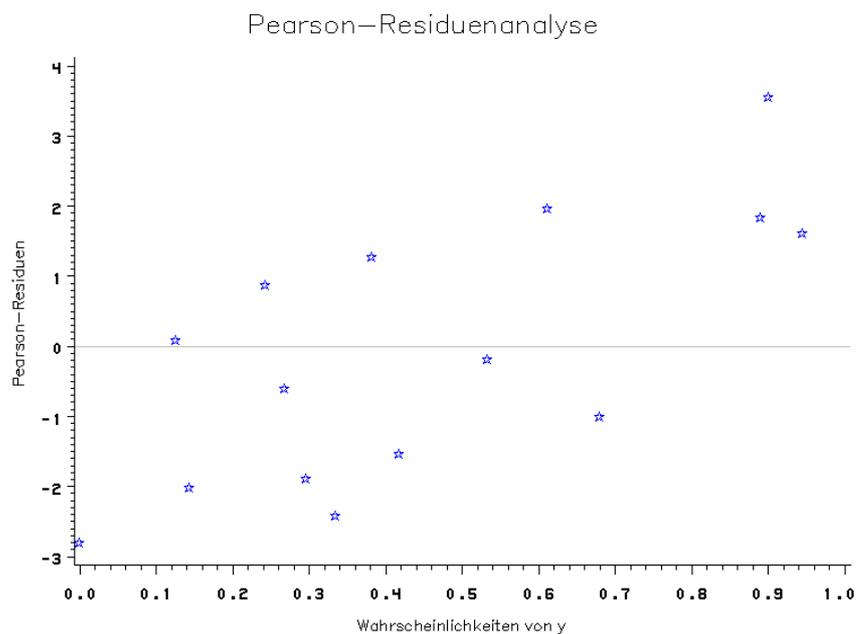
Das zweite Nutzungsjahr hat sowohl ein negatives als auch positives Vorzeichen, was eine Interpretation nicht zulässt.

Für die Odds-Ratios folgt dann $\text{Lower-Odds-Ratio} = \exp(\text{Lower}_{\text{Parameter}})$ usw. . Daraus ergibt sich die nachfolgende Übersicht (S. 89). Dort erkennt man, dass die Schwankungsbreite sehr groß ist. So kann ein strenger Winter um den Faktor 1.671 oder 14.121 mehr Einfluss als ein milder Winter auf den schlechtesten Zustand im Frühjahr haben. Jedoch bleibt das Größenverhältnis der Ausprägungen bestehen (Ein strenger Winter hat immer mehr Auswirkungen auf die Pflanze als ein normaler).

Ergebnis 3: Fortsetzung

Variable	Unit	O.-Ratio	Lower	Upper
NUTZGS1	1.0000	0.594	0.369	0.956
NUTZGS2	1.0000	0.850	0.449	1.611
WINTER1	1.0000	4.858	1.671	14.121
WINTER2	1.0000	2.889	1.140	7.321
FUSRMR1	1.0000	8.479	4.475	16.066
FUSRMR2	1.0000	2.710	1.745	4.209

Residuenanalyse:



Es sind 4 Werte mit einem Pearson-Residuum über zwei, von denen alle eine Häufigkeit von eins bzw. eine Wahrscheinlichkeit über 90%, besitzen, was die Aussagekraft relativiert, da man diese für eine gute Teststatistik ausschließen sollte. Allerdings würde sich dadurch auch keine bessere Anpassung ergeben (das RSquare wäre 0.12 und es wäre eine Verzerrung hin zum normalen Winter als auch zum zweiten Nutzungsjahr).

Ergebnis 3: Fortsetzung

Association of Predicted Probabilities and Observed Responses

Concordant = 66.3%	Somers' D = 0.423
Discordant = 24.0%	Gamma = 0.468
Tied = 9.7%	Tau-a = 0.206
(53040 pairs)	c = 0.711

Diese Statistiken gehen von der insgesamt möglichen Anzahl von Kombinationen aus. Es gibt also $195 \times 272 = 53040$ Paare. Davon werden 66.3% überschätzt, 24.0% unterschätzt und 9.7% weder unter - noch überschätzt. Dafür liegt die Toleranzgrenze bei 0.002.

somers'D ist der prozentuale Anteil der Differenz von Überschätzten zu Unterschätzten. Für Gamma, Tau-a und c siehe SAS USER'S GUIDE [12], S. 1091.

Sie beruhen alle auf die Beziehungen von Überschätzten, Unterschätzten und Tieds und sollten möglichst hohe Werte besitzen.

Die vorhandenen Werte bestätigen die Residuenanalyse, da die Regression zu ca. 50% als richtig angesehen werden kann.

Ergebnis 4: Logitanalyse (Haupteffekte mit Interaktionen) Zielvariable: Differenz über Winter

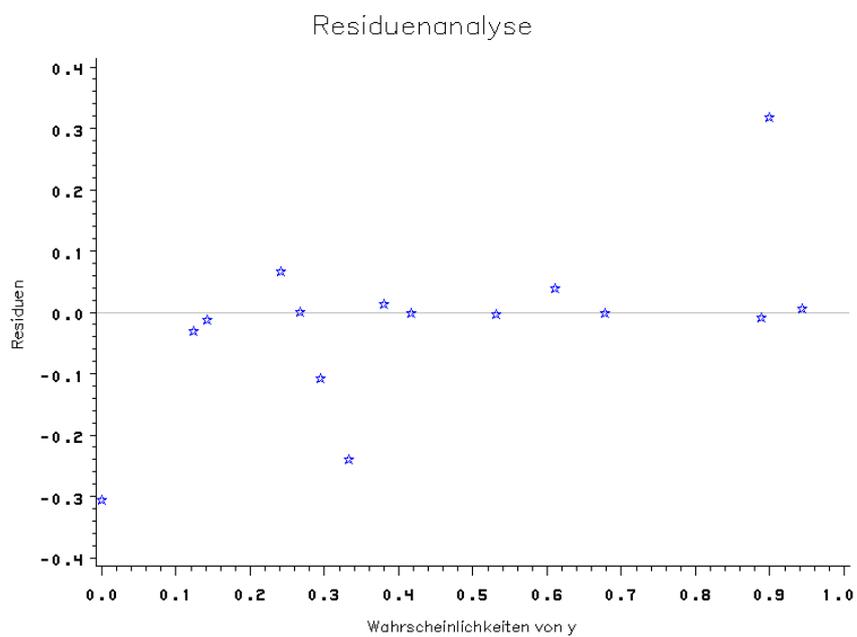
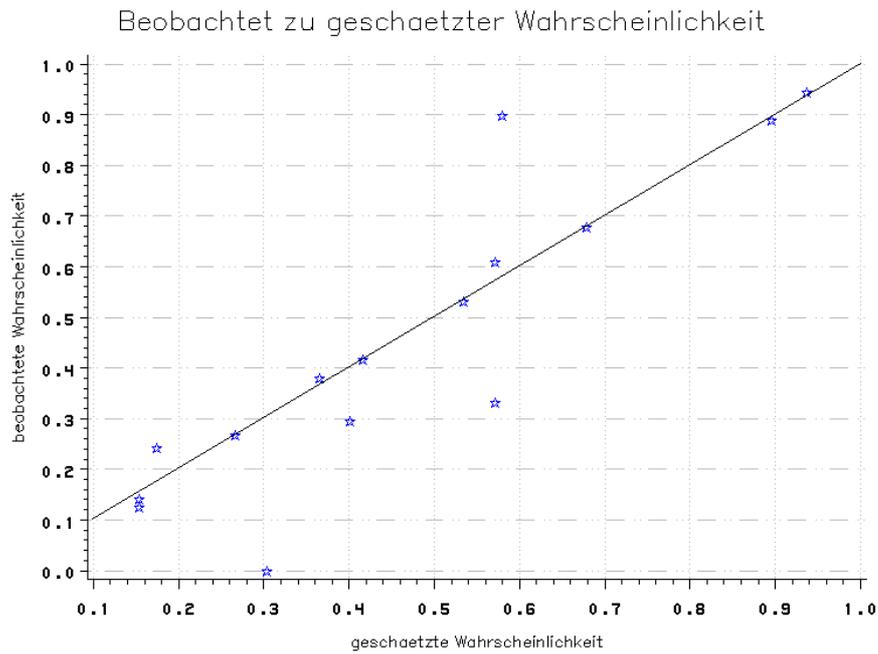
Criterion	Intercept Only	and Covariates	Chi-Square for Covariates
AIC	636.645	557.543	.
SC	640.792	594.860	.
-2 LOG L Score	634.645	539.543	95.103 with 8 DF (p=0.0001) 87.799 with 8 DF (p=0.0001)
		RSquare = 0.1842	Max-rescaled RSquare = 0.2480

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.8850	0.5693	10.9632	0.0009	.	.
NUTZGS1	1	-0.5380	0.2564	4.4033	0.0359	-0.144213	0.584
NUTZGS2	1	0.1842	0.3749	0.2415	0.6231	0.046395	1.202
WINTER1	1	1.9884	0.6076	10.7091	0.0011	0.281236	7.304
WINTER2	1	0.8742	0.5057	2.9880	0.0839	0.171209	2.397
FUSRMR2	1	1.1510	0.2361	23.7719	0.0001	0.312330	3.161
WIFU12	1	-1.7752	0.7701	5.3138	0.0212	-0.155023	0.169
WIFU21	1	3.7077	0.6331	34.3022	0.0001	0.703722	40.758
WIFUN212	1	-2.1339	0.8091	6.9563	0.0084	-0.279602	0.118

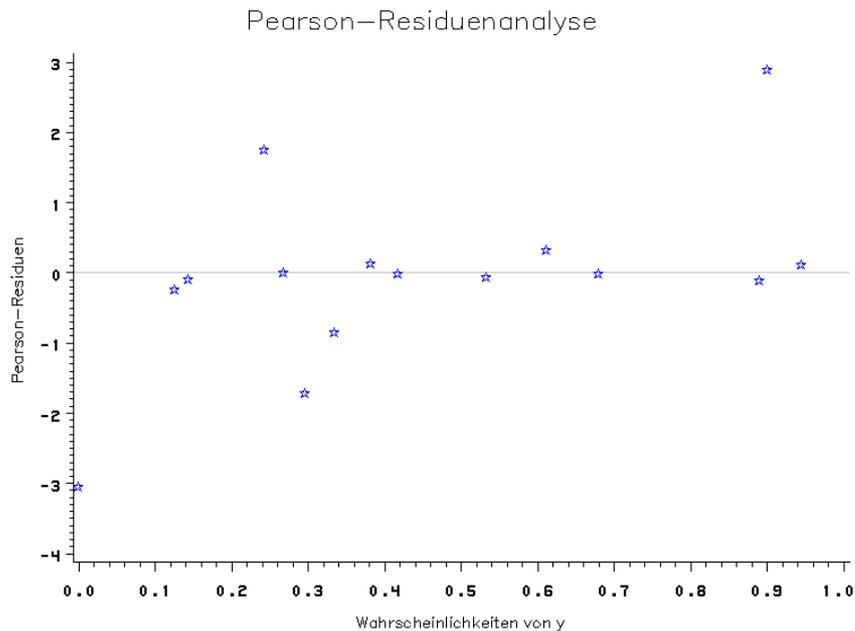
Association of Predicted Probabilities and Observed Responses

Concordant = 69.8%	Somers' D = 0.494
Discordant = 20.4%	Gamma = 0.547
Tied = 9.8%	Tau-a = 0.241
(53040 pairs)	c = 0.747

Ergebnis 4: Fortsetzung

Drei werte schätzen die wahrscheinlichkeit mit über 20% falsch ein.

Ergebnis 4: Fortsetzung



Zwei Werte sind größer Betrag zwei!

Konfidenzintervalle:

Variable	Parameter Estimate	95% Confidence Intervals wald	
		Lower	Upper
INTERCPT	-1.8850	-3.0008	-0.7692
NUTZGS1	-0.5380	-1.0405	-0.0355
NUTZGS2	0.1842	-0.5505	0.9190
WINTER1	1.9884	0.7975	3.1793
WINTER2	0.8742	-0.1170	1.8654
FUSRMR2	1.1510	0.6883	1.6137
WIFU12	-1.7752	-3.2845	-0.2658
WIFU21	3.7077	2.4669	4.9484
WIFUN212	-2.1339	-3.7196	-0.5481

Bei Winter2 liegt die untere Grenze ziemlich nahe an Null, so dass man die Parameterschätzung als auslegbar ansehen kann.

Variable	Unit	Odds Ratio	95% Confidence Intervals wald	
			Lower	Upper
NUTZGS1	1.0000	0.584	0.353	0.965
NUTZGS2	1.0000	1.202	0.577	2.507
WINTER1	1.0000	7.304	2.220	24.031
WINTER2	1.0000	2.397	0.890	6.458
FUSRMR2	1.0000	3.161	1.990	5.021
WIFU12	1.0000	0.169	0.037	0.767
WIFU21	1.0000	40.758	11.786	140.951
WIFUN212	1.0000	0.118	0.024	0.578

Ergebnis 4.1: Logitanalyse (Haupteffekte mit Interaktionen) nach der Variablenselektion

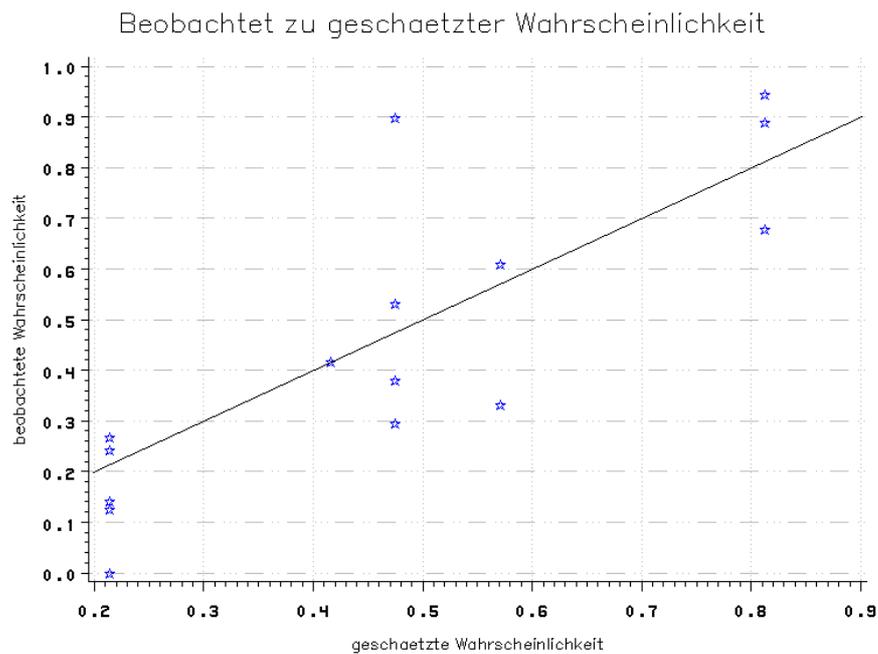
Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	636.645	563.112	.
SC	640.792	583.843	.
-2 LOG L Score	634.645	553.112	81.534 with 4 DF (p=0.0001)
	.	.	77.839 with 4 DF (p=0.0001)
	RSquare = 0.1602		Max-rescaled RSquare = 0.2156

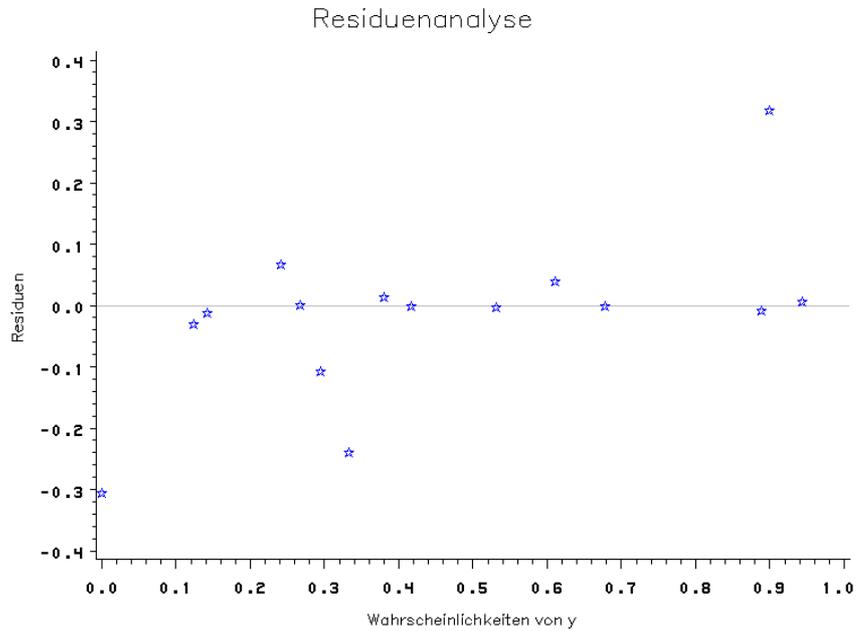
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.2971	0.1762	54.1708	0.0001	.	.
WINTER1	1	1.5847	0.4749	11.1370	0.0008	0.224141	4.878
FUSRMR2	1	1.1964	0.2312	26.7756	0.0001	0.324649	3.308
WIFU12	1	-1.8206	0.7686	5.6105	0.0179	-0.158988	0.162
WIFU21	1	2.7634	0.3655	57.1502	0.0001	0.524500	15.854

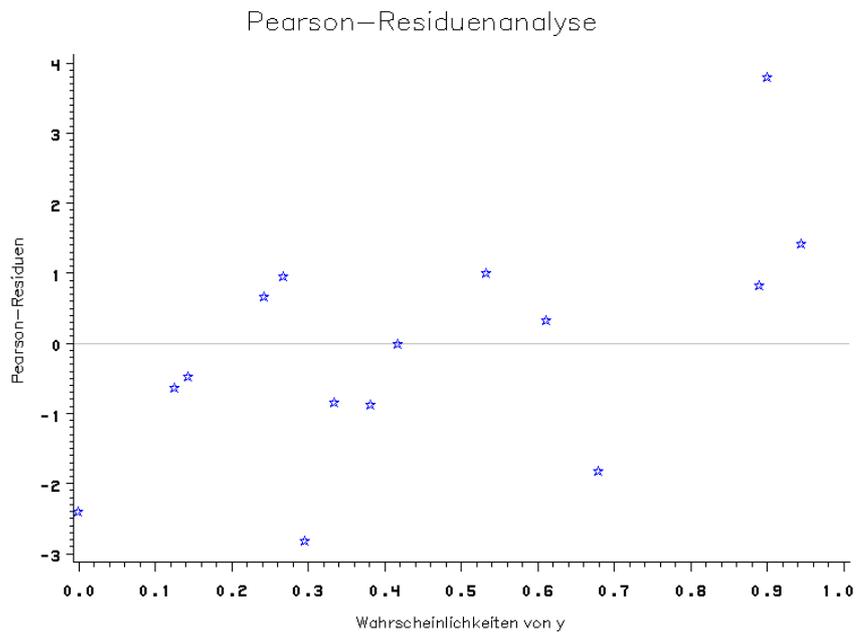
Association of Predicted Probabilities and Observed Responses

Concordant = 57.7%	Somers' D = 0.436
Discordant = 14.2%	Gamma = 0.606
Tied = 28.1%	Tau-a = 0.212
(53040 pairs)	c = 0.718

Ergebnis 4.1: Residuenanalyse:

Ergebnis 4.1: Residuenanalyse:

Drei Werte besitzen über 20% Differenz zu den beobachteten Werten.



Drei Werte sind größer Betrag zwei.

Anhang B: Mathematische Grundlagen

Die folgenden Sätze und Definitionen sind der Lektüre FAHRMAIR [6], KREYSZIG [8] und meinen Vorlesungen entnommen. Die Beweise dazu werden weggelassen. Für Abschnitt B2 war das Buch von FAHRMAIR [5] bestimmend.

B 1 Definitionen / Sätze

Def.: Seien x_1, \dots, x_N unabhängige Beobachtungen mit Dichten $f_n(x_n | \theta)$, so

folgt: Ist $\hat{\theta} = g(x_1, \dots, x_N) = g(X)$ ein Schätzer für θ , so heißt

$$E_{\theta}(\hat{\theta}) - \theta = v(\theta)$$

Bias. Für $v(\theta) = 0$ heißt g erwartungstreu.

Def.: Für jede feste positive Zahl ε , mit $\varepsilon > 0$ strebt die Wahrscheinlichkeit, dass sich die Näherung für u um weniger als ε von dem wahren Parameterwert u unterscheidet, mit wachsendem Stichprobenumfang gegen 1. Eine Schätzfunktion $g(x_1, \dots, x_n)$ mit dieser Eigenschaft heißt eine konsistente Schätzfunktion.

Def.: Sei $\frac{\partial^2 l(\theta; X)}{\partial \theta \partial \theta'}$ die Matrix der zweiten Ableitungen der log-Likelihood-

Funktion, so heißt

$$F_N(\theta) = E\left(-\frac{\partial^2 l(\theta; X)}{\partial \theta \partial \theta'}\right)$$

Fisher'sche Informationsmatrix der Stichprobe x_1, \dots, x_N . Sie ist in der Regel positiv definit.

Gehört die Dichte der Stichprobe x_1, \dots, x_N einer Exponentialfamilie an,

so ist

$$F_N(\theta) = \text{cov}(t^{(N)}(X))$$

Def.: Das Maximum-Likelihood-Prinzip besteht darin einen Parameterschätzwert $\theta \in \Theta$ so zu wählen, dass für diesen Schätzwert der Beobachtung eine maximale Wahrscheinlichkeit zukommt.

Zur Bestimmung der lokalen Maxima wird die erste Ableitung der log-Likelihood-Funktion gleich null gesetzt (mit zweiter Ableitung negativ definit).

Seien x_1, \dots, x_N die Stichprobenwerte einer N-fachen Beobachtungsreihe mit gemeinsamer Dichte $f(x_1, \dots, x_N | \theta)$, so heißt

$$L(\theta; x_1, \dots, x_N) = \prod_{n=1}^N f_n(x_n / \theta)$$

Likelihood-Funktion und

$$l(\theta; x_1, \dots, x_N) = \sum_{n=1}^N \ln f_n(x_n / \theta)$$

log-Likelihood-Funktion.

Def.: Die Maximalzahl der linear unabhängigen Spaltenvektoren bzw. der linear unabhängigen Zeilenvektoren einer $(n \times m)$ -Matrix X heißt **Rang** von X .

Es gilt: $rg(X) \leq \min\{n, m\}$

Falls $rg(X) = \min\{n, m\}$ besitzt die Matrix X vollen Rang.

Def.: Ein Zufallsvektor \mathbf{x} bzw. seine Verteilung gehört zu einer Exponentialfamilie¹, wenn die Dichte von der Form

$$f(\mathbf{x} / \boldsymbol{\tau}) = c(\mathbf{x}) \exp \left\{ \sum_{j=1}^m \varphi_j(\boldsymbol{\tau}) t_j(\mathbf{x}) \right\} / a(\boldsymbol{\tau})$$

ist. Dabei ist $\boldsymbol{\tau}$ ein Parametervektor, die Funktionen φ_j , t_j sind reelwertig, $a(\boldsymbol{\tau}) > 0$ und $c(\mathbf{x}) \geq 0$.

Sind die Funktionen $1, t_1(\mathbf{x}), \dots, t_m(\mathbf{x})$ bzw. $1, \varphi_1(\boldsymbol{\tau}), \dots, \varphi_m(\boldsymbol{\tau})$ linear unabhängig, dann bezeichnet man $\boldsymbol{\theta}_1 = (\varphi_1(\boldsymbol{\tau}), \dots, \varphi_m(\boldsymbol{\tau})) = \boldsymbol{\varphi}_m(\boldsymbol{\tau})$ als natürliche (kanonische) Parameter. Mit $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$ und $\mathbf{t}(\mathbf{x}) = (t_1(\mathbf{x}), \dots, t_m(\mathbf{x}))'$ ergibt sich die Dichte zu

$$f(\mathbf{x} / \boldsymbol{\theta}) = c(\mathbf{x}) \exp \{ \boldsymbol{\theta}' \mathbf{t}(\mathbf{x}) \} / a(\boldsymbol{\theta})$$

für $b(\boldsymbol{\theta}) = \ln a(\boldsymbol{\theta}) \quad \Rightarrow \quad f(\mathbf{x} / \boldsymbol{\theta}) = c(\mathbf{x}) \exp \{ \boldsymbol{\theta}' \mathbf{t}(\mathbf{x}) - b(\boldsymbol{\theta}) \}$

und man bezeichnet die Exponentialfamilie als m-parametrig.

Es gelten folgende Eigenschaften:

$$E(\mathbf{x}) = b'(\boldsymbol{\theta}) = \boldsymbol{\mu}$$

$$\text{cov}(\mathbf{x}) = b''(\boldsymbol{\theta})$$

und es existiert eine Umkehrabbildung $g^{-1}(\boldsymbol{\theta}) = \boldsymbol{\mu}$ und $\boldsymbol{\theta} = g(\boldsymbol{\mu})$.

Mit der $\boldsymbol{\mu}$ -Parametrisierung erhält man dann:

$$\bar{f}(\mathbf{x} / \boldsymbol{\mu}) = c(\mathbf{x}) \exp \{ \mathbf{g}(\boldsymbol{\mu})' \mathbf{t}(\mathbf{x}) - b(\mathbf{g}(\boldsymbol{\mu})) \}$$

Es gelten die Identitäten:

$$\mathbf{g}'(\boldsymbol{\mu}) = (\text{cov}(\mathbf{x}))^{-1} = \mathbf{diag} \left\{ \frac{1}{\sigma^2(\boldsymbol{\mu})} \right\}$$

$$\boldsymbol{\mu} = \text{cov}(\mathbf{x}) \frac{\partial b(\mathbf{g}(\mathbf{x}))}{\partial \boldsymbol{\mu}} = \mathbf{g}'(\boldsymbol{\mu})^{-1} \frac{\partial b(\mathbf{g}(\mathbf{x}))}{\partial \boldsymbol{\mu}}$$

Def: Für ein optimales Ergebnis der Schätzung mit der **Kleinste Quadrat Schätzung (OLS)** sollten die einzelnen Elemente des Zufallsfehlers idealerweise unkorreliert und eine gemeinsame Varianz besitzen.

Sie besteht darin, einen Parameterschätzer $\hat{\beta}$ für β so zu bestimmen, dass die Summe der Fehlerquadrate minimal wird:

$$\sum_i^I \varepsilon_i^2 = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

$\hat{\beta}$ genügt dann der Normalgleichung:

$$\begin{aligned} (\mathbf{X}'\mathbf{X})\hat{\beta} &= \mathbf{X}'\mathbf{y} \\ \Rightarrow \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned}$$

(vorausgesetzt \mathbf{X} hat vollen Rang)

Satz: Seien x_1, \dots, x_N unabhängig und identisch verteilt. Unter geeigneten Regularitätsannahmen (Existenz, Stetigkeit der Ableitungen), ist dann der **ML-Schätzer konsistent und asymptotisch normalverteilt**:

$$\sqrt{N}(\hat{\theta}_N - \theta) \rightarrow N(\theta, (\mathbf{F}(\theta))^{-1}) \quad \text{bzw.} \quad \hat{\theta}_N \sim N(\theta, (\mathbf{F}(\theta))^{-1} / N)$$

Satz: Das System $\mathbf{X}\beta = \mathbf{y}$ besitzt genau dann eine **Lösung**, wenn

$$rg(\mathbf{X}, \mathbf{y}) = rg(\mathbf{X}) \quad \text{gilt. Dies}$$

bedeutet: $rg(\mathbf{X}) = n$ für $\mathbf{X} \in M_{n,m}$

Sie ist **eindeutig** falls die Matrix \mathbf{X} vollen Rang ($rg(\mathbf{X}) = m$) besitzt.

Satz: Es gelte $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ mit $E(\mathbf{y}) = \mathbf{X}\beta$ und $cov(\mathbf{y}) = \sigma^2 \mathbf{I}$. Dann folgt:

$\hat{\beta}$ und $\hat{\sigma}^2$ sind **unverzerrt**.

$$cov(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

B 2 GLM (Logitanalyse)

B 2.1 Charakterisierung

Das Generalisierte lineare Modell (GLM) lässt sich folgendermaßen charakterisieren:

Sei (y_n, \mathbf{x}_n) mit $n = 1, \dots, N$ gleich die Daten der Beobachtungen von (y, \mathbf{x}) für jedes Element n aus einer Stichprobe des Umfangs N , mit $\mathbf{x} = (x_1, \dots, x_p)$.

Wobei p die Anzahl der unabhängigen Variablen darstellt.

Zufallskomponente: Die Responsvariable y_n wird als unabhängige Zufallsvariable aufgefasst. Deren Dichte einer Exponentialfamilien (Normal-, Binomial-, Poisson-, Gammaverteilung) mit natürlichem Parameter θ_n , Dispersionsparameter ϕ , Gewicht w_n angehört.

$$f(y_n / \theta_n, \phi, w_n) = c(y_n, \phi, w_n) \exp\left\{ \frac{y_n \theta_n - b(\theta_n)}{\phi} w_n \right\}. \quad (1.1)$$

Wobei $b(\cdot)$ und $c(\cdot)$ die jeweiligen Exponentialfamilien spezifizieren.

Der natürliche Parameter θ_n ist mit dem Erwartungswert μ_n durch die Beziehung

$$\mu_n = b'(\theta_n) \quad \text{verbunden.}$$

Für die Varianz von y_n gilt:

$$\sigma^2(\mu_n) = b''(\theta_n) \frac{\phi}{w_n}$$

Da $b''(\theta_n)$ positiv definit ist, ist μ_n eindeutig und es existiert eine Umkehrabbildung g mit

$$g^{-1}(\theta_n) = \mu_n$$

Systematische Komponente: Die erklärenden Variablen \mathbf{x}_n beeinflussen y_n als Linearkombination:

$$y_n = \mathbf{x}_n' \boldsymbol{\beta} = \beta_0 + \beta_1 x_{n1} + \dots + \beta_p x_{np}$$

Link: Der Erwartungswert $E(y_n) = \mu_n$ ist mit einer hinreichend oft differenzierbaren Linkfunktion

$$g(\mu_n) = y_n = \mathbf{x}_n' \boldsymbol{\beta} \quad \text{verbunden.}$$

Linkfunktionen, welche den natürlichen Parameter gleich dem linearen Prädiktor benutzen, nennt man kanonische (natürliche) Linkfunktion ($\theta_n = g(\mu_n) = y_n$).

Es gilt dann, die in der Erwartungswertparametrisierung erhaltene Funktion:

$$g_n^{-1}(\theta_n) = \mu_n = b'(\theta_n)$$

Setzt man $\theta_n = g(\mu_n) = \mathbf{x}_n' \boldsymbol{\beta}$ und $t_n(y_n) = y_n \mathbf{x}_n$ in

$$\bar{f}(\mathbf{x} / \boldsymbol{\mu}) = c(\mathbf{x}) \exp\{\mathbf{g}(\boldsymbol{\mu})' \mathbf{t}(\mathbf{x}) - b(\mathbf{g}(\boldsymbol{\mu}))\}$$

ein, so erhält man

$$\bar{f}(y_n / \boldsymbol{\beta}) = c(y_n) \exp\{\boldsymbol{\beta}' y_n \mathbf{x}_n - b(\mathbf{x}_n' \boldsymbol{\beta})\}.$$

bzw. mit (1.1)

$$\bar{f}(y_n / \boldsymbol{\beta}) = c(y_n) \exp\left\{\frac{\boldsymbol{\beta}' y_n \mathbf{x}_n - b(\mathbf{x}_n' \boldsymbol{\beta})}{\phi} w_n\right\}. \quad (1.2)$$

Natürliche Linkfunktionen besitzen vorteilhafte Eigenschaften bezüglich Existenz und Eindeutigkeit. Bei anderen Linkfunktionen sind diese zu prüfen.

B 2.2 Parameterschätzung: Maximum-Likelihood-Schätzung (ML-Schätzung)

Sei \mathbf{X} die Designmatrix .

Unter der Voraussetzung, dass eine natürliche Linkfunktion verwendet wird erhält man aus (1.2) die log-Likelihood-Funktion:

$$l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) = \sum_{n=1}^N \ln \bar{f}_n(y_n / \boldsymbol{\beta}) = \sum_{n=1}^N \left\{ \frac{\boldsymbol{\beta}' y_n \mathbf{x}_n - b(\mathbf{x}_n' \boldsymbol{\beta})}{\phi} w_n \right\} + \underbrace{C}_{\text{Konstante}}$$

Das Maximum dieser Funktion liefert das Nullsetzen der ersten partiellen Ableitung. Für $i = 0, \dots, p$ gilt für die n -te Beobachtung:

$$\frac{\partial l_n}{\partial \beta_i} = \frac{w_n}{\phi} \cdot \frac{(y_n - \mu_n)}{l'_n(\theta_n) = y_n - b'(\theta_n)} \cdot \frac{1}{\underbrace{\text{var}(y_n)}_{g'(\mu_n)}} \cdot \frac{[g'(\mu_n)]^{-1}}{\mu'_n(y_n)} \cdot \frac{x_{ni}}{y'_n(\beta_i)}$$

Mit

$$l = \sum_{n=1}^N l_n, \quad \mathbf{y} = (y_1, \dots, y_N)', \quad \boldsymbol{\mu} = (\mu_1, \dots, \mu_N)', \quad \mathbf{g} = (g(\mu_1), \dots, g(\mu_N))', \quad \mathbf{w} = (w_1, \dots, w_N)',$$

$$\mathbf{V} = \{\text{cov}(\mathbf{y})\}^{-1} = \text{diag}\{1/\sigma^2(\boldsymbol{\mu})\}, \quad \mathbf{G} = \text{diag}(\boldsymbol{\mu}'(\mathbf{y})), \quad \mathbf{W} = \text{diag}(\mathbf{w}), \quad \phi = \text{const} = 1$$

folgt die Score-Funktion (\mathbf{G} , \mathbf{V} und $\boldsymbol{\mu}$ hängen von $\boldsymbol{\beta}$ ab) in kompakterer Darstellung:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{s}(\boldsymbol{\beta}) = \mathbf{s}(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) = \mathbf{X}' \mathbf{G} \mathbf{V} \mathbf{W} \cdot (\mathbf{y} - \boldsymbol{\mu})$$

Man kann diese Gleichung für natürliche Linkfunktionen noch vereinfachen zu:

$$\mathbf{s}(\boldsymbol{\beta}) = \mathbf{X}' \mathbf{W} \cdot (\mathbf{y} - \boldsymbol{\mu}) \quad (1.3)$$

Diese Gleichung (1.3) nochmals abgeleitet und das Vorzeichen gewechselt ergibt dann die Fisher'sche Informationsmatrix:

$$\mathbf{F}_{(N)}(\boldsymbol{\beta}) = -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \mathbf{X}' \mathbf{W} \text{cov}(\mathbf{y}) \mathbf{X} \quad (1.4)$$

Schätzung von $\hat{\boldsymbol{\beta}}$:

Newton-Raphson-Methode

Die Newton-Raphson-Methode produziert als Nebeneffekt die asymptotischen Varianzen und Kovarianzen der ML-Schätzer.

Mit der Taylor-Entwicklung um β_k ergibt sich aus der Score-Funktion:

$$s(\beta) = \frac{\partial l}{\partial \beta}(\beta_k) = \frac{\partial l}{\partial \beta}(\hat{\beta}_k) + \frac{\partial^2 l}{\partial \beta \partial \beta'}(\hat{\beta}_k)(\beta - \hat{\beta}_k) = 0, \quad k = 0, 1, 2, \dots$$

daraus folgt:

$$\beta = \hat{\beta}_k - \left(\frac{\partial^2 l}{\partial \beta \partial \beta'}(\hat{\beta}_k) \right)^{-1} \frac{\partial l}{\partial \beta}(\hat{\beta}_k)$$

bzw.

$$\hat{\beta}_{k+1} = \hat{\beta}_k - \lambda \left(\frac{\partial^2 l}{\partial \beta \partial \beta'}(\hat{\beta}_k) \right)^{-1} \frac{\partial l}{\partial \beta}(\hat{\beta}_k)$$

und

$$\text{cov}(\hat{\beta}_{k+1}) = \left(\frac{\partial^2 l}{\partial \beta \partial \beta'}(\hat{\beta}_k) \right)^{-1}$$

Fisher-Scoring-Verfahren

Der Unterschied zur Newton-Raphson-Methode ist die Erwartungswertbildung der zweiten Ableitung der log-Likelihood-Funktion.

$$\hat{\beta}_{(k+1)} = \hat{\beta}_k + \lambda F^{-1}(\hat{\beta}_{(k)}) s(\hat{\beta}_{(k)})$$

Für große N als auch natürliche Linkfunktionen werden sich beide Verfahren, außer beim Rechenaufwand, kaum unterscheiden.

Anhang C: Bezeichnungen Notationen

C 1 Notationen

Vektoren \mathbf{x} \mathbf{v} usw. Werden in Kleinbuchstaben und fett gedruckt

Matrizen \mathbf{A} \mathbf{B} usw. Werden in Großbuchstaben und fett gedruckt

σ^2 Varianz

σ Standardabweichung

μ Erwartungswert

$f(x)$ Dichte

$F(x)$ Verteilungsfunktion

\cdot' bzw. $\partial \cdot / \partial \cdot$ Erste Ableitung einer Funktion: Hochkommas mit einfacher Schriftdicke

\cdot' bzw. \mathbf{x}' Transponierte eines Vektors bzw. Matrix: Hochkommas werden fett gedruckt

$\hat{\cdot}$ bzw. \hat{x} Schätzer werden mit Dach dargestellt

$\tilde{\cdot}$ bzw. \tilde{x} Variable nachdem eine Beobachtung weggelassen wurde:

$rg(\mathbf{A})$ Rang der Matrix A

Abkürzungen

LBP Landesamt für Bodenkultur und Pflanzenbau

WD Deutsches Weidelgras

MSA-Wert Prüfgröße des Kaiser-Mayer-Olkin-Kriteriums, S. 27

C 2 Bezeichnungen

Diskrimination	Zuordnung neuer Objekte zu einer Grundgesamtheit
Effekt	relative Unterschied zweier Populationen
Effizienz	Präzision der Schätzung, ausgedrückt durch den Standardfehler
Kategoriale Daten	Ordinale und nominale Daten
Lineares Modell	Ein Modell welches in seinen Parametern linear ist.
Metrische Daten	Eindeutig skalierte, messbare Daten mit einem Nullpunkt
Nominale Daten	Gleichwertige Daten welche nur eingeteilt aber nicht geordnet werden können z. B. blau, grün, Religion etc.
Ordinale Daten	Daten welche irgendwie geordnet sind, Abstände zwischen den Skalierungen sind willkürlich z. B. Noten oder niedrig, mittel, hoch usw.
orthonormierte Spalten a und b	$a \cdot b = 0 \wedge a = 1, b = 1$
redundant	berechenbar
restringiert	eingeschränkt
Score-Funktion	Spaltenvektor der ersten Ableitungen der log-Likelihood-Funktion
Signifikanzniveau	Irrtumswahrscheinlichkeit
Standardisierung	Auf eine Normalverteilung transformieren, üblicherweise mit $(x - \bar{x})/\sigma$

Literaturverzeichnis

- [1] AGRESTI 1996: An Introduction to Categorical Data Analysis, John Wiley & Sons, Inc., New York
- [2] ANDREß, HAGENAARS und KÜHNEL 1997: Analyse von Tabellen und kategorialen Daten, Springer-Verlag Berlin
- [3] BACKHAUS, ERICHSON, PLINKE und WEIBER 1994: Multivariate Analysemethoden, Springer-Verlag Berlin, 7. Auflage
- [4] BORTZ 1993: Statistik für Sozialwissenschaftler, Springer-Verlag Berlin, 4. Auflage
- [5] FAHRMEIR und HAMERLE 1984: Multivariate statistische Verfahren, Walter de Gruyter Berlin · New York
- [6] FAHRMEIR, HAMERLE und TUTZ 1996: Multivariate statistische Verfahren, Walter de Gruyter, Berlin · New York, 2. Auflage
- [7] HARTUNG und ELPELT 1995: Multivariate Statistik, Oldenbourg Verlag, München, 5. Auflage
- [8] KREYSZIG 1975: Statistische Methoden und ihre Anwendung, Vandenhoeck & Ruprecht , Göttingen, 7. Auflage
- [9] MYERS 1990: Classical and modern regression with applications, Duxbury Press, Wadsworth Publishing Company, Belmont, California, second edition
- [10] SACHS 1984: Angewandte Statistik, Springer-Verlag Berlin, 6. Auflage
- [11] LOGISTIC REGRESSION EXAMPLES USING THE SAS SYSTEM 1995: Version 6, First Edition, SAS Institute Inc., Cary, NC, USA, 163 pp.
- [12] SAS / STAT USER'S GUIDE 1989: Version 6, Fourth Edition, Volume 2, Cary, NC: SAS Institute Inc., 846 pp.
- [13] SAS TECHNICAL REPORT P-243, SAS/STAT SOFTWARE 1993: The Genmod Procedure, Release 6.09, Cary, NC: SAS Institute Inc., 88pp.
- [14] SAS USER'S GUIDE STATISTICS 1985: Version 5 Edition. Cary, NC: SAS Institute Inc., 956 pp.
- [15] SCHUEMER, STRÖHLEIN und GOGOLOK 1990: Datenverarbeitung und statistische Auswertung mit SAS, Band II: komplexe statistische Analyseverfahren, Gustav Fischer Verlag, Stuttgart