Modellierung der Zahlungsschwierigkeiten von Privatkunden einer Bank mittels logistischer Regression

Diplomarbeit

Zur Erlangung des Grades einer Diplom-Kauffrau
an der Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

vorgelegt von

Nadia Landmann

(Matrikel-Nr. 301102)

Prüfer: Prof. Dr. Bernd Rönz

Berlin, 19. Dezember 2000

Inhaltsverzeichnis

Abkürzungsverzeichnis **3**
SPSS-Output-Verzeichnis **4**
Tabellenverzeichnis **6**

1. Einleitung **7**

2. Beschreibung der Daten und der Variablen **10**

2.1. Die Daten und ihre Modifikation 10
2.2. Die abhängige Variable 10
2.3. Die erklärenden Variablen 11
2.3.1. Geschlecht 12
2.3.2. Region 12
2.3.3. Alter 13
2.3.4. Einkommen 15
2.3.5. Vermögen bei der Bank 15
2.3.6. Dauer der Kundenverbindung 16

3. Die logistische Regression allgemein **17**

4. Einfache logistische Regressionen **21**

4.1. Warnung 1 und Geschlecht 21
4.2. Warnung 1 und Region 29
4.3. Warnung 1 und Alter 30
4.3.1. Alter als nicht gruppierte Variable 30
4.3.2. Alter als gruppierte Variable 32
4.4. Warnung 1 und Einkommen 35
4.5. Warnung 1 und Vermögen 37
4.6. Warnung 1 und Dauer der Kundenverbindung 38

5. Multiple logistische Regression 42

5.1. Ausgangsmodell (Methode Einschluß) 43

5.2. Ausgangsmodell (Methode Vorwärts) 51

5.3. Multiples Modell mit veränderten Bezugskategorien der Variablen ALTER_G und DKV_G 58

5.4. Überprüfung und Berücksichtigung von Wechselwirkungen zwischen den Variablen REGION_1 und DKV_G 61

5.5. Identifikation von Ausreißern und das daraus folgende endgültige Modell dieser Untersuchung 66

5.5.1. Identifikation von Ausreißern 66

5.5.2. Die letzte multiple logistische Regression als Ergebnis dieser Untersuchung 72

6. Modelldiagnose 78

6.1. Prüfung der Linearität der erklärenden Variablen 78

6.2. Modellvalidität (Modelvalidity) 82

7. Zusammenfassung und Ausblick 85

Anhang A 88

Anhang B 89

Literaturverzeichnis 91

Abkürzungsverzeichnis

bzw.	beziehungsweise
c.p.	ceteris paribus
ca.	circa
d.h.	das heißt
DKV	Dauer der Kundenverbindung
Hrsg.	Herausgeber
i.d.R.	in der Regel
ID	Identifikationsnummer
SPSS	Superior Performing Software System
u.a.	unter anderem
u.ä.	und ähnliche
vgl.	vergleiche
z.B.	zum Beispiel

SPSS-Output-Verzeichnis

SPSS-Output 2.1:	Häufigkeitstabellen für REGION_1 und REGION_2	13
SPSS-Output 2.2:	Häufigkeiten der Kategorien der Variablen ALTER_G	14
SPSS-Output 2.3:	Häufigkeiten der Kategorien der Variablen DKV_G	16
SPSS-Output 4.1:	Logistische Regression mit Warnung 1 und Geschlecht	22
SPSS-Output 4.2:	Logistische Regression mit Warnung 1 und Region	29
SPSS-Output 4.3:	Logistische Regression mit Warnung 1 und Alter	30
SPSS-Output 4.4:	Logistische Regression mit Warnung 1 und Alter gruppiert; Indikator-Codierung (Bezugskategorie = die erste Kategorie)	33
SPSS-Output 4.5:	Logistische Regression mit Warnung 1 und bekanntes Einkommen	35
SPSS-Output 4.6:	Scatterplot bekanntes Einkommen gegen geschätzte Erfolgswahrscheinlichkeit ($\hat{\pi}_k$)	37
SPSS-Output 4.7:	Logistische Regression mit Warnung 1 und Vermögen bei der Bank	37
SPSS-Output 4.8:	Logistische Regression mit Warnung 1 und DKV gruppiert; Indikator-Codierung (Bezugskategorie = die letzte Kategorie)	39
SPSS-Output 4.9:	Scatterplot DKV_G gegen geschätzte Wahrscheinlichkeit ($\hat{\pi}_k$)	40
SPSS-Output 5.1:	Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die erste Kategorie), bekanntes Einkommen, Vermögen bei der Bank und DKV gruppiert (Bezugskategorie = die letzte Kategorie); Methode Einschluß	43
SPSS-Output 5.2:	Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die erste Kategorie), bekanntes Einkommen, Vermögen bei der Bank und DKV gruppiert (Bezugskategorie = die letzte Kategorie); Methode Vorwärts (Wald)	51
SPSS-Output 5.3:	Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die letzte Kategorie), bekanntes Einkommen, Vermögen bei der Bank und DKV gruppiert (Bezugskategorie = die erste Kategorie); Methode Einschluß	59
SPSS-Output 5.4:	Chi-Quadrat-Unabhängigkeitstest nach Pearson zur Überprüfung von Wechselwirkungen zwischen den Variablen REGION_1 und DKV_G	61
SPSS-Output 5.5:	Cramer-V zur Überprüfung der Stärke der Beziehung zwischen den Variablen REGION_1 und DKV_G	62
SPSS-Output 5.6:	Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die letzte Kategorie), bekanntes Einkommen und Vermögen bei der Bank; Methode Einschluß	63
SPSS-Output 5.7:	Scatterplot Identifikationsnummern (ID) gegen standardisierte Residuen	67
SPSS-Output 5.8:	Scatterplot Identifikationsnummern (ID) gegen Devianceabweichung	67

SPSS-Output 5.9: Scatterplot Identifikationsnummern (ID) gegen Cook's Distance 68

SPSS-Output 5.10: Die fünf größten Werte der Cook's Distance und die Identifikationsnummern der dazugehörigen Fälle 69

SPSS-Output 5.11: Scatterplot Identifikationsnummern (ID) gegen DFBETA für Vermögen 69

SPSS-Output 5.12: Scatterplot Identifikationsnummern (ID) gegen Vermögen bei der Bank, gegeben Warnung 1 70

SPSS-Output 5.13: Die kleinsten Werte der geschätzten Wahrscheinlichkeiten, gegeben Warnung 1 71

SPSS-Output 5.14: Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die letzte Kategorie), bekanntes Einkommen und Vermögen bei der Bank; Methode Einschluß
- ohne Fälle: ID = 10.226 und ID = 10.951 - 72

SPSS-Output 6.1: Kreuztabelle Warnung 1 und Alter gruppiert mit den beobachteten bedingten relativen Häufigkeiten 79

SPSS-Output 6.2: Scatterplot Altersklassen der Variablen ALTER_G - log(p/1-p) 79

SPSS-Output 6.3: Kreuztabelle Warnung 1 und Einkommen gruppiert mit den beobachteten bedingten relativen Häufigkeiten 80

SPSS-Output 6.4: Scatterplot Einkommensklassen der Variablen EINK_G - log(p/1-p) 81

SPSS-Output 6.5: Kreuztabelle Warnung 1 und Vermögen gruppiert mit den beobachteten bedingten relativen Häufigkeiten 82

SPSS-Output 6.6: Scatterplot Vermögensklassen der Variablen VERM_G - log(p/1-p) 82

Tabellenverzeichnis

Tabelle 4.1: Codierung der Altersklassen, Gruppierung des Alters und geschätzte Erfolgswahrscheinlichkeit $\hat{\pi}_k$ je Altersklasse 35

Tabelle 4.2: Codierung der DKV, Gruppierung der DKV und geschätzte Erfolgswahrscheinlichkeit $\hat{\pi}_k$ je Gruppe der DKV 41

Tabelle 5.1: Vergleich der geschätzten Regressionskoeffizienten der einfachen Modelle des 4. Kapitels mit denjenigen des multiplen Modells im SPSS-Output 5.1 48

Tabelle 5.2: Variablennamen, Wert der R-Statistik und unter "Sig" angegebenes Ergebnis des Wald-Tests je Variable, absteigend sortiert nach der Höhe des Absolutwertes der R-Statistik 50

Tabelle 5.3: Vergleich der geschätzten Regressionskoeffizienten der zwei multiplen Modelle der SPSS-Outputs 5.1 und 5.2 58

Tabelle 5.4: Vergleich der geschätzten Regressionskoeffizienten der zwei multiplen Modelle der SPSS-Outputs 5.3 und 5.6 65

Tabelle 5.5: Variablen und dazugehörige Werte der R-Statistik aus dem SPSS-Output 5.6, absteigend sortiert nach der Höhe des Absolutwertes der R-Statistik, und die Werte der R-Statistik für die gleichen Variablen aus dem SPSS-Output 5.3 65

Tabelle 5.6: Vergleich der geschätzten Regressionskoeffizienten der zwei multiplen Modelle der SPSS-Outputs 5.6 und 5.14 75

Tabelle 5.7: Variablen und dazugehörige Werte der R-Statistik aus dem SPSS-Output 5.14, absteigend sortiert nach der Höhe des Absolutwertes der R-Statistik 77

Tabelle 6.1: Gruppierung von EINK (Einkommen) in die neue Variable EINK_G 80

Tabelle 6.2: Gruppierung von VERMOEGE (Vermögen bei der Bank) in die neue Variable VERM_G 81

Tabelle 6.3: Modellvalidität mit verschiedenen Schwellenwerten 84

1. Einleitung

Im Bankensektor gewinnt das Problem der Identifikation und des Managements verschiedenartiger Risiken zunehmend an Bedeutung. Gründe dafür sind einerseits die besonders in diesem Bereich stattfindende starke Globalisierung. Sie führt zu immer komplexeren Aufgabenstellungen und Risiken. Andererseits findet ein enormer und rasch schärfer werdender Wettbewerb zwischen den Finanzinstituten statt. Dadurch sind alle Anbieter auf diesem Markt gezwungen, zunehmend transparenter zu agieren. Ein wichtiger Punkt hierbei ist die Fähigkeit, in kürzester Zeit angemessene Preise stellen zu können. Dazu ist es notwendig, sehr schnell und effektiv die mit den jeweiligen Produkten verbundenen Risiken einschätzen zu können.

Ein wesentlicher Risikofaktor ist die Zahlungsfähigkeit der Kunden. Entscheidend für deren Beurteilung ist die Kenntnis über die Faktoren, von denen die Zahlungsschwierigkeiten eines Kunden generell abhängen können.

Bei dieser Problemstellung können sowohl Firmenkunden als auch Privatkunden betrachtet werden. Bei der ersten Gruppe sind mögliche Einflußfaktoren die Branche und Größe des Unternehmens sowie die Bilanzkennzahlen. Bei der zweiten dagegen können für die Beurteilung der Zahlungsfähigkeit Faktoren wie Alter, Geschlecht, Einkommen, Familienstand u.ä. von Bedeutung sein.

Von großem Interesse für die Finanzinstitute ist hierbei nicht nur die Identifizierung der in Frage kommenden Einflußfaktoren, sondern auch das Ausmaß ihrer Wirkung auf das Auftreten von Zahlungsschwierigkeiten.

Wegen der hohen Geschwindigkeit des Marktes wird es immer wichtiger, daß die Mitarbeiter einer Bank bei jedem Kunden anhand weniger Eckdaten eine relativ verläßliche Prognose über seine Zahlungsfähigkeit machen können. So können sie schnell zu einer Aussage über das aus Sicht der Bank mit diesem Kunden verbundene Risiko gelangen. Dadurch wird viel Zeit und Aufwand gespart, weil dann eine umfangreiche Analyse nur bei Kunden durchzuführen ist, die aufgrund der Risikoeinstufung zu den gefährdeten Gruppen gehören, aber dennoch ein Darlehen oder eine sonstige Form der Finanzierung anstreben. Eine verläßliche Risikoeinschätzung ermöglicht zusätzlich eine individuelle und

kundenorientierte Preisgestaltung. Dies erhöht die Wettbewerbsfähigkeit und die Kundenzufriedenheit.

Um eine solche Risikobeurteilung realisieren zu können, ist es notwendig, die über Jahre hinweg in Form von Kundendaten gesammelten Erfahrungen im Hinblick auf diese Problemstellung auszuwerten. Probleme bei dieser Datenauswertung ergeben sich aus der Tatsache, daß die abhängige Variable, formuliert zum Beispiel als "*Auftreten von Zahlungsschwierigkeiten*", eine binäre (ja/nein) oder eine mehrkategoriale Variable (z.B. keine, wenig, große Zahlungsschwierigkeiten) ist. Hier können eine Reihe von sonst üblichen statistischen Verfahren wie z.B. die lineare Regression nicht angewandt werden, da einige wesentliche Voraussetzungen (u.a. normalverteilte Fehlervariablen) nicht erfüllt sind.

Eine geeignete Methode für die Auswertung ist die logistische Regression, soweit die Fragestellung so gestellt ist, daß die Y-Variable binär ist und damit das Vorhandensein oder Nichtvorhandensein einer Eigenschaft oder eines Ereignisses ausdrückt.

Ziel der vorliegenden Arbeit ist daher, die Zahlungsschwierigkeiten von Privatkunden einer Bank mittels der logistischen Regression unter Verwendung des Statistikprogramms SPSS in der Version 9.0 zu modellieren, um herauszufinden, ob und wenn ja, in welchem Ausmaß bestimmte Faktoren Einfluß auf die Wahrscheinlichkeit haben, daß Zahlungsschwierigkeiten auftreten. Ein weiteres Ziel ist es zu überprüfen, ob die logistische Regression, die neben der Diskriminanzanalyse im Bereich des Credit Scoring[1] eine entscheidende Rolle spielt, auch für eine Fragestellung, bei der es nicht unmittelbar um eine Kreditwürdigkeitsprüfung geht, hilfreich ist.

Im nachfolgenden Abschnitt werden die Daten, ihre Modifikation und die bei der Untersuchung benutzten Variablen vorgestellt. Im Anschluß wird allgemein das Modell der logistischen Regression erläutert. Im vierten Abschnitt wird eine Reihe von einfachen logistischen Regressionen betrachtet, um den Einfluß je einer erklärenden Variablen auf die abhängige Variable zu untersuchen. An dieser Stelle wird ein erster Überblick gegeben, bevor dann im fünften Abschnitt auf die Schätzung von ausgewählten multiplen Modellen übergegangen wird. Hierbei werden die möglichen Zusammenhänge zwischen der abhängigen Variablen und mehreren erklärenden Variablen untersucht. Dabei können sich Ver-

[1] Zu Credit Scoring siehe u.a. Fahrmeir, L., Hamerle, A., S. 334 - 339.

änderungen im Verhältnis zu den festgestellten Ergebnissen der einfachen logistischen Regressionen ergeben, da zwischen den im jeweiligen Modell enthaltenen erklärenden Variablen Assoziationen auftreten können. Diese Wechselwirkungen erschweren die Auswertung der Ergebnisse. Dennoch liegt der Schwerpunkt dieser Untersuchung bei den multiplen Modellen, da sie in größerem Maße der Realität entsprechen, in der ebenfalls mehrere Faktoren gleichzeitig auf die Zahlungsfähigkeit des Kunden einwirken.

Neben der Modellschätzung beginnt die Modelldiagnose bereits im fünften Abschnitt mit der Identifikation von Ausreißern. Nachdem anschließend das endgültige multiple Modell als Resultat dieser Untersuchung vorgestellt wird, wird die Modelldiagnostik im sechsten Abschnitt mit der Überprüfung der X-Variablen auf Linearität in der link Funktion sowie der Modellvalidität fortgesetzt.

Im letzten Abschnitt erfolgen eine Zusammenfassung und einige abschließende Bemerkungen.

2. Beschreibung der Daten und der Variablen

2.1. Die Daten und ihre Modifikation

Die von einer großen deutschen Bank zur Verfügung gestellten Daten sind eine Zufallsstichprobe aus dem dortigen Datenbestand. Sie umfaßten 12.515 Fälle (Kunden). Zunächst wurde die Variable LFD_NR mit einer fortlaufenden Numerierung von 1 bis 12.515 eingefügt. Im Anschluß wurden die Fälle gelöscht, die unter der Variablen ALTER einen Wert von "999" aufwiesen. Hierbei handelt es sich bei dieser Bank um sogenannte "Gemeinschaftskunden".

Ein "Gemeinschaftskunde" entsteht durch den Verbund von zwei oder mehreren Personen zum Zwecke der gemeinsamen Kontoführung. Um dies erkennbar zu machen, wird ein fiktives Geburtsdatum eingegeben, wodurch bei dem Alter ein Wert von "999" errechnet wird. Da, wie bereits erläutert, die "Gemeinschaftskunden" keine einzelnen Personen sind, wäre es falsch, sie mit den übrigen Kunden zu vergleichen. Aus diesem Grund wurden die betroffenen 1.136 Fälle aus den Daten entfernt. Somit verbleiben für die Untersuchung 11.379 einzelne Kunden.

Im Anschluß wurde die Variable ID als fallbezogene Identifikationsnummer hinzugefügt. Sie ist ebenfalls eine fortlaufende Numerierung von 1 bis 11.379 und dient der Identifikation, der Datenzuordnung und der Datenüberprüfung.

Die hierdurch entstandene Datei wurde mit Hilfe des SPSS-Zufallsgenerators in zwei Stichproben mit einem Umfang von 5.684 bzw. 5.695 Kunden getrennt. Alle Untersuchungen in den Abschnitten 4., 5. und 6.1. werden mittels der ersten Datei mit 5.684 Personen durchgeführt. Anschließend wird der zweite Teil der ursprünglichen Datei im Abschnitt 6.2. für die Überprüfung der Modellvalidität verwendet.

2.2. Die abhängige Variable

Ein Hinweis darauf, daß bei einem Kunden Zahlungsschwierigkeiten aufgetreten sind, gibt die in der Datei enthaltene Variable ANZ_WAR1. Sie umfaßt die Anzahl der pro Kunde im System der Bank eingetragenen Warnungen 1. Dies ist die höchste Warnstufe und wird vergeben, wenn erhebliche Zahlungsschwierigkeiten aufgetreten sind. Das kann zum

Beispiel dann der Fall sein, wenn davon ausgegangen wird, daß der Kunde das von ihm aufgenommene Darlehen nicht mehr zurückzahlen kann. Ein weiterer Grund kann die Überziehung des Girokontos sein, wenn dies in erheblichem Maße über dem eingeräumten bzw. geduldeten Dispositionskredit hinaus geschieht. Daher ist die Vergabe von Warnung 1 nicht nur in Verbindung mit Krediten im engeren Sinne zu sehen. Es liegen keine Informationen darüber vor, weshalb bei dem jeweiligen Kunden eine oder mehrere solche Warnungen vorhanden sind. Um eine Verwechslung zu vermeiden, wird daher im weiteren von Zahlungsschwierigkeiten als Oberbegriff und nicht von Kreditschwierigkeiten gesprochen.

Es kommt selten vor, daß Warnung 1 mehr als einmal bei demselben Kunden vergeben wird. Daher wird diese Variable in eine neue binäre Variable WARN_1 umgewandelt. Ein Wert von "1" bedeutet hierbei das Vorhandensein einer oder mehrerer Warnungen 1. Ein Wert von "0" bedeutet dagegen, daß der Kunde keine Warnung 1 erhalten hat. WARN_1 wird als abhängige Variable in dieser Untersuchung festgelegt.

Warnungen werden von dem jeweils zuständigen Kundenbetreuer in das System eingegeben. Allerdings gibt es keine genauen Regeln, nach denen dies geschieht. Daher kann gesagt werden, daß die Warnungen in gewisser Weise subjektiv vergeben werden. Eine bessere Kenntnis über den vorhandenen Datenbestand der Bank und damit nicht nur über die eigenen Erfahrungen mit den Kunden, sondern auch über die der Kollegen kann sehr nützlich sein. Dies sowie die Beachtung statistisch überprüfter Einflußgrößen in bezug auf das Auftreten von Zahlungsschwierigkeiten können zu einer kundenorientierten Betreuung führen, die jedoch besser die individuellen Risiken berücksichtigt. Außerdem kann eine einheitliche Vorgehensweise z.B. bei der Vergabe von Warnungen erzielt werden. Zum Erreichen solcher Ziele können die Ergebnisse dieser Arbeit beitragen.

2.3. Die erklärenden Variablen

Als unabhängige (erklärende) Variablen $X_1,..., X_6$ werden folgende Variablen in die Analyse einbezogen:

- → Geschlecht
- → Region
- → Alter
- → Einkommen
- → Vermögen bei dieser Bank
- → Dauer der Kundenverbindung

2.3.1. Geschlecht

Das Geschlecht, ursprünglich mit dem Variablennamen GESCHLEC, ist als eine nominalskalierte binäre Stringvariable mit "m" für männlich und "w" für weiblich erfaßt. Für die nachfolgenden Untersuchungen wird diese aus Gründen der Vereinfachung und der Vergleichbarkeit in die neue numerische Variable SEX mit den Ausprägungen "0" für männlich und "1" für weiblich umgewandelt. Diese Zuordnung der Werte führt dazu, daß in den logistischen Regressionen, wenn diese X-Variable enthalten ist, innerhalb von SEX die männlichen Kunden als Bezugskategorie festgelegt werden. Der Grund dafür ist, daß SPSS bei binären Variablen als Bezugskategorie immer die Ausprägung mit dem niedrigeren Wert (in diesem Fall "0") wählt. Daher sind die Ergebnisse in bezug auf diese Variable jeweils als Vergleich der weiblichen Kunden zu den männlichen Kunden zu interpretieren.[2]

2.3.2. Region

Die zur Verfügung stehenden Daten bieten keine explizite Variable "Region", im Sinne von wohnhaft in den neuen oder in den alten Bundesländern.[3] Diese Information kann dennoch aus zwei anderen vorhandenen Variablen RISIKO und STATUS abgeleitet werden. Beide Variablen enthalten jeweils einen Buchstaben "O" für Ost oder "W" für West sowie eine Ziffer von Null bis Neun, die ein bestimmtes Niveau an Risiko oder Status ausdrückt, das hier nicht von Interesse ist.

Für die regionale Zuordnung wurde eine neue Codierung vorgenommen. Aus der ordinalskalierten Variablen RISIKO entstand die neue nominalskalierte Variable REGION_1, indem alle Werte von "O0" bis "O9" mit "0", alle Werte von "W0" bis "W9" mit "1" und alle fehlenden Werte von RISIKO durch den Wert "5" ersetzt wurden. Damit hat die neue Variable REGION_1 folgende Ausprägungen:

- "0" für alle Kunden, die laut dem letzten Informationsstand in den neuen Bundesländern wohnen;
- "1" für alle, die laut diesem in den alten Bundesländern wohnen;
- "5" für die Kunden, bei denen keine solchen Angaben vorliegen.

---

[2] Siehe hierzu u.a. Abschnitt 4.1.

[3] Die folgenden Begriffe werden im weiteren synonym angewandt: "neue Bundesländer", "Ost" und "ehemaliger Osten" als Bezeichnung für die frühere DDR sowie "alte Bundesländer" und "West" für die frühere Bundesrepublik Deutschland.

Es ist hier besonders darauf hinzuweisen, daß die Variable REGION_1 nicht die ursprüngliche Herkunft des Kunden, sondern seinen Wohnsitz angibt.

Der Wert "5" wird als ein benutzerdefinierter Missing-Wert vereinbart und dadurch bei allen durchgeführten Analysen nicht mit einbezogen.

Die Codierung der Variablen STATUS in die neue Variable REGION_2 erfolgte auf die gleiche Weise. Im Anschluß wurden die neuen Variablen REGION_1 und REGION_2 verglichen. Wie der SPSS-Output 2.1 zeigt, stimmen diese vollkommen überein. Da nur eine Variable notwendig ist, wird in dieser Arbeit stets die Variable REGION_1 verwendet.

SPSS-Output 2.1: Häufigkeitstabellen für REGION_1 und REGION_2

REGION_1: Region (aus RISIKO)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	wohnhaft im Osten	4072	35,8	41,0	41,0
	wohnhaft im Westen	5851	51,4	59,0	100,0
	Gesamt	9923	87,2	100,0	
Fehlend	unbekannte Region	1456	12,8		
Gesamt		11379	100,0		

REGION_2: Region (aus STATUS)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	wohnhaft im Osten	4072	35,8	41,0	41,0
	wohnhaft im Westen	5851	51,4	59,0	100,0
	Gesamt	9923	87,2	100,0	
Fehlend	unbekannte Region	1456	12,8		
Gesamt		11379	100,0		

2.3.3. Alter

Die Variable ALTER ist im Gegensatz zu den zwei vorgenannten Variablen metrisch skaliert und gibt das tatsächliche Alter des jeweiligen Kunden in Lebensjahren an. Ein Wert von "0" bedeutet, daß der Kunde unter einem Jahr alt ist, ein Wert von "1", daß es sich um einen Kunden im Alter von eins bis unter zwei Jahren handelt usw.

Diese Variable hat eine besondere Ausprägung: "119". Dieser Wert bedeutet, daß bei dem Kunden keine Angaben über das Geburtsdatum vorhanden sind und somit das Alter nicht errechnet werden kann. Hierfür kann es mehrere Ursachen geben, die sich im Einzelfall nicht ermitteln lassen. Zum Beispiel war bei der Einführung des Eingabeprogramms in der Bank kurzzeitig das Geburtsdatum nicht als Pflichtfeld definiert. Ein weiterer Grund kann die Tatsache sein, daß im Zuge der Wiedervereinigung bei manchen Kunden aus den neuen Bundesländern das Geburtsdatum nicht bekannt war und durch einen fiktiven Wert ersetzt wurde. Unabhängig von der Ursache für das Fehlen der Altersangabe muß die Ausprägung "119" als Missing-Wert definiert werden.

Vorgreifend auf Abschnitt 4.3.2. ist hier zu erwähnen, daß eine Gruppierung der Variablen ALTER in eine neue Variable ALTER_G ("_G" für die Gruppierung) vorgenommen wurde. Es handelt sich hierbei um eine Gruppierung in sechs Altersklassen. Die jeweiligen Klassengrenzen sind dem SPSS-Output 2.2 zu entnehmen. Der Wert "119" wurde außerhalb dieser Gruppierung gelassen, durch den Wert "20" ersetzt und erneut als Missing definiert. Um dies zu ermöglichen, wurde für die obere Klassengrenze der letzten Altersklasse der Wert "118" gewählt. In der gesamten Datei mit 11.379 Fällen hat das Alter den höchsten Wert von "111". Somit ist gewährleistet, daß in der letzten Altersklasse "70-118 Lebensjahre" alle Kunden enthalten sind, die 70-jährig oder älter sind. Mit Ausnahme des folgenden SPSS-Outputs wird diese Altersklasse im weiteren als "70 Lebensjahre und älter" bezeichnet.

SPSS-Output 2.2: Häufigkeiten der Kategorien der Variablen ALTER_G

Alter gruppiert

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0-29 Lebensjahre	3481	30,6	31,4	31,4
	30-39 Lebensjahre	3067	27,0	27,7	59,0
	40-49 Lebensjahre	1852	16,3	16,7	75,7
	50-59 Lebensjahre	1146	10,1	10,3	86,1
	60-69 Lebensjahre	773	6,8	7,0	93,0
	70-118 Lebensjahre	772	6,8	7,0	100,0
	Gesamt	11091	97,5	100,0	
Fehlend	119 = Missing	288	2,5		
Gesamt		11379	100,0		

2.3.4. Einkommen

Bei der Variablen mit dem Variablennamen EINK handelt es sich nicht direkt um das Einkommen, sondern um einen Teil der aus Sicht der Bank bekannten Einkünfte. Es werden zunächst die Eingänge auf dem Privatgirokonto bzw. auf den Privatgirokonten des Kunden bei der Bank ermittelt. Daraus wird ein Durchschnitt für die letzten zwölf Monate gebildet, in Hunderter gruppiert und unter EINK angegeben. Ein Wert von z.B. "15" bedeutet, daß bei dieser Person in den letzten zwölf Monaten im Durchschnitt monatlich zwischen 1.451 und 1.550 DM auf dem Privatgirokonto bei dieser Bank eingegangen sind.

Es ist allerdings nicht ausgeschlossen, daß der Kunde sein Gehaltskonto oder sonstige Einkünfte bei einer anderen Bank hat und dadurch wahrscheinlich ein wesentlicher Teil seines Einkommens unter dieser Variablen nicht erfaßt wird. Er kann aber auch weitere Einkünfte haben, die zwar bei dieser Bank aber nicht auf seinem Privatgirokonto eingehen und somit ebenfalls nicht ausgewiesen werden. Ein Wert von "0" bedeutet damit nicht zwangsläufig, daß die Person kein Einkommen hat.

Es geht hier also um das aus Sicht der Bank bekannte Einkommen, und zwar um das Ganze, falls Einkünfte nur auf dem Privatgirokonto eingehen oder um einen Teil davon, soweit Einzahlungen auf andere Konten dieser Bank stattfinden. Diese Variable wird im weiteren auch als "bekanntes Einkommen" bezeichnet.

In den beiden Stichproben befinden sich jeweils vier Kunden, die unter EINK einen Wert größer Null aufweisen, aber kein Privatgirokonto haben. Das sind Personen, die ihr Privatgirokonto bei diesem Geldinstitut gekündigt haben. Da sie aber für einen gewissen Zeitraum in den letzten zwölf Monaten ein solches Konto bei der Bank besaßen, wird für sie ebenfalls ein solcher monatlicher Durchschnittswert ermittelt und ausgewiesen.

2.3.5. Vermögen bei der Bank

Unter der Variablen VERMOEGE wird das ganze Vermögen des Kunden bei dieser Bank angegeben. Dieses umfaßt, soweit vorhanden, die Guthaben auf Privatgiro-, Geschäftsgiro- und Kreditkartenkonten, Termingelder, Sparkonten, Sparbriefe, Depotkonten usw. Das Vermögen wird hierbei pfenniggenau angegeben.

2.3.6. Dauer der Kundenverbindung

Die Variable DAUER_KV gibt die Dauer der Kundenverbindung[4] in Monaten an. Diese Dauer wurde erst mit der Einführung dieses EDV-Systems vor 206 Monaten (ca. 17 Jahre) erfaßt. Damit ist das der höchste Wert dieser Variablen in der Datei, selbst wenn die Person tatsächlich länger Kunde ist. Gleichzeitig ist das mit 27,5 % der am häufigsten vertretene Wert. Zwei weitere Häufungen ergeben sich bei 116 und 117 Monaten. In dieser Zeit erfolgte die technische Zusammenführung der Systeme aufgrund der Währungsunion, wodurch Kunden aus der früheren DDR hinzukamen.

Wegen der Häufigkeitsverteilung dieser Variablen erscheint es für die Problemstellung dieser Arbeit sinnvoll, eine Gruppierung der Dauer der Kundenverbindung in folgender Weise vorzunehmen:

SPSS-Output 2.3: Häufigkeiten der Kategorien der Variablen DKV_G

DKV gruppiert

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	DKV <= 61 Monate	1148	10,1	10,1	10,1
	DKV = 62 - 105 Monate	1143	10,0	10,0	20,1
	DKV = 106 - 115 Monate	1059	9,3	9,3	29,4
	DKV = 116 - 117 Monate	2871	25,2	25,2	54,7
	DKV = 118 - 132 Monate	432	3,8	3,8	58,5
	DKV = 133 - 156 Monate	523	4,6	4,6	63,1
	DKV = 157 - 180 Monate	545	4,8	4,8	67,9
	DKV = 181 - 205 Monate	525	4,6	4,6	72,5
	DKV = 206 Monate und länger	3133	27,5	27,5	100,0
	Gesamt	11379	100,0	100,0	

Die so gruppierte Variable mit dem Variablennamen DKV_G und nicht die ungruppierte Variable DAUER_KV wird in die Untersuchungen dieser Arbeit mit einbezogen.

[4] Dauer der Kundenverbindung wird im folgenden mit DKV abgekürzt.

3. Die logistische Regression allgemein

Wie bereits erwähnt, wird in dieser Arbeit das Auftreten von Zahlungsschwierigkeiten durch das Auftreten der Warnung 1 modelliert. Hierbei ist die abhängige Variable, zunächst symbolisiert mit Z, eine nominalskalierte binäre Zufallsvariable. Für Z gilt:

$$Z = \begin{cases} 1 & \text{wenn Warnung 1 auftritt (= Erfolg)} \\ 0 & \text{wenn keine Warnung 1 auftritt (= Mißerfolg)} \end{cases}$$

mit den Wahrscheinlichkeiten $P(Z = 1) = \pi$ und $P(Z = 0) = 1 - \pi$. Ihre Wahrscheinlichkeitsverteilung ist:

$$f(z;\pi) = P(Z = z) = \pi^z (1-\pi)^{1-z} \quad , \quad z = 0, 1 \, , \qquad \text{Gleichung 3.1}$$

bekannt als Bernoulli-Verteilung, mit dem Erwartungswert $E(Z) = \pi$ und der Varianz $Var(Z) = \pi(1-\pi)$.

Da es sich nicht nur um eine, sondern um N statistische Einheiten (Kunden) handelt, an denen unter gleichen Bedingungen und unabhängig voneinander das Auftreten von Warnung 1 beobachtet wurde, existieren N Zufallsvariablen Z_i (i = 1,…,N). Für sie gelten analog die Wahrscheinlichkeiten $P(Z_i = 1) = \pi_i$ und $P(Z_i = 0) = 1 - \pi_i$, d.h. die Wahrscheinlichkeitsverteilung ist:

$$f(z_i;\pi_i) = P(Z_i = z_i) = \pi_i^{z_i} (1-\pi_i)^{1-z_i} \quad , \qquad z_i = 0, 1 \, , \qquad i = 1,\ldots,N \qquad \text{Gleichung 3.2}$$

mit $E(Z_i) = \pi_i$ und $Var(Z_i) = \pi_i(1-\pi_i)$.[5]

Neben der abhängigen Variablen gibt es p erklärende Variablen X_j, mit j = 1,…,p. Die Werte der X-Variablen sind fest. Sie werden in einem Vektor $\mathbf{x} = (x_1 \ldots x_p)^T$ zusammengefaßt. Da die X-Variablen unterschiedliche Werte haben, ergeben sich q verschiedene Vektoren $\mathbf{x}_k = (x_{k1} \ldots x_{kp})^T$, k = 1,...,q. Jede Gruppe k von statistischen Einheiten (Fällen) ist durch denselben Vektor $\mathbf{x}_k$ gekennzeichnet.

[5] Hierzu sowie zu den weiteren Ausführungen vgl. u.a. Rönz, B. (1997b), S. 75 - 112; Collett, D., S. 18 f. und 56 ff.

Es wird angenommen, daß die Erfolgswahrscheinlichkeit π_k in jeder Gruppe konstant ist. Dann folgt die response Variable Y_k (k = 1,…,q) als Anzahl der Erfolge in der k-ten Gruppe der Binomialverteilung, d.h. $Y_k \sim B(n_k; \pi_k)$ mit

$$f(y_k;\pi_k) = \binom{n_k}{y_k} \pi_k^{y_k} (1-\pi_k)^{n_k - y_k} \quad . \qquad \text{Gleichung 3.3}$$

Nach einigen Umformungen kann die letzte Gleichung geschrieben werden als:

$$f(y_k;\pi_k) = \exp\left[y_k \log\left(\frac{\pi_k}{1-\pi_k} \right) + n_k \log(1-\pi_k) + \log\binom{n_k}{y_k} \right] . \qquad \text{Gleichung 3.4}$$

Das Ziel der statistischen Analyse ist, die unbekannten Erfolgswahrscheinlichkeiten π_k (also die jeweilige Wahrscheinlichkeit, daß in der k-ten Gruppe Warnung 1 eintritt) zu schätzen und herauszufinden, wodurch das Auftreten von Warnung 1 und damit die Wahrscheinlichkeiten beeinflußt werden. Somit ergibt sich für alle Y_k die Likelihood-Funktion als:

$$L(\pi_1,...,\pi_q; y_1,...,y_q) = \prod_{k=1}^{q} f(y_k;\pi_k) = \prod_{k=1}^{q} \binom{n_k}{y_k} \pi_k^{y_k} (1-\pi_k)^{n_k - y_k} \qquad \text{Gleichung 3.5}$$

und die log-Likelihood-Funktion als:

$$l(\pi_1,...,\pi_q; y_1,...,y_q) = \sum_{k=1}^{q} \left[y_k \log\left(\frac{\pi_k}{1-\pi_k} \right) + n_k \log(1-\pi_k) + \log\binom{n_k}{y_k} \right] . \qquad \text{Gleichung 3.6}$$

Die Abhängigkeit der Erfolgswahrscheinlichkeiten π_k von den erklärenden Variablen kann folgendermaßen zum Ausdruck gebracht werden:

$$g(\pi_k) = \eta_k = \mathbf{x}_k^T \boldsymbol{\beta} = \Sigma_j x_{kj}\beta_j , \quad k = 1,\ldots,q , \quad j = 1,...,p . \qquad \text{Gleichung 3.7}$$

Hierbei wird g als link Funktion bezeichnet und $\boldsymbol{\beta} = (\beta_1 \ldots \beta_p)^T$ ist ein Vektor unbekannter Parameter. Zunächst werden diese Parameter geschätzt, um dann mit diesen die Wahrscheinlichkeiten π_k zu schätzen. Dazu muß die link Funktion spezifiziert werden, indem zur Absicherung von $0 \le \pi \le 1$ die Modellierung von π als Verteilungsfunktion erfolgt.[6] Als Verteilungsfunktion können unter anderem die Normalverteilung oder die logistische Verteilung gewählt werden. Im ersten Fall handelt es sich um das sogenannte Probit-Modell, im zweiten um das Logit-Modell. Es kann gezeigt werden, daß sich die Schätzergebnisse dieser zwei Modelle für π_k wenig von einander unterscheiden.[7]

In dieser Arbeit wird das Logit-Modell (logistische Regression) gewählt, da es sich in zahlreichen praktischen Untersuchungen als geeignet herausgestellt hat. Somit ergibt sich unter Verwendung der Gleichung 3.7 und Vernachlässigung des Laufindexes k folgende logistische Verteilung:

$$\pi = g^{-1}(\eta) = g^{-1}\left(\sum_j x_j \beta_j\right) = \frac{\exp(\eta)}{1+\exp(\eta)} = \frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{1+e^{\eta}} \quad . \qquad \text{Gleichung 3.8}$$

Diese Gleichung wird nach e^{η} aufgelöst:

$$e^{\eta} = \exp(\eta) = \frac{\pi}{1-\pi} \quad , \qquad \text{Gleichung 3.9}$$

worin das Verhältnis der Erfolgswahrscheinlichkeit π zur Mißerfolgswahrscheinlichkeit $1-\pi$ als odds (Chancen) eines Erfolges bezeichnet wird. Wird dieses Verhältnis anschließend logarithmiert, ergibt sich unmittelbar die link Funktion (auch logit Funktion genannt):

$$\eta = g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \sum_j x_j \beta_j \quad , \qquad \text{Gleichung 3.10}$$

die linear in den Parametern ist. Der Logarithmus von den odds wird auch log odds genannt.

---

[6] Siehe hierzu Gleichung 3.8.

[7] Vgl. Rönz, B. (1997b), S. 114 f. und Collett, D., S. 54 f.

Zur Schätzung des Logit-Modells wird die Gleichung 3.10 in die log-Likelihood-Funktion (Gleichung 3.6) eingesetzt. Somit werden die unbekannten Wahrscheinlichkeiten π_k durch den Vektor $\boldsymbol{\beta}$ ersetzt:

$$l(\beta;y)=\sum_{k=1}^{q}\left\{\sum_{j=1}^{p} y_k x_{kj}\beta_j - n_k \log\left[1+\exp\left(\sum_{j=1}^{p} x_{kj}\beta_j\right)\right]+\log\binom{n_k}{y_k}\right\}. \qquad \text{Gleichung 3.11}$$

Gleichung 3.11 wird nach β_j abgeleitet:

$$\frac{\partial l}{\partial \beta_j}=\sum_{k=1}^{q}\left\{y_k x_{kj} - n_k x_{kj}\left[\frac{\exp\left(\sum_{j=1}^{p} x_{kj}\beta_j\right)}{1+\exp\left(\sum_{j=1}^{p} x_{kj}\beta_j\right)}\right]\right\}=\sum_{k=1}^{q} x_{kj}\left(y_k - n_k\pi_k\right) \qquad \text{Gleichung 3.12}$$

und anschließend Null gesetzt. Dies führt zu nichtlinearen Gleichungen, die iterativ bei Vorgabe der Startwerte für $\boldsymbol{\beta}$ gelöst werden. Die hiermit geschätzten Regressionskoeffizienten b_j werden dann in Gleichung 3.8 eingesetzt, um dadurch die Schätzung der Erfolgswahrscheinlichkeiten π_k zu erhalten.

4. Einfache logistische Regressionen

Alle in diesem Kapitel behandelten Modelle sind einfache logistische Regressionen mit je einer erklärenden Variablen und WARN_1 (Auftreten von Warnung 1) als abhängige Variable.

Das erste Modell wird im Abschnitt 4.1. sehr ausführlich behandelt, weil an diesem Beispiel, in Ergänzung zum 3. Kapitel, einige theoretische Hintergründe verdeutlicht sowie die für die Untersuchung wichtigen Maßzahlen und deren Berechnungsweise erläutert werden. Dabei handelt es sich zum Teil auch um Erläuterungen, die erst später behandelte Modelle betreffen, aber sinngemäß zu diesen theoretischen Ausführungen gehören. Sie werden auch deshalb bereits hier aufgeführt, weil sie sich anhand einer einfachen logistischen Regression relativ übersichtlich darstellen lassen. Damit wird es möglich, sich später hauptsächlich auf die Interpretation der Ergebnisse zu konzentrieren.

Bevor zur eigentlichen Untersuchung übergegangen wird, ist, um Manipulationen zu vermeiden, das Signifikanzniveau α zur Überprüfung von Testergebnissen vorzugeben. In dieser Arbeit wird als Entscheidungsgrundlage stets ein $\alpha = 0{,}05$ verwendet.[8]

4.1. Warnung 1 und Geschlecht

Als erste unabhängige Variable wird Geschlecht mit dem Variablennamen SEX und den Ausprägungen "0" für männlich und "1" für weiblich in die Untersuchung einbezogen. Bis auf weiteres wird für die Aufnahme der X-Variablen in die Regressionsfunktion die Methode Einschluß (Enter) gewählt.[9] Der SPSS-Output 4.1 enthält die Ergebnisse dieses ersten Modells.[10]

[8] Zur näheren Erläuterung, wie Testentscheidungen mit Hilfe des vorgegebenen Signifikanzniveaus unter SPSS erfolgen, siehe Anhang A.

[9] Erst bei Modellen mit mehr als einer X-Variablen können auch andere Methoden benutzt werden, bei denen die erklärenden Variablen nicht wie bei "Einschluß" in einem Schritt aufgenommen werden, sondern schrittweise in die Regressionsfunktion eingeschlossen (Vorwärts/Forward) bzw. aus dieser ausgeschlossen (Rückwärts/Backward) werden.

[10] Bei allen SPSS-Outputs wird nur der für die betreffende Auswertung relevante Teil des SPSS-Viewers übernommen.

SPSS-Output 4.1: Logistische Regression mit Warnung 1 und Geschlecht

```
-2 Log Likelihood   7864,2757
* Constant is included in the model.

Beginning Block Number  1.  Method: Enter
Variable(s) Entered on Step Number
1..       SEX       Geschl_umkodiert

Estimation terminated at iteration number 2 because
Log Likelihood decreased by less than ,01 percent.

-2 Log Likelihood     7554,777
Cox & Snell - R^2         ,053
Nagelkerke - R^2          ,071

                    Chi-Square   df  Significance
Model                 309,499     1         ,0000

Classification Table for WARN_1
The Cut Value is ,50               Predicted
                        keine Warnung 1  Warnung 1 vorhan  Percent Correct
                                  k             W
                       ┌──────────────┬──────────────┐
Observed               │              │              │
keine Warnung 1    k   │      1650    │      1340    │   55,18%
                       ├──────────────┼──────────────┤
Warnung 1 vorhan   W   │       865    │      1829    │   67,89%
                       └──────────────┴──────────────┘
                                              Overall  61,21%

---------------------- Variables in the Equation -----------------------
Variable            B      S.E.       Wald    df      Sig       R    Exp(B)
SEX           -,9568     ,0553   299,6697     1    ,0000  -,1946     ,3841
Constant       ,3111     ,0360    74,8504     1    ,0000
```

Zunächst ist die Güte der Anpassung der logistischen Regression von Interesse. Sie kann in einem relativen Sinne mit Hilfe der Reduktion in der Deviance ΔD beurteilt werden. Für die Ermittlung von ΔD wird zuerst für ein Anfangsmodell M_0, das nur eine Konstante und keine erklärenden Variablen enthält, der maximale Wert der log-Likelihood-Funktion (Gleichung 3.6) berechnet, d.h. der Wert an der Stelle $\hat{\pi}_0$. Dieser wird dann mit "-2" multipliziert. Der daraus resultierende Wert von $-2l_0(\hat{\pi}_0; y)$, auch als -2LL$_0$ bezeichnet, ist im SPSS-Output 4.1 als erster "-2 Log Likelihood"-Wert in Höhe von 7.864,2757 angegeben. Danach wird in gleicher Weise für das Modell M_1 mit einer Konstanten und der Variablen SEX als erklärende Variable der Wert $-2l_1(\hat{\pi}_1; y)$, auch als -2LL$_1$ bezeichnet, ermittelt. Er ist hier als zweiter "-2 Log Likelihood"-Wert in Höhe von 7.554,777 angegeben. Anschließend wird die Reduktion in der Deviance als Differenz dieser zwei Werte berechnet:[11]

[11] Vgl. u.a. Rönz, B. (1997b), S. 90 - 93 sowie Collett, D., S. 62 - 69.

$$\begin{aligned}\Delta D &= 2\left[l_1\left(\hat{\pi}_1;y\right)-l_0\left(\hat{\pi}_0;y\right)\right]\\ &= \left[-2\,l_0\left(\hat{\pi}_0;y\right)\right]+\left[2\,l_1\left(\hat{\pi}_1;y\right)\right]\\ &= \left[-2\,l_0\left(\hat{\pi}_0;y\right)\right]-\left[-2\,l_1\left(\hat{\pi}_1;y\right)\right].\end{aligned}$$ Gleichung 4.1[12]

Unter "Chi-Square" wird die als "Model" bezeichnete ΔD in Höhe von 309,499 angegeben.

Grundsätzlich führt ein gutes Modell zu einem niedrigen Wert des -2LL.[13, 14] Dies und Gleichung 4.1 machen deutlich, daß ein Wert von ΔD größer Null (wie in diesem Fall) eine Verbesserung des Anfangsmodells durch die Hinzunahme der erklärenden Variablen (hier SEX) bedeutet. Das Modell M_1 ist hier also besser an die Daten angepaßt als das Modell M_0. Diese Verbesserung ist um so größer, je größer die Reduktion in der Deviance ist. Um eine Aussage darüber treffen zu können, ob die Verbesserung signifikant ist, wird der Wert von ΔD mit dem Quantil der Chi-Quadrat-Verteilung für das in dieser Arbeit in Höhe von 0,05 vorgegebene Signifikanzniveau und für df = 1 Freiheitsgrad[15] verglichen.[16] Das unter "Significance" angegebene α_{sign} ist kleiner als 0,05. Daher ist die Verbesserung signifikant, oder anders ausgedrückt, trägt das Geschlecht zur Erklärung des Auftretens von Warnung 1 bei.

---

[12] Gleichung 4.1 ergibt sich unmittelbar aus der Definition von ΔD als:

$$\begin{aligned}\Delta D = D_0 - D_1 &= 2\left[l_{\max}\left(\hat{\pi}_{\max};y\right)-l_o\left(\hat{\pi}_0;y\right)\right]-2\left[l_{\max}\left(\hat{\pi}_{\max};y\right)-l_1\left(\hat{\pi}_1;y\right)\right]\\ &= 2\left[l_1\left(\hat{\pi}_1;y\right)-l_0\left(\hat{\pi}_0;y\right)\right].\end{aligned}$$

Darin ist

$D_0 = 2\left[l_{\max}\left(\hat{\pi}_{\max};y\right)-l_0\left(\hat{\pi}_0;y\right)\right]$	die Deviance für das Modell M_0, das nur eine Konstante enthält;
$D_1 = 2\left[l_{\max}\left(\hat{\pi}_{\max};y\right)-l_1\left(\hat{\pi}_1;y\right)\right]$	die Deviance für das Modell M_1 mit einer Konstanten und p erklärenden Variablen;
$l_{\max}\left(\hat{\pi}_{\max};y\right)$	der maximale Wert der log-Likelihood-Funktion gemäß Gleichung 3.6 für ein sogenanntes maximales (full) Modell, das weder eine Konstante noch erklärende Variablen enthält;
$l_0\left(\hat{\pi}_0;y\right)$	der maximale Wert der log-Likelihood-Funktion gemäß Gleichung 3.6 für das Modell M_0 und
$l_1\left(\hat{\pi}_1;y\right)$	der maximale Wert der log-Likelihood-Funktion gemäß Gleichung 3.6 für das Modell M_1.

[13] Vgl. SPSS Regression Models™ 9.0, S. 45.

[14] Mit -2LL ist hier allgemein der mit "-2" multiplizierte Wert der log-Likelihood-Funktion für das jeweilige Modell, d.h. für M_0, M_1 oder für ein anderes Modell, bezeichnet.

[15] Die Anzahl der Freiheitsgrade ist 1, da das Modell M_1 genau einen Parameter mehr enthält als das Modell M_0. Allgemein entspricht df der Differenz in der Anzahl der X-Variablen in M_1 und M_0.

[16] Vgl. u.a. Rönz, B. (1997b), S. 97; Dobson, A. J., S. 61 f.

Die Maßzahlen Cox & Snell R^2 und Nagelkerke $\widetilde{R}^2$ quantifizieren den durch die logistische Regression erklärten Anteil der Variation der abhängigen Variablen Y. Sie sind dem Bestimmtheitsmaß einer linearen Regression ähnlich.[17] An dieser Stelle ist jedoch zu betonen, daß bei einer logistischen Regression die Variation von Y anders als bei der linearen Regression definiert ist, da Y nicht metrisch, sondern nominalskaliert und dichotom ist.

Cox & Snell R^2 wird berechnet als:[18]

$$R^2 = 1 - \left[\frac{L_0}{L_1} \right]^{2/N}, \qquad \text{Gleichung 4.2}$$

wobei L_0 der maximale Wert der Likelihood-Funktion des Modells M_0 mit nur einer Konstanten, L_1 der maximale Wert der Likelihood-Funktion des Modells M_1 mit p erklärenden Variablen und N der Umfang der Stichprobe ist.

Es gilt $0 \leq R^2 < 1$, d.h. diese Maßzahl kann den maximalen Wert von 1 nicht erreichen. Daher wird die Anwendung von Nagelkerke $\widetilde{R}^2$ empfohlen, für das $0 \leq \widetilde{R}^2 \leq 1$ gilt. Der Wert von Nagelkerke $\widetilde{R}^2$ ergibt sich als Transformation von Cox & Snell R^2:

$$\widetilde{R}^2 = \frac{R^2}{R^2_{\max}}, \qquad \text{Gleichung 4.3}$$

mit $R^2_{\max} = 1 - [L_0]^{2/N}$.

Im SPSS-Output 4.1 ist ein Nagelkerke $\widetilde{R}^2$ in Höhe von 0,071 angegeben. Es werden also 7,1 % der Variation der response Variablen Auftreten von Warnung 1 durch die im Modell M_1 enthaltene unabhängige Variable Geschlecht erklärt. Dies ist relativ wenig. Die Aufnahme der Variablen SEX in das Modell trägt zwar zur Verbesserung der logistischen Regression bei, aber es gibt vermutlich noch eine Reihe von weiteren Einflußfaktoren, die für das Auftreten der Warnung 1 eine noch wesentlichere Rolle spielen können. Danach unter den in dieser Datei vorhandenen Variablen zu suchen, ist

[17] Vgl. Bühl, A., Zöfel, P. (2000), S. 358.
[18] Vgl. SPSS Regression Models™ 9.0, S. 46.

das Ziel in den nächsten Abschnitten dieser Arbeit. Zunächst aber werden die Ergebnisse des ersten Modells weiter erläutert.

Unter "Classification Table for WARN_1" erfolgt im SPSS-Output 4.1 eine Gegenüberstellung zwischen den beobachteten und den aufgrund des Modells vorhergesagten Gruppenzugehörigkeiten. Die Ermittlung der Gruppenzugehörigkeit erfolgt mit Hilfe des "Cut Value" = 0,5, d.h. die Fälle mit $\hat{\pi}_k < 0{,}5$ werden der Gruppe "keine Warnung 1", alle anderen der Gruppe "Warnung 1 vorhanden" zugeordnet.[19] Aus der Klassifikationstabelle folgt, daß 1.650 Kunden, die keine Warnung 1 erhalten haben, vom Modell korrekt zugeordnet wurden. Dagegen wurden 865 Kunden vom Modell falsch klassifiziert, da für sie keine Warnung 1 erwartet wurde, obwohl sie tatsächlich mindestens eine solche haben. Die Zelle rechts oben in der Klassifikationstabelle gibt an, daß das Modell für 1.340 Fälle ein positives Ereignis geschätzt hat, obwohl sie tatsächlich keine Warnung 1 erhalten haben. Die restlichen 1.829 Kunden wurden vom Modell korrekt als Personen mit mindestens einer Warnung 1 identifiziert. Insgesamt wurden also 61,21 % aller 5.684 Fälle vom Modell M_1 richtig erkannt.

Die Klassifikationstabelle gibt keine Auskunft darüber, ob bei den oben genannten $(865 + 1.340) = 2.205$ falsch zugeordneten Fällen die geschätzte Wahrscheinlichkeit des Auftretens mindestens einer Warnung 1 nahe 50 % oder weit entfernt von 50 % lag. Offensichtlich kann man mittels der Klassifikationstabelle nur erste Eindrücke über die Güte des Modells gewinnen. Die Überprüfung dieser ist aber erst mit Hilfe anderer Größen wie z.B. mit ΔD möglich.

Am Ende des SPSS-Outputs 4.1 werden unter "Variables in the Equation" u.a. die geschätzten Regressionskoeffizienten, ihre Standardfehler und die Ergebnisse der anhand dieser Werte durchgeführten Signifikanzüberprüfung[20] angegeben. Es ist daraus zu entnehmen, daß die Konstante der Regression b_0 einen Wert von 0,3111 und einen Standardfehler von 0,036 hat. Dagegen hat der Koeffizient b_1 einen Wert von -0,9568 mit einem Standardfehler von 0,0553. Damit ergibt sich für die link Funktion, die allgemein die Verknüpfung der abhängigen Variablen mit der/den erklärenden Variablen darstellt, folgendes:

[19] Zu "Cut Value" siehe SPSS Regression Models™ 9.0, S. 42 f.
[20] Siehe Gleichung 4.6.

$g(\hat{\pi}_k) = \hat{\eta}_k = \mathbf{x}_k{}^T \mathbf{b} = 0{,}3111 - 0{,}9568*(\text{sex})$. Gleichung 4.4

Um die Wahrscheinlichkeit des Auftretens von Warnung 1 π_k in Abhängigkeit von der erklärenden Variablen Geschlecht zu schätzen, wird das letzte Ergebnis in Gleichung 3.8 eingesetzt:

$$\hat{\pi}_k = g^{-1}(\hat{\eta}_k) = \frac{1}{1+e^{-\hat{\eta}_k}} = \frac{1}{1+e^{-[0{,}3111 - 0{,}9568*(sex)]}} \quad .$$ Gleichung 4.5

Somit ergibt sich für alle Frauen (sex = 1) ein $\hat{\eta}_k$ von -0,6457 und für alle Männer (sex = 0) ein $\hat{\eta}_k$ von 0,3111. Die Wahrscheinlichkeit, daß bei einer Frau Warnung 1 auftritt, wird auf 0,3440 geschätzt, bei einem Mann dagegen auf 0,5772. Die geschätzte Erfolgswahrscheinlichkeit für das Auftreten von Warnung 1, als Signal für das Vorliegen von erheblichen Zahlungsschwierigkeiten, ist also bei den Männern höher als bei den Frauen.

Unter Berücksichtigung von $\{\pi(1) / [1-\pi(1)]\}$ und $\{\pi(0) / [1-\pi(0)]\}$ als odds für x = 1 (weiblich) und x = 0 (männlich) und gemäß der Gleichung 3.9 gilt für das Modell M_1:

$$\frac{\pi(1)}{1-\pi(1)} = \exp(\beta_0 + \beta_1) \qquad \text{sowie} \qquad \frac{\pi(0)}{1-\pi(0)} = \exp(\beta_0) \; .$$

Das Verhältnis der letzten zwei Gleichungen ist das sogenannte odds ratio und wird weiter mit OR symbolisiert:

$$OR = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1) \; .$$

Dieses odds ratio ist ein Assoziationsmaß und gibt an, wieviel wahrscheinlicher es ist, daß das interessierende Ereignis (hier das Auftreten von Warnung 1) unter den Fällen mit x = 1

im Gegensatz zu den Fällen mit x = 0 eintritt.[21] OR kann direkt aus dem logistischen Modell geschätzt werden.

Im SPSS-Output 4.1 ist das OR unter "Exp(B)" in Höhe von 0,3841 angegeben. Damit ist es für eine Frau, wie bereits mit der Berechnung von $\hat{\pi}_k$ gezeigt wurde, weniger wahrscheinlich eine Warnung 1 zu erhalten. Anders ausgedrückt, ausgehend von der vorliegenden Stichprobe ist es für einen Mann $(1 / 0{,}3841) = 2{,}6$-mal wahrscheinlicher, eine Warnung 1 zu erhalten als für eine Frau.

Nun ist die konkrete Interpretation von β_j von Interesse. Besonders ist darauf hinzuweisen, daß die Interpretation der Regressionskoeffizienten hier anders als bei einer linearen Regression ist. Der Koeffizient β_j einer linearen Regression gibt an, wie sich der Wert der abhängigen Variablen im Durchschnitt verändert, wenn die unabhängige Variable X_j um eine Einheit erhöht wird. Um dagegen die Interpretation der Koeffizienten einer logistischen Regression zu verstehen, sei an Gleichung 3.10 unter Berücksichtigung des Laufindexes k erinnert:

$$\eta_k = g(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right) = \sum_j x_{kj}\beta_j \quad .$$

Der Koeffizient β_j einer logistischen Regression kann somit allgemein als die Veränderung in den log odds interpretiert werden, wenn die Variable X_j um eine Einheit erhöht wird und alle anderen erklärenden Variablen konstant bleiben. Der im vorliegenden Modell geschätzte Koeffizient $b_1 = -0{,}9568$ bedeutet somit, daß die log odds bei den Frauen (x = 1) um 0,9568 niedriger sind als bei den Männern (x = 0).

Der sogenannte Wald-Test[22] ermöglicht die Überprüfung der einzelnen Regressionskoeffizienten β_j auf Signifikanz. Bei diesem Test lautet die Nullhypothese H_0: $\beta_j = 0$. Die Teststatistik ist wie folgt definiert:

- wenn die zu b_j dazugehörige X_j Variable nicht mehrkategorial ist, gilt:

[21] Vgl. Hosmer, D. W., Lemeshow, S., S. 40 f.
[22] Vgl. u.a. Rönz, B. (1997b), S. 97 sowie Kleinbaum, D. G., S. 134 f.

$$W = [b_j / s(b_j)]^2 \,, \qquad \text{Gleichung 4.6}$$

worin W chi-quadrat-verteilt mit df = 1 Freiheitsgrad ist;

- wenn dagegen X_j mehrkategorial ist[23] und h Ausprägungen hat, werden für diese h-1 Kontrast-Variablen erzeugt, dann gilt:

$$W = \mathbf{b}_I^T \mathbf{C}^{-1} \mathbf{b}_I \,. \qquad \text{Gleichung 4.7}$$

Hierbei ist $\mathbf{b}_I$ der Vektor der Maximum-Likelihood-Schätzung der Koeffizienten der h-1 Kontrast-Variablen und $\mathbf{C}$ die asymptotische Kovarianzmatrix dieser Koeffizienten. W ist in der Gleichung 4.7 chi-quadrat-verteilt mit df = h-1 Freiheitsgraden.

Aus Gleichung 4.6 folgt, daß der Wert der Wald-Statistik $W = [(-0{,}9568) / 0{,}0553]^2 = 299{,}36$ ist. Die Differenz zum Wert im SPSS-Output 4.1 basiert auf einem Rundungsfehler. Unter "Sig" ist die Überschreitungswahrscheinlichkeit $\alpha_{sing} = 0{,}0000$ angegeben. Daraus folgt, daß aufgrund dieser Stichprobe und zum vorgegebenen Signifikanzniveau von 0,05 die Nullhypothese ($\beta_1 = 0$) abgelehnt wird. Damit hat Geschlecht einen signifikanten Einfluß auf das Auftreten der Warnung 1.

Der Wald-Test weist dennoch eine unangenehme Eigenschaft auf. Wenn der Absolutwert des Regressionskoeffizienten sehr groß ist, ist auch der Wert seines geschätzten Standardfehlers groß. Der dadurch zu kleine Wert der Wald-Statistik kann dazu führen, daß die Nullhypothese fälschlicherweise beibehalten wird. Es wird deshalb empfohlen, den oben beschriebenen Test mittels der Reduktion in der Deviance dem Wald-Test vorzuziehen. Allerdings sind beide Tests nicht ganz identisch. Bei dem Wald-Test wird geprüft, ob jeder einzelne Koeffizient β_j signifikant verschieden von Null ist. Ziel des Tests mittels ΔD ist es dagegen zu untersuchen, ob die zusätzlich aufgenommenen X-Variablen gemeinsam wesentlich zur Erklärung der log odds und damit zur Erklärung der Wahrscheinlichkeiten π_k beitragen. In dem Fall, daß es sich um ein Modell mit nur einer erklärenden Variablen handelt, sind dennoch die Aussagen beider Tests gleichbedeutend.

[23] Siehe beispielsweise das Modell im Abschnitt 4.3.2.

4.2. Warnung 1 und Region

In der zweiten vorgestellten einfachen logistischen Regression (SPSS-Output 4.2) wird die nominalskalierte binäre Variable REGION_1 als erklärende Variable gewählt. Unter der Variablen REGION_1 gibt es in der für die Untersuchung verwendeten Stichprobe mit einem Umfang von 5.684 Kunden 747 Fälle, bei denen nicht bekannt ist, ob sie im ehemaligen Osten oder Westen wohnen. Wie bereits unter 2.3.2. erläutert, sind diese Fälle als Missings definiert und werden automatisch nicht berücksichtigt. Daher basiert das Modell der logistischen Regression hier nur auf 4.937 Fällen.[24]

SPSS-Output 4.2: Logistische Regression mit Warnung 1 und Region

```
Number of selected cases:                 5684
Number rejected because of missing data:   747
Number of cases included in the analysis: 4937

-2 Log Likelihood   6843,4772
* Constant is included in the model.

Beginning Block Number  1.  Method: Enter
Variable(s) Entered on Step Number
1..       REGION_1  Region (aus Risiko)

Estimation terminated at iteration number 2 because
Log Likelihood decreased by less than ,01 percent.

-2 Log Likelihood     6689,997
Cox & Snell - R^2         ,031
Nagelkerke - R^2          ,041

                    Chi-Square   df  Significance
Model                 153,480     1         ,0000

Classification Table for WARN_1
The Cut Value is ,50              Predicted
                       keine Warnung 1  Warnung 1 vorhan  Percent Correct
                                     k                W
                      ┌──────────────┬────────────────┐
Observed              │              │                │
keine Warnung 1    k  │       1236   │        1261    │   49,50%
                      ├──────────────┼────────────────┤
Warnung 1 vorhan   W  │        786   │        1654    │   67,79%
                      └──────────────┴────────────────┘
                                                Overall  58,54%

---------------------- Variables in the Equation -----------------------
Variable            B      S.E.       Wald    df      Sig       R    Exp(B)
REGION_1       ,7240     ,0590   150,6562     1    ,0000   ,1474    2,0626
Constant      -,4527     ,0456    98,4530     1    ,0000
```

Der Wald-Test und der Test mittels ΔD geben hier an, daß die Region (im Sinne von wohnhaft in den neuen oder alten Bundesländern) einen signifikanten Einfluß auf das

[24] Siehe im SPSS-Output 4.2 unter "Number of cases included in the analysis".

Auftreten von Warnung 1 hat. Allerdings ist das Nagelkerke $\tilde{R}^2$ mit einem Wert von 0,041 geringer als im vorhergehenden Modell. Durch die Variable REGION_1 werden also nur 4,1 % der Variation der abhängigen Variablen erklärt. Auch der Anteil der richtig identifizierten Fälle ist mit 58,54 % laut Klassifikationstabelle etwas niedriger als im Modell mit Geschlecht als erklärende Variable.

Aus den Ergebnissen kann die link Funktion wie folgt berechnet werden:

$$\hat{\eta}_k = (-0{,}4527) + 0{,}7240*(\text{region\_1}) \; . \qquad \text{Gleichung 4.8}$$

Der Regressionskoeffizient $b_1 = 0{,}7240$ gibt an, daß die log odds bei den in den alten Bundesländern wohnenden Kunden um 0,7240 höher sind als bei denen, die in den neuen Bundesländern wohnen. Durch das Einsetzen der Ergebnisse von Gleichung 4.8 in die Gleichung 3.8 ergibt sich die Erfolgswahrscheinlichkeit für die Gruppe der im Westen wohnenden Kunden mit einem Wert von 0,5674. Für die im Osten wohnenden Kunden wird dagegen eine niedrigere Wahrscheinlichkeit für das Auftreten von Warnung 1 in Höhe von 0,3887 berechnet. Anders ausgedrückt, ausgehend von der vorliegenden Stichprobe ist das Risiko für das Auftreten einer Warnung 1 bzw. von erheblichen Zahlungsschwierigkeiten für einen im Westen wohnenden Kunden ca. 2-mal höher (OR = Exp(B) = 2,0626) als für einen im ehemaligen Osten wohnenden Kunden.

4.3. Warnung 1 und Alter

4.3.1. Alter als nicht gruppierte Variable

Die Untersuchung wird mit der Variablen ALTER als erklärende Variable fortgesetzt. Dies führt zu folgendem Output:

SPSS-Output 4.3: Logistische Regression mit Warnung 1 und Alter

```
Number of selected cases:                5684
Number rejected because of missing data:  140
Number of cases included in the analysis: 5544

-2 Log Likelihood   7680,9977
* Constant is included in the model.

Beginning Block Number  1.  Method: Enter
Variable(s) Entered on Step Number
1..       ALTER
```

```
Estimation terminated at iteration number 2 because
Log Likelihood decreased by less than ,01 percent.

-2 Log Likelihood     7544,232
Cox & Snell - R^2         ,024
Nagelkerke - R^2          ,032

                    Chi-Square   df  Significance
Model                 136,766     1         ,0000

Classification Table for WARN_1
The Cut Value is ,50               Predicted
                       keine Warnung 1  Warnung 1 vorhan  Percent Correct
                                 k              W
                      ┌──────────────┬──────────────┐
Observed              │              │              │
keine Warnung 1    k  │       1549   │       1303   │    54,31%
                      ├──────────────┼──────────────┤
Warnung 1 vorhan   W  │       1157   │       1535   │    57,02%
                      ├──────────────┼──────────────┤
                                                     Overall  55,63%

---------------------- Variables in the Equation -----------------------
Variable            B      S.E.       Wald    df      Sig       R    Exp(B)
ALTER         -,0180     ,0016   130,2963     1    ,0000  -,1292     ,9822
Constant       ,6465     ,0670    93,1694     1    ,0000
```

Auch in diesem Modell werden nicht alle 5.684 Fälle, sondern nur 5.544 Fälle bei der Schätzung mit einbezogen. Dies ist darauf zurückzuführen, daß bei 140 Kunden das Alter unbekannt ist und sie daher als Missings definiert wurden.

Das Alter hat nach dem Test mittels ΔD und dem Wald-Test ebenfalls einen signifikanten Einfluß auf das Auftreten der Warnung 1. Das Nagelkerke $\tilde{R}^2$ ist aber mit einem Wert von 0,032 noch geringer als bei den zwei vorhergehenden Modellen. Damit erklärt das Alter nur 3,2 % von der Variation der Y-Variablen. Der Anteil der vom Modell richtig klassifizierten Fälle hat sich dementsprechend auf 55,63 % verringert.

In diesem Modell ist der Absolutwert des geschätzten Regressionskoeffizienten b_1 relativ gering ($b_1 = -0{,}018$) und gibt an, daß die log odds um 0,018 sinken, wenn sich das Alter um eine Einheit (1 Lebensjahr) erhöht. Der Koeffizient b_1 ist also der Anstieg des linearen Terms und ist somit für jeden Wert von ALTER:

$$b_1 = g[\hat{\pi}_k(x+1)] - g[\hat{\pi}_k(x)] = \log[OR(x+1;x)]\,.$$[25]

[25] Vgl. Rönz, B. (1997b), S. 106.

Daraus kann das odds ratio berechnet werden als:

$OR(x+1;x) = \exp(b_1) = 0{,}9822$.

Dies bedeutet, daß mit dem Erhöhen des Alters um 1 Jahr das Risiko für das Auftreten einer Warnung 1 in einem relativ geringem Maße sinkt.

4.3.2. Alter als gruppierte Variable

In diesem Kapitel wird eine einfache logistische Regression mit der im Abschnitt 2.3.3. beschriebenen Variablen ALTER_G als erklärende Variable betrachtet und mit dem Modell unter 4.3.1. verglichen. Da die Variable ALTER_G eine kategoriale Variable mit h = 6 Kategorien ist, ist eine neue Codierung erforderlich. Dazu müssen (h-1) = 5 neue Kontrast-Variablen erstellt werden, die im folgenden mit K symbolisiert werden. Die Anzahl der Kontrast-Variablen ist genau um eins kleiner als die Anzahl der vorhandenen Kategorien, da eine der Kategorien als Bezugskategorie dient. Dadurch ergibt sich das Modell der logistischen Regression allgemein als:

$$g(\pi_k) = \eta_k = \beta_o + \beta_{k;1} K_{k1} + \beta_{k;2} K_{k2} + ... + \beta_{k;h-1} K_{k\,h-1} ,$$

wobei der erste Index bei $\beta_{k;i}$ die kategoriale Variable X_k (k = 1,...,q) und der zweite Index die Kontrast-Variable symbolisiert.

Grundsätzlich gibt es verschiedene Möglichkeiten bezüglich der Codierung der Kontrast-Variablen. Dennoch wird in dieser Arbeit ausschließlich die am häufigsten verwendete Methode, die Indikator-Codierung, angewandt. Sie ist der herkömmlichen Codierung in Dummy-Variablen (0;1-Variablen) sehr ähnlich. Die Bezugskategorie enthält in allen h-1 Kontrast-Variablen den Wert Null. Alle anderen Kategorien erhalten in je einer Kontrast-Variablen eine Eins und in den übrigen Kontrast-Variablen eine Null. Damit erfolgt die Interpretation der Ergebnisse jeder der h-1 Kategorien immer im Vergleich zur Bezugskategorie. Diese Interpretation ist nicht nur einfach, sondern auch für die hier behandelte Fragestellung sachgerecht.

In diesem Teilabschnitt wird als Bezugskategorie die erste Kategorie, die Gruppe der bis unter 30-jährigen, definiert. Diese ist im SPSS-Output 4.4 unter "Parameter" dadurch gekennzeichnet, daß sie bei allen fünf Kontrast-Variablen einen Wert von Null hat.

SPSS-Output 4.4: Logistische Regression mit Warnung 1 und Alter gruppiert; Indikator-Codierung (Bezugskategorie = die erste Kategorie)[26]

```
Number of selected cases:                 5684
Number rejected because of missing data:   140
Number of cases included in the analysis: 5544
```

			Parameter Coding				
	Value	Freq	(1)	(2)	(3)	(4)	(5)
ALTER_G							
0-29 Lebensjahre	1	1728	,000	,000	,000	,000	,000
30-39 Lebensjahre	2	1558	1,000	,000	,000	,000	,000
40-49 Lebensjahre	3	909	,000	1,000	,000	,000	,000
50-59 Lebensjahre	4	573	,000	,000	1,000	,000	,000
60-69 Lebensjahre	5	379	,000	,000	,000	1,000	,000
70 Lebensjahre und älter	6	397	,000	,000	,000	,000	1,000

```
-2 Log Likelihood   7680,9977
* Constant is included in the model.

Beginning Block Number  1.  Method: Enter
Variable(s) Entered on Step Number
1..       ALTER_G  Alter gruppiert

Estimation terminated at iteration number 3 because
Log Likelihood decreased by less than ,01 percent.

-2 Log Likelihood     7065,293
Cox & Snell - R^2         ,105
Nagelkerke - R^2          ,140
```

	Chi-Square	df	Significance
Model	615,704	5	,0000

```
Classification Table for WARN_1
The Cut Value is ,50
```

		Predicted: keine Warnung 1 k	Predicted: Warnung 1 vorhan W	Percent Correct
Observed				
keine Warnung 1	k	1902	950	66,69%
Warnung 1 vorhan	W	1175	1517	56,35%
			Overall	61,67%

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
ALTER_G			451,9698	5	,0000	,2399	
ALTER_G(1)	,7246	,0717	102,0582	1	,0000	,1141	2,0640
ALTER_G(2)	,3162	,0823	14,7620	1	,0001	,0408	1,3719
ALTER_G(3)	-,2256	,0974	5,3610	1	,0206	-,0209	,7981
ALTER_G(4)	-1,2324	,1353	82,9367	1	,0000	-,1027	,2916
ALTER_G(5)	-2,1960	,1808	147,5466	1	,0000	-,1377	,1113
Constant	-,1019	,0482	4,4776	1	,0343		

[26] In SPSS werden hierfür unter "*Kategorial*" die Variable ALTER_G als kategoriale Variable und unter "*Kontrast*" die Indikator-Codierung gewählt.

Im Vergleich zum vorhergehenden Modell mit dem nicht gruppierten Alter ist der Wert von "$-2LL_1$" auf 7.065,293 gesunken. Damit ist der Wert von ΔD höher. Dies ändert aber nichts an dem Ergebnis, daß Alter einen signifikanten Einfluß auf das Auftreten von Warnung 1 hat. Was sich dennoch ändert, d.h. sich um fast 11 Prozentpunkte verbessert, ist die durch das Nagelkerke $\widetilde{R}^2$ angegebene Anpassung der logistischen Regression an die Daten. Durch die Modifikation der erklärenden Variablen ALTER werden nun 14,0 % statt 3,2 % der Variation von Y erklärt.

Unter "Variables in the Equation" sind, abgesehen von dem Regressionskoeffizienten der Bezugskategorie, die fünf geschätzten $b_{k;i}$, deren Werte von Kategorie zu Kategorie stets sinken, und die Konstante b_0 angegeben. Der darauffolgende Wald-Test überprüft zunächst, ob die Variable ALTER_G insgesamt einen signifikanten Einfluß auf die abhängige Variable hat. Diese Frage wird hier positiv beantwortet. Danach wird untersucht, ob die Koeffizienten $b_{k;i}$ der einzelnen Kategorien verschieden von Null sind. Auch hier wird die Nullhypothese zum vorgegebenen Signifikanzniveau und aufgrund der vorliegenden Stichprobe abgelehnt.

Bezüglich der Interpretation von Exp(B) ist festzustellen, daß aufgrund der Stichprobe das Auftreten von Warnung 1 in der Gruppe der 30- bis unter 40-jährigen als 2-mal wahrscheinlicher geschätzt wird als in der Gruppe der bis unter 30-jährigen. In der Gruppe der 40- bis unter 50-jährigen ist dies im Vergleich zur Referenzkategorie nur 1,37-mal der Fall. In den anderen drei Gruppen tritt Warnung 1 jeweils seltener als bei der Bezugskategorie auf. Somit kann gesagt werden, daß bei der zweiten und dritten Altersklasse, der 30- bis unter 40-jährigen bzw. der 40- bis unter 50-jährigen, im Vergleich zur Bezugskategorie das Risiko für das Auftreten von Warnung 1 höher ($Exp(B) > 1$) und bei den anderen Altersklassen niedriger ($Exp(B) < 1$) ist. Abgesehen von der ersten Altersklasse, d.h. von der Bezugskategorie in diesem Modell, sinkt die geschätzte Wahrscheinlichkeit und damit das Risiko für das Auftreten des Ereignisses mit steigender Altersgruppe. Um dies zu verdeutlichen, sind in der nachfolgenden Tabelle 4.1 die geschätzten Wahrscheinlichkeiten $\hat{\pi}_k$ für die sechs Altersklassen angegeben:

Tabelle 4.1: Codierung der Altersklassen, Gruppierung des Alters und geschätzte Erfolgswahrscheinlichkeit $\hat{\pi}_k$ je Altersklasse

Codierung der Altersklassen	Gruppierung des Alters	$\hat{\pi}_k$
1	0 - 29 Lebensjahre	0,47454
2	30 - 39 Lebensjahre	0,65083
3	40 - 49 Lebensjahre	0,55336
4	50 - 59 Lebensjahre	0,41885
5	60 - 69 Lebensjahre	0,20845
6	70 Lebensjahre und älter	0,09130

4.4. Warnung 1 und Einkommen

Als nächste einfache logistische Regression wird ein Modell mit EINK als erklärende Variable betrachtet. Wie unter 2.3.4. bereits erläutert, handelt es sich bei dieser Variablen um die durchschnittlichen monatlichen Einkünfte der Kunden, die auf ihren Privatgirokonten bei dieser Bank eingehen und möglicherweise geringer sind als ihr tatsächliches Einkommen. Die wesentlichen Ergebnisse des Modells sind im SPSS-Output 4.5 enthalten.

SPSS-Output 4.5: Logistische Regression mit Warnung 1 und bekanntes Einkommen

```
-2 Log Likelihood   7864,2757
* Constant is included in the model.

Beginning Block Number  1.  Method: Enter
Variable(s) Entered on Step Number
1..       EINK      bekanntes Einkommen

Estimation terminated at iteration number 6 because
parameter estimates changed by less than ,001

-2 Log Likelihood     6687,757
Cox & Snell - R^2         ,187
Nagelkerke - R^2          ,250

                    Chi-Square   df  Significance
Model                1176,519     1         ,0000

Classification Table for WARN_1
The Cut Value is ,50               Predicted
                       keine Warnung 1  Warnung 1 vorhan  Percent Correct
                                  k          W
                      ┌──────────────┬──────────────┐
Observed              │              │              │
keine Warnung 1    k  │      1077    │      1913    │   36,02%
                      ├──────────────┼──────────────┤
Warnung 1 vorhan   W  │       103    │      2591    │   96,18%
                      └──────────────┴──────────────┘
                                              Overall  64,53%
```

---------------------- Variables in the Equation -----------------------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
EINK	-,2143	,0126	290,9997	1	,0000	-,1917	,8071
Constant	,2922	,0298	96,3402	1	,0000		

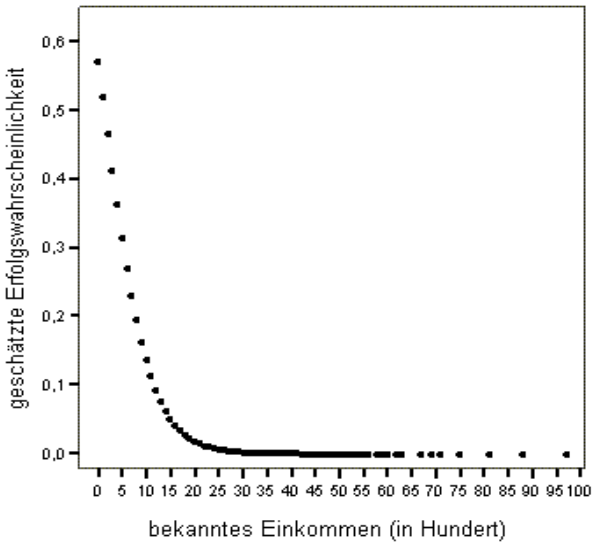
Auch das bekannte Einkommen hat nach dem Wald-Test und dem Test mittels ΔD einen signifikanten Einfluß auf Y. Allerdings erklärt diese einzelne X-Variable bereits 25 % der Variation der abhängigen Variablen.

Aus den unter "Variables in the Equation" aufgeführten Ergebnissen ergeben sich die geschätzten Erfolgswahrscheinlichkeiten als:

$$\hat{\pi}_k = \frac{1}{1 + e^{-[0{,}2922 - 0{,}2143*(eink)]}} \ .$$

Die Wahrscheinlichkeit für das Auftreten einer Warnung 1 bei einem bekannten Einkommen von Null wird somit auf 0,5725 (d.h. 57,25 %) geschätzt. Für die vorliegende Stichprobe bedeutet dies, daß bei nahezu 60 % der Kunden ohne bekanntes Einkommen mindestens eine Warnung 1 und damit erhebliche Zahlungsschwierigkeiten beobachtet wurden. Dagegen wird $\hat{\pi}_k$ bei einem Kunden mit durchschnittlichen monatlichen Eingängen zwischen 951 und 1.050 DM (EINK = 10) auf dem Privatgirokonto nur auf etwa 13,58 % geschätzt. Bei Erhöhung dieser Eingänge auf durchschnittlich 2.051 bis 2.150 DM (EINK = 21) hat $\hat{\pi}_k$ einen relativ geringen Wert von nur noch 1,47 %. Um die Veränderung von $\hat{\pi}_k$ mit steigendem Einkommen zu verdeutlichen, wird das bekannte Einkommen im SPSS-Output 4.6 gegen die in diesem Modell geschätzten Erfolgswahrscheinlichkeiten geplottet.

SPSS-Output 4.6: Scatterplot bekanntes Einkommen gegen geschätzte Erfolgswahrscheinlichkeit ($\hat{\pi}_k$)



Abschließend kann zusammengefaßt werden, daß mit steigendem Wert des bekannten Einkommens die geschätzte Wahrscheinlichkeit für das Auftreten von Warnung 1 kontinuierlich sinkt. Dieser Rückgang ist anfangs sehr stark und später nur geringfügig. Ab einem gewissen Wert des bekannten Einkommens strebt $\hat{\pi}_k$ gegen Null.

4.5. Warnung 1 und Vermögen

Das nächste betrachtete Modell, angegeben im nachfolgenden SPSS-Output, enthält als erklärende Variable das Vermögen des Kunden bei der Bank.

SPSS-Output 4.7: Logistische Regression mit Warnung 1 und Vermögen bei der Bank

```
-2 Log Likelihood   7864,2757
* Constant is included in the model.

Beginning Block Number  1.  Method: Enter
Variable(s) Entered on Step Number
1..       VERMOEGE  Vermögen bei der Bank

Estimation terminated at iteration number 7 because
Log Likelihood decreased by less than ,01 percent.

-2 Log Likelihood     6973,664
Cox & Snell - R^2         ,145
Nagelkerke - R^2          ,194

                    Chi-Square   df  Significance
Model                 890,612     1         ,0000
```

Classification Table for WARN_1

The Cut Value is ,50		Predicted		
		keine Warnung 1 k	Warnung 1 vorhan W	Percent Correct
Observed				
keine Warnung 1	k	1336	1654	44,68%
Warnung 1 vorhan	W	176	2518	93,47%
			Overall	67,80%

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
VERMOEGE	-,0002	1,339E-05	247,4308	1	,0000	-,1767	,9998
Constant	,2785	,0306	82,6526	1	,0000		

Das bei der Bank vorhandene Vermögen des Kunden hat nach dem Wald-Test und dem Test mittels ΔD, wie alle vorhergehenden X-Variablen, einen signifikanten Einfluß auf das Auftreten von Warnung 1. Die Variable erklärt 19,4 % der Variation der abhängigen Variablen. Dies ist zwar weniger als im Modell im Abschnitt 4.4., aber deutlich mehr als in den einfachen logistischen Regressionen mit Geschlecht, Region oder Alter als unabhängige Variablen. Bezüglich der Klassifikationstabelle ist die Anzahl der anhand des Modells richtig zugeordneten Fälle (67,80 %) sogar das bis jetzt höchste Ergebnis.

Der Regressionskoeffizient b_1 = -0,0002 gibt an, daß mit dem Erhöhen des Vermögens um eine Einheit, hier ein Pfennig, die log odds um 0,0002 sinken. Der Wert des odds ratio von 0,9998 bedeutet, daß die Wahrscheinlichkeit eine Warnung 1 zu bekommen, bei Erhöhung des Vermögens um einen Pfennig nahezu unverändert bleibt.

4.6. Warnung 1 und Dauer der Kundenverbindung

In der letzten hier behandelten einfachen logistischen Regression, wird der Einfluß, den die Dauer der Kundenverbindung auf das Auftreten von Warnung 1 bzw. von erheblichen Zahlungsschwierigkeiten eines Kunden ausübt, untersucht. Wie bereits im Abschnitt 2.3.6. erläutert, wird hier die in 9 Gruppen zusammengefaßte Dauer der Kundenverbindung mit dem Variablennamen DKV_G verwendet. Da es sich um eine mehrkategoriale Variable handelt, wird DKV_G als solche in SPSS festgelegt, wobei als Bezugskategorie die letzte Kategorie dient, d.h. die Gruppe derjenigen, die am längsten Kunden der Bank sind. Die Ergebnisse sind im SPSS-Output 4.8 angegeben.

SPSS-Output 4.8: Logistische Regression mit Warnung 1 und DKV gruppiert; Indikator-Codierung (Bezugskategorie = die letzte Kategorie)

			Parameter Coding				
	Value	Freq	(1)	(2)	(3)	(4)	(5)
DKV_G							
DKV <= 61 Monate	1	570	1,000	,000	,000	,000	,000
DKV = 62 - 105 Monate	2	570	,000	1,000	,000	,000	,000
DKV = 106 - 115 Monate	3	549	,000	,000	1,000	,000	,000
DKV = 116 - 117 Monate	4	1421	,000	,000	,000	1,000	,000
DKV = 118 - 132 Monate	5	213	,000	,000	,000	,000	1,000
DKV = 133 - 156 Monate	6	282	,000	,000	,000	,000	,000
DKV = 157 - 180 Monate	7	275	,000	,000	,000	,000	,000
DKV = 181 - 205 Monate	8	231	,000	,000	,000	,000	,000
DKV = 206 Monate und länger	9	1573	,000	,000	,000	,000	,000

		(6)	(7)	(8)
DKV_G				
DKV <= 61 Monate	1	,000	,000	,000
DKV = 62 - 105 Monate	2	,000	,000	,000
DKV = 106 - 115 Monate	3	,000	,000	,000
DKV = 116 - 117 Monate	4	,000	,000	,000
DKV = 118 - 132 Monate	5	,000	,000	,000
DKV = 133 - 156 Monate	6	1,000	,000	,000
DKV = 157 - 180 Monate	7	,000	1,000	,000
DKV = 181 - 205 Monate	8	,000	,000	1,000
DKV = 206 Monate und länger	9	,000	,000	,000

-2 Log Likelihood 7864,2757
* Constant is included in the model.

Beginning Block Number 1. Method: Enter
Variable(s) Entered on Step Number
1.. DKV_G DKV gruppiert

Estimation terminated at iteration number 3 because
parameter estimates changed by less than ,001

-2 Log Likelihood 7442,097
Cox & Snell - R^2 ,072
Nagelkerke - R^2 ,096

	Chi-Square	df	Significance
Model	422,178	8	,0000

Classification Table for WARN_1
The Cut Value is ,50

		Predicted		
		keine Warnung 1 k	Warnung 1 vorhan W	Percent Correct
Observed				
keine Warnung 1	k	1647	1343	55,08%
Warnung 1 vorhan	W	893	1801	66,85%
			Overall	60,66%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
DKV_G			388,6287	8	,0000	,2177	
DKV_G(1)	-,2827	,0984	8,2575	1	,0041	-,0282	,7537
DKV_G(2)	,3257	,0989	10,8523	1	,0010	,0336	1,3850
DKV_G(3)	-,2374	,0996	5,6839	1	,0171	-,0216	,7887
DKV_G(4)	-,9587	,0776	152,6048	1	,0000	-,1384	,3834
DKV_G(5)	,5881	,1521	14,9601	1	,0001	,0406	1,8007
DKV_G(6)	,6398	,1354	22,3231	1	,0000	,0508	1,8961
DKV_G(7)	,6336	,1368	21,4596	1	,0000	,0497	1,8844
DKV_G(8)	,9375	,1554	36,3839	1	,0000	,0661	2,5536
Constant	,0216	,0504	,1837	1	,6682		

Der Test mittels ΔD gibt an, daß alle Kontrast-Variablen, d.h. alle Gruppen der Variablen DKV_G, gemeinsam einen signifikanten Einfluß auf die abhängige Variable ausüben. Es ist aber von Interesse, ob jede einzelne dieser Gruppen auch einen solchen Einfluß auf das Auftreten von Warnung 1 ausübt. Nach den Ergebnissen des Wald-Tests ist dies positiv zu beantworten, weil die Regressionskoeffizienten für jede der acht Kontrast-Variablen zum vorgegebenen Signifikanzniveau verschieden von Null sind. Dennoch wird in diesem Modell mit 9,6 % nur relativ wenig von der Variation von Y erklärt.

Bezüglich des Risikos des Auftretens von Warnung 1 für die einzelnen Gruppen dieser Variablen im Vergleich zur Bezugskategorie kann kein Trend festgestellt werden. Um dies zu verdeutlichen, wird im SPSS-Output 4.9 ein Scatterplot der Gruppen von DKV_G gegen die geschätzten Wahrscheinlichkeiten für das Auftreten von Warnung 1 gezeigt. Anschließend ist in Tabelle 4.2 diese Erfolgswahrscheinlichkeit je Gruppe der Variablen DKV_G angegeben.

SPSS-Output 4.9: Scatterplot DKV_G gegen geschätzte Wahrscheinlichkeit ($\hat{\pi}_k$)

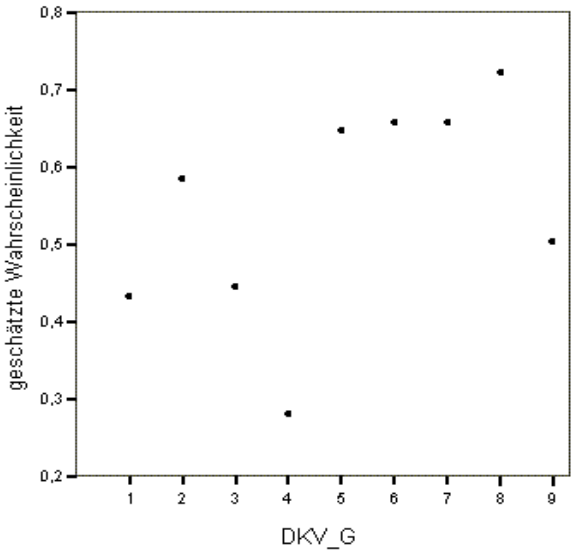


Tabelle 4.2: Codierung der DKV, Gruppierung der DKV und geschätzte Erfolgswahrscheinlichkeit $\hat{\pi}_k$ je Gruppe der DKV

Codierung der DKV	Gruppierung der DKV	$\hat{\pi}_k$
1	DKV <= 61 Monate	0,43509
2	DKV = 62 - 105 Monate	0,58596
3	DKV = 106 - 115 Monate	0,44627
4	DKV = 116 - 117 Monate	0,28149
5	DKV = 118 - 132 Monate	0,64789
6	DKV = 133 - 156 Monate	0,65957
7	DKV = 157 - 180 Monate	0,65818
8	DKV = 181 - 205 Monate	0,72294
9	DKV = 206 Monate und länger	0,50540

Die geschätzte Wahrscheinlichkeit für das Auftreten des interessierenden Ereignisses bzw. das Risiko für das Auftreten von erheblichen Zahlungsschwierigkeiten ist im Vergleich zur Bezugskategorie in den folgenden drei Gruppen:

- Dauer der Kundenverbindung ≤ 61 Monate
- Dauer der Kundenverbindung von 106 bis einschließlich 115 Monate
- Dauer der Kundenverbindung von 116 bis einschließlich 117 Monate

niedriger, in allen anderen Gruppen aber höher. Es fällt auch auf, daß die geschätzte Erfolgswahrscheinlichkeit für die Bezugskategorie sehr nahe bei 50 % liegt. Dies bedeutet, daß in der Stichprobe ungefähr die Hälfte der Kunden mit einer Dauer der Kundenverbindung von 206 Monaten und länger mindestens eine Warnung 1 erhalten haben. Daher ist es bei solchen Kunden schwierig, eine verläßliche Prognose zu machen, ob bei ihnen Warnung 1 auftritt oder nicht.

Zum 4. Kapitel kann abschließend gesagt werden, daß die behandelten einfachen Modelle insoweit nicht zufriedenstellend sind, als daß in jedem von ihnen zu wenig von der Variation der abhängigen Variablen erklärt wird. Dennoch sind sie nicht ohne Bedeutung, da sie den Einfluß je einer erklärenden Variablen auf das Auftreten der Warnung 1 darstellen. Somit verschaffen sie einen ersten Überblick und geben möglicherweise Ideen und Anregungen für die Auswertung der im nächsten Kapitel betrachteten multiplen logistischen Regressionen.

5. Multiple logistische Regression

In diesem Kapitel wird nach Modellen gesucht, die eher der Realität entsprechen, in der Zahlungsschwierigkeiten vermutlich nicht nur von einer Variablen abhängen. Zu diesem Zweck werden gleichzeitig mehrere erklärende Variablen in die Regression aufgenommen. Dabei handelt es sich bis auf weiteres um folgende Variablen:

SEX	Geschlecht umkodiert
REGION_1	Region (aus Risiko)
ALTER_G	Alter gruppiert
EINK	bekanntes Einkommen
VERMOEGE	Vermögen bei der Bank
DKV_G	Dauer der Kundenverbindung gruppiert.

Die Auswahl der Modelle aus der Vielzahl möglicher multipler logistischer Regressionen erfolgt nach zwei Kriterien. Einerseits wird angestrebt, daß möglichst alle Regressionskoeffizienten signifikant verschieden von Null sind und somit jede einzelne erklärende Variable einen signifikanten Einfluß auf das Auftreten von Warnung 1 ausübt. Andererseits soll eine bessere Anpassung an die Daten erreicht werden, d.h. die X-Variablen sollen gemeinsam möglichst viel von der Variation der abhängigen Variablen erklären. Hierbei wird schrittweise vorgegangen, indem zunächst das erste Ziel (signifikante Koeffizienten) und anschließend das zweite, d.h. ein möglichst hoher Wert des Nagelkerke $\widetilde{R}^2$, verfolgt wird.

Es wird an dieser Stelle darauf hingewiesen, daß in allen multiplen Modellen bis auf weiteres nicht alle 5.684 Fälle der ersten Stichprobe, sondern nur 4.852 einbezogen werden. Wie bereits erläutert, sind bei den Variablen REGION_1 der Wert "5" (Region unbekannt) und bei ALTER_G der Wert "20" (Alter unbekannt) als Missing-Werte definiert. Werden diese zwei X-Variablen bei der Schätzung des jeweiligen multiplen Modells verwendet, werden hierbei die betroffenen 832 Personen nicht berücksichtigt.

5.1. Ausgangsmodell (Methode Einschluß)

Als erstes Modell wird eine multiple logistische Regression betrachtet, bei der alle erklärenden Variablen in einem Schritt in das Modell aufgenommen werden. Hierzu wird die Methode Einschluß gewählt. Zwei der nichtbinären erklärenden Variablen werden zusätzlich als kategorial bestimmt. Das sind, wie bei den in den Abschnitten 4.3.2. bzw. 4.6. behandelten einfachen Regressionen, einerseits das in sechs Klassen gruppierte Alter mit dem Variablennamen ALTER_G, andererseits die in neun Gruppen eingeteilte Dauer der Kundenverbindung mit dem Variablennamen DKV_G. Als Bezugskategorie wird bei ALTER_G die erste Gruppe, die der unter 30-jährigen, gewählt. Bei DKV_G wird dagegen die letzte Kategorie, d.h. 206 Monate und länger, als Referenzkategorie bestimmt. Wegen der zusätzlichen Kontrast-Variablen (0;1-Variablen)[27] sind in diesem Modell nicht 6, sondern 17 erklärende Variablen enthalten. Der nachfolgende SPSS-Output 5.1 beinhaltet die wesentlichen Ergebnisse dieses Modells.

SPSS-Output 5.1: Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die erste Kategorie), bekanntes Einkommen, Vermögen bei der Bank und DKV gruppiert (Bezugskategorie = die letzte Kategorie); Methode Einschluß

```
Number of selected cases:                  5684
Number rejected because of missing data:    832
Number of cases included in the analysis: 4852
```

			Parameter Coding				
	Value	Freq	(1)	(2)	(3)	(4)	(5)
DKV_G							
DKV <= 61 Monate	1	494	1,000	,000	,000	,000	,000
DKV = 62 - 105 Monate	2	496	,000	1,000	,000	,000	,000
DKV = 106 - 115 Monate	3	453	,000	,000	1,000	,000	,000
DKV = 116 - 117 Monate	4	1192	,000	,000	,000	1,000	,000
DKV = 118 - 132 Monate	5	197	,000	,000	,000	,000	1,000
DKV = 133 - 156 Monate	6	254	,000	,000	,000	,000	,000
DKV = 157 - 180 Monate	7	247	,000	,000	,000	,000	,000
DKV = 181 - 205 Monate	8	207	,000	,000	,000	,000	,000
DKV = 206 Monate und länger	9	1312	,000	,000	,000	,000	,000

		(6)	(7)	(8)
DKV_G				
DKV <= 61 Monate	1	,000	,000	,000
DKV = 62 - 105 Monate	2	,000	,000	,000
DKV = 106 - 115 Monate	3	,000	,000	,000
DKV = 116 - 117 Monate	4	,000	,000	,000
DKV = 118 - 132 Monate	5	,000	,000	,000
DKV = 133 - 156 Monate	6	1,000	,000	,000
DKV = 157 - 180 Monate	7	,000	1,000	,000
DKV = 181 - 205 Monate	8	,000	,000	1,000
DKV = 206 Monate und länger	9	,000	,000	,000

[27] Die Kontrast-Variablen sind hier 0;1-Variablen, da auch bei den multiplen Modellen ausschließlich die Indikator-Codierung angewendet wird.

	Value	Freq	Parameter Coding (1)	(2)	(3)	(4)	(5)
ALTER_G							
0-29 Lebensjahre	1	1529	,000	,000	,000	,000	,000
30-39 Lebensjahre	2	1374	1,000	,000	,000	,000	,000
40-49 Lebensjahre	3	770	,000	1,000	,000	,000	,000
50-59 Lebensjahre	4	504	,000	,000	1,000	,000	,000
60-69 Lebensjahre	5	336	,000	,000	,000	1,000	,000
70 Lebensjahre und älter	6	339	,000	,000	,000	,000	1,000

-2 Log Likelihood 6726,1609
* Constant is included in the model.

Beginning Block Number 1. Method: Enter
Variable(s) Entered on Step Number
1.. SEX Geschl_umkodiert
REGION_1 Region (aus Risiko)
ALTER_G Alter gruppiert
EINK bekanntes Einkommen
VERMOEGE Vermögen bei der Bank
DKV_G DKV gruppiert

Estimation terminated at iteration number 7 because
parameter estimates changed by less than ,001

-2 Log Likelihood 4708,837
Cox & Snell - R^2 ,340
Nagelkerke - R^2 ,454

	Chi-Square	df	Significance
Model	2017,324	17	,0000

Classification Table for WARN_1
The Cut Value is ,50

Observed		Predicted: keine Warnung 1 k	Predicted: Warnung 1 vorhan W	Percent Correct
keine Warnung 1	k	1643	770	68,09%
Warnung 1 vorhan	W	345	2094	85,85%
			Overall	77,02%

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
SEX	-,7693	,0723	113,3576	1	,0000	-,1287	,4633
REGION_1	,1281	,1080	1,4073	1	,2355	,0000	1,1367
ALTER_G			197,1700	5	,0000	,1668	
ALTER_G(1)	,8408	,0901	87,0915	1	,0000	,1125	2,3182
ALTER_G(2)	,6748	,1090	38,3277	1	,0000	,0735	1,9637
ALTER_G(3)	,2137	,1267	2,8456	1	,0916	,0112	1,2382
ALTER_G(4)	-,4803	,1768	7,3772	1	,0066	-,0283	,6186
ALTER_G(5)	-1,3888	,2190	40,2145	1	,0000	-,0754	,2494
EINK	-,1760	,0131	181,7234	1	,0000	-,1635	,8386
VERMOEGE	-,0001	1,168E-05	90,4980	1	,0000	-,1147	,9999
DKV_G			95,3427	8	,0000	,1086	
DKV_G(1)	-,6713	,1307	26,3808	1	,0000	-,0602	,5111
DKV_G(2)	,0400	,1384	,0834	1	,7727	,0000	1,0408
DKV_G(3)	-,0948	,1613	,3454	1	,5567	,0000	,9096
DKV_G(4)	-,8244	,1425	33,4596	1	,0000	-,0684	,4385
DKV_G(5)	,1890	,1921	,9676	1	,3253	,0000	1,2080
DKV_G(6)	,1999	,1692	1,3971	1	,2372	,0000	1,2213
DKV_G(7)	,2587	,1732	2,2305	1	,1353	,0059	1,2952
DKV_G(8)	,3957	,1965	4,0562	1	,0440	,0175	1,4854
Constant	,7806	,1452	28,9141	1	,0000		

Die Klassifikationstabelle gibt an, daß das Modell 77,02 % aller Fälle bezüglich des Auftretens bzw. des Nichtauftretens des Ereignisses richtig zugeordnet hat. Nach dem Nagelkerke $\hat{R}^2$ erklären die X-Variablen gemeinsam 45,4 % der Variation der abhängigen Variablen. Der Wert des Nagelkerke $\hat{R}^2$ ist hier im Vergleich zu den im 4. Kapitel behandelten einfachen logistischen Regressionen deutlich höher und veranschaulicht somit den Vorteil der multiplen Regression gegenüber den einfachen Modellen mit nur einer erklärenden Variablen. Allerdings können Wechselwirkungen zwischen den unabhängigen Variablen zu Problemen bezüglich der Schätzung und der Interpretation der Ergebnisse in multiplen Modellen führen. Daher wird auf diese Problematik später gesondert eingegangen.

Mittels der Reduktion in der Deviance ΔD wird bei multiplen Modellen getestet, ob alle Regressionskoeffizienten, mit Ausnahme der Konstante, gleich Null sind. Zum vorgegebenen Signifikanzniveau von 0,05 wird diese Hypothese hier abgelehnt. Somit üben alle 17 im Modell enthaltenen unabhängigen Variablen gemeinsam einen signifikanten Einfluß auf das Auftreten von Warnung 1 aus. Dies bedeutet jedoch nicht, daß auch jede einzelne X-Variable wesentlich zur Erklärung der Variation von Y beiträgt. Aufschluß hierzu geben die Ergebnisse des Wald-Tests für die einzelnen Regressionskoeffizienten. Für dieses Modell wird festgestellt, daß die Regressionskoeffizienten der Variablen Geschlecht, bekanntes Einkommen und Vermögen bei der Bank signifikant verschieden von Null sind, der Koeffizient von Region dagegen nicht.

Für eine als kategorial definierte Variable, soweit sie mehr als zwei Kategorien hat, wird neben der Signifikanz der einzelnen Kategorien auch die der Variablen als Ganzes überprüft. Dem SPSS-Output 5.1 ist bei "Variables in the Equation" unter "Sig" für ALTER_G und DKV_G zu entnehmen, daß das Alter und die Dauer der Kundenverbindung einen signifikanten Einfluß auf die abhängige Variable ausüben. Allerdings trifft das nicht für jede einzelne Kategorie zu. Bei der Dauer der Kundenverbindung ist die Signifikanz nur auf drei Gruppen (DKV_G(1), DKV_G(4) und DKV_G(8)) zurückzuführen, bei dem Alter auf vier von insgesamt fünf Gruppen (ALTER_G(1), ALTER_G(2), ALTER_G(4) und ALTER_G(5)). Bei den übrigen Kategorien (DKV_G(2), DKV_G(3), DKV_G(5), DKV_G(6), DKV_G(7)

und ALTER_G(3)) kann die Nullhypothese $\beta_j = 0$ auf dem vorgegebenen Signifikanzniveau nicht abgelehnt werden.[28]

Es ist also festzuhalten, daß die Koeffizienten von 7 aller 17 Variablen nicht signifikant zum vorgegebenen Niveau von 0,05 sind. Daraus ergeben sich folgende Überlegungen, die in den weiteren Abschnitten näher betrachtet werden: Zum einen ist zu untersuchen, ob alle bzw. wenn nicht, welche Variablen bei einem schrittweisen Vorgehen in die Regression aufgenommen werden. Vermutlich wird z.B. REGION_1 wegen des nichtsignifikanten Regressionskoeffizienten zu den Variablen gehören, die bei der Modellierung nicht berücksichtigt werden. Bei der gewählten Methode Einschluß besteht nicht die Möglichkeit, Faktoren, die als X-Variablen gewählt wurden, im Zuge der Schätzung aus dem Modell auszuschließen. Es muß also eine andere Methode, z.B. die unter 5.2. verwendete Vorwärts (Wald), gewählt werden, um eine Antwort auf diese Frage zu finden. Zum anderen ist in bezug auf die kategorialen Variablen ALTER_G und DKV_G der Frage nachzugehen, wie die Signifikanz möglichst aller Kategorien erreicht werden kann. Eine andere Gruppierung kann zu dem gewünschten Ergebnis führen. Die passende Gruppierung zu finden, kann zeitaufwendig sein, da dies mit wiederholtem Testen verbunden ist. Hilfreich kann aber auch eine Veränderung der Bezugskategorie sein, wobei sich dann die Interpretation entsprechend ändert. Im Abschnitt 5.3. wird die zweite Möglichkeit gewählt, um zu prüfen, ob eine Signifikanzverbesserung der einzelnen Kategorien von Alter und Dauer der Kundenverbindung auf diesem einfachen Weg zu erreichen ist.

Neben der Signifikanzüberprüfung ist die Interpretation der Regressionskoeffizienten von besonderem Interesse. Wie bereits erläutert, gibt β_j die Veränderung in den log odds an, wenn die Variable X_j um eine Einheit erhöht wird und alle anderen erklärenden Variablen konstant bleiben. Somit bedeutet zum Beispiel der Koeffizient der Variablen SEX mit einem Wert von -0,7693, daß die log odds bei den Frauen um 0,7693 niedriger sind als bei den Männern. Daher ist die Wahrscheinlichkeit für das Auftreten von Zahlungsschwierigkeiten bei den männlichen Kunden c.p. höher als bei den weiblichen. Weiter zeigen die Ergebnisse, daß die log odds und damit das Risiko des Auftretens von Warnung 1 mit dem Erhöhen des Einkommens bzw. des Vermögens c.p. sinken, weil die dazugehörigen Koeffizienten negativ sind. Ähnliches wurde in den entsprechenden einfachen Modellen

[28] Hierbei wird die jeweilige Referenzkategorie nicht betrachtet, weil durch die Indikator-Codierung deren Regressionskoeffizient automatisch Null gesetzt wird.

festgestellt, da auch dort die Koeffizienten der Variablen SEX, EINK und VERMOEGE negativ waren. In bezug auf die Region zeigt sich erneut, daß für die im Westen wohnenden Kunden bei sonst gleichen Merkmalen eine höhere Wahrscheinlichkeit für das Auftreten des Ereignisses geschätzt wird als bei den im Osten wohnenden.

Bei den zwei letzten Altersklassen, der 60- bis unter 70-jährigen (ALTER_G(4)) sowie der 70-jährigen und älteren Kunden (ALTER_G(5)), sind die Koeffizienten ebenfalls negativ. Somit sind diese Gruppen im Vergleich zur Referenzkategorie mit sinkenden log odds verbunden, wobei die log odds bei ALTER_G(5) stärker sinken als bei ALTER_G(4). Dies bedeutet, daß bei den Kunden, die 60-jährig oder älter sind, im Vergleich zu den Kunden unter 30 Jahren Warnung 1 und damit Zahlungsschwierigkeiten seltener auftreten, wobei ab einem Alter von 70 Jahren das Risiko noch einmal deutlich zurückgeht. Die Regressionskoeffizienten der übrigen drei Altersklassen sind positiv. Daher sind die log odds sowie die Wahrscheinlichkeit für das Auftreten einer Warnung 1 bei diesen Gruppen im Vergleich zur Bezugskategorie höher. Das Auftreten des Ereignisses ist in der Gruppe der 30- bis unter 40-jährigen 2,3-mal (siehe Exp(B)) wahrscheinlicher, in der Gruppe der 40- bis unter 50-jährigen ca. 2-mal wahrscheinlicher und in der Gruppe der 50- bis unter 60-jährigen 1,2-mal wahrscheinlicher als in der Gruppe der bis unter 30-jährigen Kunden.

Bei der Dauer der Kundenverbindung, der zweiten nichtbinären kategorialen Variablen in diesem Modell, ist das Risiko des Auftretens des interessierenden Ereignisses bei drei Kategorien, d.h. DKV bis einschließlich 61 Monate (DKV_G(1)), zwischen 106 und 115 Monaten (DKV_G(3)) bzw. von 116 bis einschließlich 117 Monaten (DKV_G(4)), geringer, bei den übrigen Kategorien dagegen höher als bei der Referenzkategorie, d.h. als bei denjenigen Personen, die am längsten Kunden sind. Allerdings ist das Auftreten von Warnung 1 in den fünf Kategorien, in denen dies häufiger als bei der Referenzkategorie geschieht, von 1,04-mal bis nur maximal 1,49-mal wahrscheinlicher. Daher ist in diesen Gruppen von einem eher geringfügig höheren Risiko auszugehen.

In bezug auf die Werte der geschätzten Regressionskoeffizienten fällt besonders auf, daß der Koeffizient der Variablen VERMOEGE mit einem Wert von -0,0001 absolut gesehen sehr gering ist. Ein ähnlich niedriger Koeffizient wurde für diese Variable auch im einfachen Modell des Abschnittes 4.5. ermittelt. $b_{vermoege}$ gibt die Veränderung in den log odds bei Erhöhung des Vermögens um einen Pfennig an. Somit erklären die pfenniggenauen

Angaben, aber auch die hohen Ausprägungen dieser Variablen (fünf- bis siebenstellige Beträge) den geringen Absolutwert des Koeffizienten.[29]

Es ist sicherlich von Interesse, die Regressionskoeffizienten dieses multiplen Modells mit den Regressionskoeffizienten der im 4. Kapitel behandelten einfachen Modelle zu vergleichen, um festzustellen, ob sich größere Veränderungen ergeben haben und bei welchen Variablen dies der Fall ist. Information darüber gibt Tabelle 5.1.

Tabelle 5.1: Vergleich der geschätzten Regressionskoeffizienten der einfachen Modelle des 4. Kapitels mit denjenigen des multiplen Modells im SPSS-Output 5.1

	Regressions-koeffizienten der einfachen Modelle im 4. Kapitel	Regressions-koeffizienten im multiplen Modell unter SPSS-Output 5.1	Absolute Differenz	Differenz in Prozent
SEX	- 0,9568	- 0,7693	0,1875	19,60
REGION_1	0,7240	0,1281*	- 0,5959	- 82,31
ALTER_G(1)	0,7246	0,8408	0,1162	16,04
ALTER_G(2)	0,3162	0,6748	0,3586	113,41
ALTER_G(3)	- 0,2256	0,2137*	0,4393	194,73
ALTER_G(4)	- 1,2324	- 0,4803	0,7521	61,03
ALTER_G(5)	- 2,1960	- 1,3888	0,8072	36,76
EINK	- 0,2143	- 0,1760	0,0383	17,87
VERMOEGE	- 0,0002	- 0,0001	0,0001	50,00
DKV_G(1)	- 0,2827	- 0,6713	- 0,3886	- 137,46
DKV_G(2)	0,3257	0,0400*	- 0,2857	- 87,72
DKV_G(3)	- 0,2374	- 0,0948*	0,1426	60,07
DKV_G(4)	- 0,9587	- 0,8244	0,1343	14,01
DKV_G(5)	0,5881	0,1890*	- 0,3991	- 67,86
DKV_G(6)	0,6398	0,1999*	- 0,4399	- 68,76
DKV_G(7)	0,6336	0,2587*	- 0,3749	- 59,17
DKV_G(8)	0,9375	0,3957	- 0,5418	- 57,79

* Dieser Koeffizient ist nicht signifikant.

Der Vergleich zeigt, daß bei den meisten Regressionskoeffizienten, gemessen an dem jeweiligen Wert, erhebliche Veränderungen eingetreten sind. Die größte Veränderung betrifft die Altersklasse der 50- bis unter 60-jährigen (ALTER_G(3)). Sie beträgt fast 200 % und ist sogar mit einem Vorzeichenwechsel verbunden. Die großen Differenzen zwischen den Regressionskoeffizienten in den einfachen Modellen und in diesem multiplen Modell deuten

[29] Dies wurde überprüft, indem das Vermögen analog dem Einkommen auf 100 DM gerundet (d.h. VERMOEGE wurde durch 100 geteilt) und die Schätzung noch einmal durchgeführt wurde. Als Ergebnis wurde dann ein Koeffizient für VERMOEGE in Höhe von -0,0111 geschätzt. Sein Absolutwert stieg also um das 111-fache und damit in etwa gleich wie die Erfassungseinheit.

auf Wechselwirkungen zwischen den erklärenden Variablen hin und zeigen, daß einfache Regressionen bei der hier behandelten Problemstellung nur einen ersten Überblick vermitteln können. Nur multiple Modelle können aber der komplexen Realität näher kommen. Daher wird die Suche nach einem adäquaten Modell unter den multiplen Regressionen fortgesetzt. Allerdings ist mit einem schrittweisen Vorgehen zu überprüfen, ob alle erklärenden Variablen in das Modell aufgenommen werden oder z.B. starke Wechselwirkungen dies verhindern.

Nach diesen umfangreichen Erläuterungen und Interpretationen bezüglich der Regressionskoeffizienten darf natürlich nicht vergessen werden, daß das letztendliche Ziel einer solchen Schätzung darin liegt, mit ihrer Hilfe die Erfolgswahrscheinlichkeit $\hat{\pi}_k$ zu bestimmen. Diese kann für jeden einzelnen Kunden in folgender Weise berechnet werden:

$$\hat{\pi}_k = \frac{1}{1+e^{-[\,0{,}7806 - 0{,}7693*(\text{sex}) + 0{,}1281*(\text{region\_1}) + 0{,}8408*(\text{alter\_g(1)}) + \ldots + 0{,}3957*(\text{dkv\_g(8)})\,]}} \quad .$$

So ergibt sich zum Beispiel die Wahrscheinlichkeit für das Auftreten von Warnung 1 für eine Person mit den Merkmalen: männlich, wohnhaft in den neuen Bundesländern, 45 Jahre alt (hier ALTER_G(2)), kein bekanntes Einkommen, kein Vermögen bei der Bank und seit 117 Monaten Kunde dieser Bank (DKV_G(4)), in Höhe von 0,65. Da $\hat{\pi}_k \geq 0{,}5$ ist, wird anhand der Schätzung das Auftreten der Warnung 1 vermutet, welche bei diesem Kunden tatsächlich beobachtet wurde.

Ein weiterer wichtiger Punkt bei multiplen Modellen ist der Beitrag jeder einzelnen erklärenden Variablen zur logistischen Regression. Dies zu bestimmen ist deshalb nicht einfach, weil dieser Beitrag von den anderen erklärenden Variablen abhängt. Hierzu wird die R-Statistik verwendet, welche die partielle Korrelation zwischen der abhängigen Variablen und jeder einzelnen erklärenden Variablen angibt.[30] Diese Kennzahl kann einen Wert zwischen -1 und +1 annehmen. Wenn sie einen positiven Wert annimmt, gibt dies an, daß mit steigenden Werten dieser X-Variablen auch die Likelihood des Ereignisses, d.h. die Wahrscheinlichkeit für das Auftreten des interessierenden Ereignisses, steigt. Umgekehrt sinkt mit steigenden Werten der X-Variablen die Likelihood des Ereignisses, wenn R negativ ist. Das Vorzeichen der R-Statistik entspricht also dem Vorzeichen des jeweiligen

[30] Vgl. SPSS Regression Models™ 9.0, S. 39 f.

Regressionskoeffizienten. Somit beinhaltet nur der Absolutwert der R-Statistik zusätzliche Informationen. Große Absolutwerte deuten auf einen großen partiellen Beitrag der Variablen zu dem Modell hin. Diese Kennzahl wird nach folgender Formel berechnet:

$$R = \pm \sqrt{\left(\frac{(Wald - Statisitk) - 2 * K}{-2 * LL_0} \right)},$$

worin K die Anzahl der Freiheitsgrade der Wald-Statistik ist. Ist der Wert der Wald-Statistik geringer als 2*K, so wird die R-Statistik gleich Null gesetzt. Dies ist hier z.B. bei der Variablen REGION_1 der Fall. Anhand der R-Statistik kann in dieser multiplen logistischen Regression festgestellt werden, daß die Variable EINK (bekanntes Einkommen) den höchsten partiellen Beitrag zum Modell leistet, gefolgt von den Variablen SEX, VERMOEGE usw. Die genaue Reihenfolge kann der Tabelle 5.2 entnommen werden, in der alle 17 erklärenden Variablen absteigend nach der Höhe des Absolutwertes der R-Statistik geordnet sind.

Tabelle 5.2: Variablennamen, Wert der R-Statistik und unter "Sig" angegebenes Ergebnis des Wald-Tests je Variable, absteigend sortiert nach der Höhe des Absolutwertes der R-Statistik

	Variablenname	R-Statistik	Sig (Wald-Test)
1	EINK	- 0,1635	0,0000
2	SEX	- 0,1287	0,0000
3	VERMOEGE	- 0,1147	0,0000
4	ALTER_G(1)	0,1125	0,0000
5	ALTER_G(5)	- 0,0754	0,0000
6	ALTER_G(2)	0,0735	0,0000
7	DKV_G(4)	- 0,0684	0,0000
8	DKV_G(1)	- 0,0602	0,0000
9	ALTER_G(4)	- 0,0283	0,0066
10	DKV_G(8)	0,0175	0,0440
11	ALTER_G(3)	0,0112	**0,0916***
12	DKV_G(7)	0,0059	**0,1353**
13	REGION_1	0,0000	**0,2355**
14	DKV_G(2)	0,0000	**0,7727**
15	DKV_G(3)	0,0000	**0,5567**
16	DKV_G(5)	0,0000	**0,3253**
17	DKV_G(6)	0,0000	**0,2372**

* Fett markiert unter "Sig (Wald-Test)" sind diejenigen Werte, die höher als das vorgegebene Signifikanzniveau von 0,05 sind. Somit sind die zu diesen Variablen gehörenden Regressionskoeffizienten nicht signifikant verschieden von Null.

5.2. Ausgangsmodell (Methode Vorwärts)

Als zweite multiple logistische Regression wird ein Modell betrachtet, das dem Modell unter 5.1. bis auf einen Unterschied entspricht. Es wird an Stelle von Einschluß die Methode Vorwärts (Wald) gewählt. Dadurch werden die X-Variablen nicht in einem Schritt, sondern schrittweise nach Überprüfung bestimmter Kriterien in das Modell aufgenommen und möglicherweise wieder herausgenommen. Das Aufnahmekriterium bezieht sich auf die als Score angegebene Kennzahl.[31] Es wird diejenige Variable als erste, zweite usw. in das Modell aufgenommen, die auf der entsprechenden Stufe den höchsten Wert bei dieser Kennzahl aufweist, vorausgesetzt das dazugehörige Signifikanzniveau ist geringer als 0,05. Eine bereits aufgenommene Variable kann wieder ausgeschlossen werden, wenn das Signifikanzniveau der Wald-Statistik, unter "Variables in the Equation" angegeben, größer als 0,1 wird.[32]

SPSS-Output 5.2: Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die erste Kategorie), bekanntes Einkommen, Vermögen bei der Bank und DKV gruppiert (Bezugskategorie = die letzte Kategorie); Methode Vorwärts (Wald)

```
Number of selected cases:                  5684
Number rejected because of missing data:    832
Number of cases included in the analysis: 4852
```

	Value	Freq	Parameter Coding (1)	(2)	(3)	(4)	(5)
DKV_G							
DKV <= 61 Monate	1	494	1,000	,000	,000	,000	,000
DKV = 62 - 105 Monate	2	496	,000	1,000	,000	,000	,000
DKV = 106 - 115 Monate	3	453	,000	,000	1,000	,000	,000
DKV = 116 - 117 Monate	4	1192	,000	,000	,000	1,000	,000
DKV = 118 - 132 Monate	5	197	,000	,000	,000	,000	1,000
DKV = 133 - 156 Monate	6	254	,000	,000	,000	,000	,000
DKV = 157 - 180 Monate	7	247	,000	,000	,000	,000	,000
DKV = 181 - 205 Monate	8	207	,000	,000	,000	,000	,000
DKV = 206 Monate und länger	9	1312	,000	,000	,000	,000	,000

		(6)	(7)	(8)
DKV_G				
DKV <= 61 Monate	1	,000	,000	,000
DKV = 62 - 105 Monate	2	,000	,000	,000
DKV = 106 - 115 Monate	3	,000	,000	,000
DKV = 116 - 117 Monate	4	,000	,000	,000
DKV = 118 - 132 Monate	5	,000	,000	,000
DKV = 133 - 156 Monate	6	1,000	,000	,000
DKV = 157 - 180 Monate	7	,000	1,000	,000
DKV = 181 - 205 Monate	8	,000	,000	1,000
DKV = 206 Monate und länger	9	,000	,000	,000

[31] Siehe SPSS-Output 5.2 unter "Variables not in the Equation".

[32] Sowohl bei dem Aufnahme- wie auch bei dem Ausschlußkriterium sind die hier genannten Werte von 0,05 bzw. von 0,1 Voreinstellungen in SPSS, die in dieser Arbeit beibehalten werden.

	Value	Freq	Parameter Coding (1)	(2)	(3)	(4)	(5)
ALTER_G							
0-29 Lebensjahre	1	1529	,000	,000	,000	,000	,000
30-39 Lebensjahre	2	1374	1,000	,000	,000	,000	,000
40-49 Lebensjahre	3	770	,000	1,000	,000	,000	,000
50-59 Lebensjahre	4	504	,000	,000	1,000	,000	,000
60-69 Lebensjahre	5	336	,000	,000	,000	1,000	,000
70 Lebensjahre und älter	6	339	,000	,000	,000	,000	1,000

-2 Log Likelihood 6726,1609
* Constant is included in the model.

Estimation terminated at iteration number 1 because
Log Likelihood decreased by less than ,01 percent.

Classification Table for WARN_1
The Cut Value is ,50

		Predicted: keine Warnung 1 k	Predicted: Warnung 1 vorhan W	Percent Correct
Observed				
keine Warnung 1	k	0	2413	,00%
Warnung 1 vorhan	W	0	2439	100,00%
			Overall	50,27%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Constant	,0107	,0287	,1393	1	,7090		

Beginning Block Number 1. Method: Forward Stepwise (WALD)

---------- Variables not in the Equation ----------
Residual Chi Square not computed because of redundancies.

Variable	Score	df	Sig	R
SEX	272,2224	1	,0000	,2004
REGION_1	174,0442	1	,0000	,1599
ALTER_G	517,5563	5	,0000	,2747
ALTER_G(1)	213,5776	1	,0000	,1774
ALTER_G(2)	20,0179	1	,0000	,0518
ALTER_G(3)	8,7048	1	,0032	,0316
ALTER_G(4)	117,6358	1	,0000	,1311
ALTER_G(5)	246,5502	1	,0000	,1907
EINK	716,4679	1	,0000	,3259
VERMOEGE	178,8207	1	,0000	,1621
DKV_G	355,2549	8	,0000	,2246
DKV_G(1)	7,2323	1	,0072	,0279
DKV_G(2)	22,1631	1	,0000	,0548
DKV_G(3)	,1343	1	,7140	,0000
DKV_G(4)	265,2802	1	,0000	,1978
DKV_G(5)	21,6348	1	,0000	,0540
DKV_G(6)	32,6422	1	,0000	,0675
DKV_G(7)	25,7386	1	,0000	,0594
DKV_G(8)	40,7763	1	,0000	,0759

Variable(s) Entered on Step Number
1.. EINK bekanntes Einkommen

Estimation terminated at iteration number 5 because
Log Likelihood decreased by less than ,01 percent.

-2 Log Likelihood 5629,170
Cox & Snell - R^2 ,202
Nagelkerke - R^2 ,270

	Chi-Square	df	Significance
Model	1096,991	1	,0000

Classification Table for WARN_1
The Cut Value is ,50

Observed		Predicted: keine Warnung 1 k	Predicted: Warnung 1 vorhan W	Percent Correct
keine Warnung 1	k	888	1525	36,80%
Warnung 1 vorhan	W	82	2357	96,64%
			Overall	66,88%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
EINK	-,2159	,0129	278,5182	1	,0000	-,2028	,8058
Constant	,4358	,0326	178,2728	1	,0000		

---------- Variables not in the Equation ----------
Residual Chi Square not computed because of redundancies.

Variable	Score	df	Sig	R
SEX	173,7357	1	,0000	,1598
REGION_1	76,1157	1	,0000	,1050
ALTER_G	358,0847	5	,0000	,2275
ALTER_G(1)	160,6131	1	,0000	,1536
ALTER_G(2)	27,8261	1	,0000	,0620
ALTER_G(3)	5,3613	1	,0206	,0224
ALTER_G(4)	48,1110	1	,0000	,0828
ALTER_G(5)	152,6148	1	,0000	,1496
VERMOEGE	92,8030	1	,0000	,1162
DKV_G	209,3875	8	,0000	,1696
DKV_G(1)	14,8160	1	,0001	,0437
DKV_G(2)	9,9671	1	,0016	,0344
DKV_G(3)	1,4177	1	,2338	,0000
DKV_G(4)	147,7993	1	,0000	,1472
DKV_G(5)	12,0418	1	,0005	,0386
DKV_G(6)	15,4398	1	,0001	,0447
DKV_G(7)	12,4610	1	,0004	,0394
DKV_G(8)	25,2104	1	,0000	,0587

Variable(s) Entered on Step Number
2.. ALTER_G Alter gruppiert

Estimation terminated at iteration number 6 because
parameter estimates changed by less than ,001

-2 Log Likelihood 5261,158
Cox & Snell - R^2 ,261
Nagelkerke - R^2 ,347

	Chi-Square	df	Significance
Model	1465,003	6	,0000

Classification Table for WARN_1
The Cut Value is ,50

Observed		Predicted: keine Warnung 1 k	Predicted: Warnung 1 vorhan W	Percent Correct
keine Warnung 1	k	1164	1249	48,24%
Warnung 1 vorhan	W	196	2243	91,96%
			Overall	70,22%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
ALTER_G			309,0812	5	,0000	,2109	
ALTER_G(1)	,9299	,0836	123,7932	1	,0000	,1346	2,5341
ALTER_G(2)	,6670	,1001	44,4253	1	,0000	,0794	1,9484
ALTER_G(3)	,0413	,1126	,1343	1	,7140	,0000	1,0421
ALTER_G(4)	-,6965	,1554	20,0921	1	,0000	-,0519	,4983
ALTER_G(5)	-1,7658	,2035	75,2653	1	,0000	-,1044	,1711
EINK	-,2074	,0133	243,2520	1	,0000	-,1894	,8127
Constant	,1726	,0536	10,3469	1	,0013		

---------- Variables not in the Equation ----------

Residual Chi Square not computed because of redundancies.

Variable	Score	df	Sig	R
SEX	131,3739	1	,0000	,1387
REGION_1	76,6343	1	,0000	,1053
VERMOEGE	60,2753	1	,0000	,0931
DKV_G	184,5657	8	,0000	,1583
DKV_G(1)	17,6636	1	,0000	,0483
DKV_G(2)	6,3741	1	,0116	,0255
DKV_G(3)	,6152	1	,4328	,0000
DKV_G(4)	127,0487	1	,0000	,1364
DKV_G(5)	7,3009	1	,0069	,0281
DKV_G(6)	13,5582	1	,0002	,0415
DKV_G(7)	10,5718	1	,0011	,0357
DKV_G(8)	19,1054	1	,0000	,0504

Variable(s) Entered on Step Number
3.. DKV_G DKV gruppiert

Estimation terminated at iteration number 6 because
parameter estimates changed by less than ,001

-2 Log Likelihood	5077,138
Cox & Snell - R^2	,288
Nagelkerke - R^2	,384

	Chi-Square	df	Significance
Model	1649,023	14	,0000

Classification Table for WARN_1
The Cut Value is ,50

Observed		Predicted: keine Warnung 1 k	Predicted: Warnung 1 vorhan W	Percent Correct
keine Warnung 1	k	1569	844	65,02%
Warnung 1 vorhan	W	418	2021	82,86%
			Overall	73,99%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
ALTER_G			285,0371	5	,0000	,2022	
ALTER_G(1)	,8172	,0867	88,8761	1	,0000	,1136	2,2641
ALTER_G(2)	,5916	,1043	32,1713	1	,0000	,0670	1,8069
ALTER_G(3)	-,0055	,1183	,0022	1	,9627	,0000	,9945
ALTER_G(4)	-,8202	,1618	25,6831	1	,0000	-,0593	,4403
ALTER_G(5)	-1,9168	,2092	83,9084	1	,0000	-,1104	,1471
EINK	-,1932	,0130	220,3489	1	,0000	-,1802	,8243
DKV_G			178,8125	8	,0000	,1556	
DKV_G(1)	-,6566	,1218	29,0593	1	,0000	-,0634	,5186
DKV_G(2)	-,0221	,1243	,0317	1	,8586	,0000	,9781
DKV_G(3)	-,1715	,1315	1,7027	1	,1919	,0000	,8424
DKV_G(4)	-,9605	,0996	92,9967	1	,0000	-,1163	,3827
DKV_G(5)	,1891	,1821	1,0780	1	,2991	,0000	1,2081
DKV_G(6)	,2492	,1642	2,3025	1	,1292	,0067	1,2830
DKV_G(7)	,2008	,1649	1,4828	1	,2233	,0000	1,2223
DKV_G(8)	,4704	,1885	6,2266	1	,0126	,0251	1,6006
Constant	,4802	,0891	29,0365	1	,0000		

---------- Variables not in the Equation ----------
Residual Chi Square not computed because of redundancies.

Variable	Score	df	Sig	R
SEX	124,5882	1	,0000	,1350
REGION_1	1,1604	1	,2814	,0000
VERMOEGE	59,2269	1	,0000	,0922

Variable(s) Entered on Step Number
4.. SEX Geschl_umkodiert

Estimation terminated at iteration number 6 because
parameter estimates changed by less than ,001

-2 Log Likelihood	4953,592
Cox & Snell - R^2	,306
Nagelkerke - R^2	,408

	Chi-Square	df	Significance
Model	1772,569	15	,0000

Classification Table for WARN_1
The Cut Value is ,50

Observed		Predicted: keine Warnung 1 k	Predicted: Warnung 1 vorhan W	Percent Correct
keine Warnung 1	k	1605	808	66,51%
Warnung 1 vorhan	W	485	1954	80,11%
			Overall	73,35%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
SEX	-,7765	,0702	122,3502	1	,0000	-,1338	,4600
ALTER_G			260,2668	5	,0000	,1929	
ALTER_G(1)	,7917	,0881	80,7898	1	,0000	,1082	2,2072
ALTER_G(2)	,5930	,1060	31,3082	1	,0000	,0660	1,8094
ALTER_G(3)	-,0270	,1202	,0504	1	,8223	,0000	,9734
ALTER_G(4)	-,8717	,1647	27,9970	1	,0000	-,0622	,4183
ALTER_G(5)	-1,7608	,2121	68,9370	1	,0000	-,0998	,1719
EINK	-,1854	,0128	210,2123	1	,0000	-,1759	,8307
DKV_G			172,2710	8	,0000	,1524	
DKV_G(1)	-,6846	,1240	30,5010	1	,0000	-,0651	,5043
DKV_G(2)	-,0512	,1265	,1641	1	,6854	,0000	,9501
DKV_G(3)	-,2378	,1335	3,1702	1	,0750	-,0132	,7884
DKV_G(4)	-,9643	,1012	90,7222	1	,0000	-,1149	,3812
DKV_G(5)	,1947	,1857	1,0985	1	,2946	,0000	1,2149
DKV_G(6)	,2396	,1669	2,0600	1	,1512	,0030	1,2707
DKV_G(7)	,2160	,1673	1,6677	1	,1966	,0000	1,2411
DKV_G(8)	,4242	,1905	4,9599	1	,0259	,0210	1,5284
Constant	,8185	,0960	72,6615	1	,0000		

---------- Variables not in the Equation ----------
Residual Chi Square not computed because of redundancies.

Variable	Score	df	Sig	R
REGION_1	2,2699	1	,1319	,0063
VERMOEGE	55,9993	1	,0000	,0896

Variable(s) Entered on Step Number
5.. VERMOEGE Vermögen bei der Bank

Estimation terminated at iteration number 7 because
parameter estimates changed by less than ,001

-2 Log Likelihood	4710,241
Cox & Snell - R^2	,340
Nagelkerke - R^2	,453

	Chi-Square	df	Significance
Model	2015,920	16	,0000

Classification Table for WARN_1
The Cut Value is ,50

Observed		Predicted: keine Warnung 1 k	Predicted: Warnung 1 vorhan W	Percent Correct
keine Warnung 1	k	1646	767	68,21%
Warnung 1 vorhan	W	354	2085	85,49%
			Overall	76,90%

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
SEX	-,7656	,0722	112,5696	1	,0000	-,1282	,4650
ALTER_G			196,7538	5	,0000	,1666	
ALTER_G(1)	,8354	,0900	86,2534	1	,0000	,1119	2,3058
ALTER_G(2)	,6720	,1089	38,0503	1	,0000	,0732	1,9581
ALTER_G(3)	,2109	,1266	2,7736	1	,0958	,0107	1,2348
ALTER_G(4)	-,4866	,1767	7,5795	1	,0059	-,0288	,6147
ALTER_G(5)	-1,3960	,2189	40,6776	1	,0000	-,0758	,2476
EINK	-,1765	,0131	182,4632	1	,0000	-,1638	,8382
VERMOEGE	-,0001	1,168E-05	90,6761	1	,0000	-,1148	,9999
DKV_G			160,1823	8	,0000	,1464	
DKV_G(1)	-,7114	,1262	31,7582	1	,0000	-,0665	,4910
DKV_G(2)	-,0099	,1316	,0057	1	,9398	,0000	,9901
DKV_G(3)	-,1934	,1381	1,9623	1	,1613	,0000	,8242
DKV_G(4)	-,9396	,1044	80,9613	1	,0000	-,1083	,3908
DKV_G(5)	,1678	,1913	,7694	1	,3804	,0000	1,1827
DKV_G(6)	,1935	,1691	1,3101	1	,2524	,0000	1,2135
DKV_G(7)	,2533	,1731	2,1415	1	,1434	,0046	1,2882
DKV_G(8)	,3902	,1964	3,9499	1	,0469	,0170	1,4773
Constant	,9067	,0990	83,8269	1	,0000		

Variables not in the Equation

Residual Chi Square 1,407 with 1 df Sig = ,2355

Variable	Score	df	Sig	R
REGION_1	1,4075	1	,2355	,0000

No more variables can be deleted or added.

Dieses schrittweise Vorgehen führt zu einem sehr umfangreichen SPSS-Output. Allerdings bietet eine solche Methode wichtige Informationen darüber, ob alle Variablen im Modell enthalten sind und wenn dies nicht der Fall ist, welche Variable gar nicht aufgenommen bzw. nach ihrer Aufnahme wieder ausgeschlossen wurde.

Dem SPSS-Output 5.2 ist zunächst zu entnehmen, daß im ersten Schritt die Variable EINK das Aufnahmekriterium erfüllt und in der Modellschätzung berücksichtigt wird. Im zweiten Schritt wird die Variable ALTER_G aufgenommen, danach DKV_G, SEX und VERMOEGE. In keinem Schritt wird eine aufgenommene Variable ausgeschlossen. Einige

Kategorien der Dauer der Kundenverbindung sowie die Altersklasse der 50- bis unter 60-jährigen erfüllen nach ihrer Aufnahme in das Modell das Ausschlußkriterium, werden aber nicht herausgenommen. Eine Kategorie von DKV (DKV_G(3), d.h. 106 bis einschließlich 115 Monate) hätte sogar nach dem Einschlußkriterium nicht in das Modell aufgenommen werden dürfen. Dennoch sind alle Gruppen der Variablen ALTER_G und DKV_G in der endgültigen Regression enthalten, da bei einer drei- und mehrkategorialen Variablen für die Aufnahme in das Modell bzw. für ihren Ausschluß das jeweilige Kriterium für die Variable als Ganzes (ALTER_G bzw. DKV_G) und nicht für die einzelnen Kontrast-Variablen relevant ist.

Die wichtigste Erkenntnis dieses schrittweisen Vorgehens ist, daß REGION_1 als einzige erklärende Variable nicht in die Regression aufgenommen wird. Werden die Veränderungen des Modells durch die Nichtaufnahme dieser Variablen betrachtet, so ist durch den Vergleich der SPSS-Outputs 5.1 und 5.2 (im zweiten Fall nur die Ergebnisse für das im letzten Schritt geschätzte Modell) festzustellen, daß der Wert des Nagelkerke $\widetilde{R}^2$ und die Ergebnisse der Klassifikationstabelle nahezu gleich geblieben sind. Auch die Werte der Regressionskoeffizienten für die meisten Variablen unterscheiden sich in den zwei Modellen kaum voneinander. Dies ist der nachfolgenden Tabelle 5.3 zu entnehmen. Es fällt dennoch auf, daß die Veränderungen bei den Koeffizienten für die Kategorien der Dauer der Kundenverbindung größer ausfallen als bei den Regressionskoeffizienten der anderen Variablen. Aus der Tatsache, daß die Nichteinbeziehung der Variablen REGION_1 in das Modell zu größeren Veränderungen bezüglich der geschätzten Koeffizienten für die Variable DKV_G führt, entsteht die Vermutung, daß zwischen beiden Variablen Wechselwirkungen bestehen.

Abschließend ist festzuhalten, daß die Regressionskoeffizienten der Variablen ALTER_G(3), DKV_G(2), DKV_G(3), DKV_G(5), DKV_G(6) und DKV_G(7) weiterhin nicht signifikant verschieden von Null sind. Wie kann nun dieses Ergebnis verbessert werden? Da es sich hier um Kontrast-Variablen handelt, kann z.B. durch eine andere Gruppierung eine Signifikanzverbesserung erreicht werden. Möglicherweise reicht auch nur die Veränderung der jeweiligen Bezugskategorie aus. Da der zweite Weg schnell zu überprüfen ist, wird diesem nunmehr in dem nachfolgenden Abschnitt 5.3. nachgegangen, bevor dann im Abschnitt 5.4. überprüft wird, ob zwischen den Variablen REGION_1 und DKV_G starke Wechselwirkungen bestehen.

Tabelle 5.3: Vergleich der geschätzten Regressionskoeffizienten der zwei multiplen Modelle der SPSS-Outputs 5.1 und 5.2

	Regressions-koeffizienten im multiplen Modell unter SPSS-Output 5.1: $\mathbf{b_{M(5.1)}}$	Regressions-koeffizienten im multiplen Modell unter SPSS-Output 5.2:* $\mathbf{b_{M(5.2)}}$	Absolute Differenz zwischen $\mathbf{b_{M(5.2)}}$ und $\mathbf{b_{M(5.1)}}$	Differenz in Prozent
SEX	- 0,7693	- 0,7656	0,0037	0,48
ALTER_G(1)	0,8408	0,8354	- 0,0054	- 0,64
ALTER_G(2)	0,6748	0,6720	- 0,0028	- 0,42
ALTER_G(3)	0,2137	0,2109	- 0,0028	- 1,31
ALTER_G(4)	- 0,4803	- 0,4866	- 0,0063	- 1,31
ALTER_G(5)	- 1,3888	- 1,3960	- 0,0072	- 0,52
EINK	- 0,1760	- 0,1765	- 0,0005	- 0,28
VERMOEGE	- 0,0001	- 0,0001	0,0000	0,00
DKV_G(1)	- 0,6713	- 0,7114	- 0,0401	- 5,97
DKV_G(2)	0,0400	- 0,0099	- 0,0499	- 124,75
DKV_G(3)	- 0,0948	- 0,1934	- 0,0986	- 104,01
DKV_G(4)	- 0,8244	- 0,9396	- 0,1152	- 13,97
DKV_G(5)	0,1890	0,1678	- 0,0212	- 11,22
DKV_G(6)	0,1999	0,1935	- 0,0064	- 3,20
DKV_G(7)	0,2587	0,2533	- 0,0054	- 2,09
DKV_G(8)	0,3957	0,3902	- 0,0055	- 1,39

* Hierbei handelt es sich um die im letzten Schritt geschätzten Regressionskoeffizienten.

5.3. Multiples Modell mit veränderten Bezugskategorien der Variablen ALTER_G und DKV_G

Die nächste multiple logistische Regression im SPSS-Output 5.3 und das Modell im SPSS-Output 5.1 sind sehr ähnlich. Es werden hier nur die Referenzkategorien bei den Variablen ALTER_G und DKV_G verändert, um dem oben genannten Ziel (Signifikanz aller Regressionskoeffizienten) näher zu kommen. Innerhalb der Altersklassen wird jetzt die letzte Gruppe der 70-jährigen und älteren, unter den Kategorien für die Dauer der Kundenverbindung dagegen die erste Gruppe, DKV bis einschließlich 61 Monate, als Referenzkategorie bestimmt. Alle anderen Merkmale des ersten multiplen Modells, darunter auch die Wahl von REGION_1 als erklärende Variable sowie die Methode Einschluß, werden beibehalten. Daher sind die Ergebnisse in den beiden SPSS-Outputs 5.1 und 5.3 nahezu gleich.

SPSS-Output 5.3: Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die letzte Kategorie), bekanntes Einkommen, Vermögen bei der Bank und DKV gruppiert (Bezugskategorie = die erste Kategorie); Methode Einschluß

```
Number of selected cases:                5684
Number rejected because of missing data:  832
Number of cases included in the analysis: 4852
```

	Value	Freq	Parameter Coding (1)	(2)	(3)	(4)	(5)
DKV_G							
DKV <= 61 Monate	1	494	,000	,000	,000	,000	,000
DKV = 62 - 105 Monate	2	496	1,000	,000	,000	,000	,000
DKV = 106 - 115 Monate	3	453	,000	1,000	,000	,000	,000
DKV = 116 - 117 Monate	4	1192	,000	,000	1,000	,000	,000
DKV = 118 - 132 Monate	5	197	,000	,000	,000	1,000	,000
DKV = 133 - 156 Monate	6	254	,000	,000	,000	,000	1,000
DKV = 157 - 180 Monate	7	247	,000	,000	,000	,000	,000
DKV = 181 - 205 Monate	8	207	,000	,000	,000	,000	,000
DKV = 206 Monate und länger	9	1312	,000	,000	,000	,000	,000

		(6)	(7)	(8)
DKV_G				
DKV <= 61 Monate	1	,000	,000	,000
DKV = 62 - 105 Monate	2	,000	,000	,000
DKV = 106 - 115 Monate	3	,000	,000	,000
DKV = 116 - 117 Monate	4	,000	,000	,000
DKV = 118 - 132 Monate	5	,000	,000	,000
DKV = 133 - 156 Monate	6	,000	,000	,000
DKV = 157 - 180 Monate	7	1,000	,000	,000
DKV = 181 - 205 Monate	8	,000	1,000	,000
DKV = 206 Monate und länger	9	,000	,000	1,000

	Value	Freq	Parameter Coding (1)	(2)	(3)	(4)	(5)
ALTER_G							
0-29 Lebensjahre	1	1529	1,000	,000	,000	,000	,000
30-39 Lebensjahre	2	1374	,000	1,000	,000	,000	,000
40-49 Lebensjahre	3	770	,000	,000	1,000	,000	,000
50-59 Lebensjahre	4	504	,000	,000	,000	1,000	,000
60-69 Lebensjahre	5	336	,000	,000	,000	,000	1,000
70 Lebensjahre und älter	6	339	,000	,000	,000	,000	,000

```
-2 Log Likelihood   6726,1609
* Constant is included in the model.

Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..       SEX       Geschl_umkodiert
          REGION_1  Region (aus Risiko)
          ALTER_G   Alter gruppiert
          EINK      bekanntes Einkommen
          VERMOEGE  Vermögen bei der Bank
          DKV_G     DKV gruppiert

Estimation terminated at iteration number 7 because
parameter estimates changed by less than ,001

-2 Log Likelihood     4708,837
Cox & Snell - R^2         ,340
Nagelkerke - R^2          ,454

                    Chi-Square   df  Significance
Model                 2017,324   17         ,0000
```

Classification Table for WARN_1
The Cut Value is ,50

Observed		Predicted: keine Warnung 1 k	Predicted: Warnung 1 vorhan W	Percent Correct
keine Warnung 1	k	1643	770	68,09%
Warnung 1 vorhan	W	345	2094	85,85%
			Overall	77,02%

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
SEX	-,7693	,0723	113,3576	1	,0000	-,1287	,4633
REGION_1	,1281	,1080	1,4073	1	,2355	,0000	1,1367
ALTER_G			197,1700	5	,0000	,1668	
ALTER_G(1)	1,3888	,2190	40,2145	1	,0000	,0754	4,0099
ALTER_G(2)	2,2295	,2201	102,6500	1	,0000	,1223	9,2956
ALTER_G(3)	2,0636	,2272	82,4985	1	,0000	,1094	7,8743
ALTER_G(4)	1,6024	,2352	46,4324	1	,0000	,0813	4,9651
ALTER_G(5)	,9085	,2642	11,8270	1	,0006	,0382	2,4806
EINK	-,1760	,0131	181,7234	1	,0000	-,1635	,8386
VERMOEGE	-,0001	1,168E-05	90,4980	1	,0000	-,1147	,9999
DKV_G			95,3427	8	,0000	,1086	
DKV_G(1)	,7112	,1477	23,1986	1	,0000	,0561	2,0365
DKV_G(2)	,5765	,1635	12,4284	1	,0004	,0394	1,7798
DKV_G(3)	-,1531	,1401	1,1950	1	,2743	,0000	,8580
DKV_G(4)	,8602	,2044	17,7199	1	,0000	,0483	2,3637
DKV_G(5)	,8712	,1832	22,6241	1	,0000	,0554	2,3898
DKV_G(6)	,9300	,1877	24,5378	1	,0000	,0579	2,5344
DKV_G(7)	1,0670	,2097	25,8932	1	,0000	,0596	2,9066
DKV_G(8)	,6713	,1307	26,3808	1	,0000	,0602	1,9567
Constant	-1,2794	,2473	26,7567	1	,0000		

Erwartungsgemäß bleiben das Nagelkerke $\tilde{R}^2$, die ΔD, die Klassifikationstabelle sowie alle unter "Variables in the Equation" angegebenen Werte für die Variablen SEX, REGION_1, EINK und VERMOEGE im Vergleich zum ersten multiplen Modell unter 5.1. unverändert. Nur bei den Kontrast-Variablen für das Alter und für die Dauer der Kundenverbindung sind andere Werte der Regressionskoeffizienten, deren Standardfehler sowie der damit berechneten Statistiken zu beobachten.

Durch die Veränderung der Bezugskategorien für das Alter und für die Dauer der Kundenverbindung wurde tatsächlich eine Verbesserung bezüglich der Signifikanz erreicht. Diese wird nur noch von zwei Koeffizienten nicht erfüllt. Zum einen handelt es sich hier erneut um die Variable REGION_1, zum anderen um die Gruppe derjenigen, die seit 116 oder 117 Monaten Kunden sind (hier DKV_G(3)). Die Mehrheit der Personen mit einer solchen Dauer der Kundenverbindung sind wahrscheinlich Kunden, die in der ehemaligen DDR wohnten und nach der Wiedervereinigung in den neuen Bundesländern geblieben sind.

Daher ist es möglich, daß die Region die Dauer der Kundenverbindung zum Teil beinhaltet bzw. erklärt. Damit entsteht erneut die Frage, ob zwischen den Variablen REGION_1 und DKV_G Wechselwirkungen bestehen.

5.4. Überprüfung und Berücksichtigung von Wechselwirkungen zwischen den Variablen REGION_1 und DKV_G

Wechselwirkungen zwischen den Variablen REGION_1 und DKV_G können die Tatsache erklären, daß die im vorigen Abschnitt genannten Regressionskoeffizienten nicht signifikant sind. Die Ursache dafür liegt in der Eigenschaft von Beziehungen zwischen erklärenden Variablen, die Schätzung der dazugehörigen Regressionskoeffizienten schwächer und unzuverlässiger zu machen. Letzteres führt zur Erhöhung des Standardfehlers dieser Regressionskoeffizienten und begünstigt somit das Beibehalten der Nullhypothese, obwohl $\beta_j \neq 0$ ist.[33]

Die Variablen REGION_1 und DKV_G werden mit Hilfe eines Chi-Quadrat-Unabhängigkeitstests nach Pearson auf Wechselwirkungen überprüft. Dieser ist im SPSS-Output 5.4 angegeben. Die Anzahl der hier einbezogenen Fälle beträgt 4.937 und ist damit höher als die bisherige Anzahl von 4.852 gültigen Fällen. Die Differenz besteht aus 85 Personen, für die Region und Dauer der Kundenverbindung bekannt sind, ihr Alter dagegen nicht. Anhang B enthält einige Erläuterungen zu dem Chi-Quadrat-Unabhängigkeitstest nach Pearson. Es sei hier nur erwähnt, daß alle im Anhang B genannten Bedingungen dieses Tests in der hier behandelten Problemstellung erfüllt sind.

SPSS-Output 5.4: Chi-Quadrat-Unabhängigkeitstest nach Pearson zur Überprüfung von Wechselwirkungen zwischen den Variablen REGION_1 und DKV_G

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	2880,538[a]	8	,000
Anzahl der gültigen Fälle	4937		

a. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 80,68.

[33] Vgl. Rönz, B., Förster, E., S. 200 - 203. An dieser Stelle geht es zwar um die multiple lineare Regression, aber einige allgemeine Überlegungen gelten auch für die multiple logistische Regression.

Bei diesem Test ist die Nullhypothese auf dem vorgegebenen Signifikanzniveau abzulehnen, womit sich der Zusammenhang zwischen den Variablen DKV_G und REGION_1 quantitativ nachweisen läßt. Es stellt sich die Frage, wie stark bzw. schwach dieser Zusammenhang ist. Dazu wird das sogenannte Cramer-V angewandt, das ein Maß für die Stärke der Beziehung zweier nominal- oder höher skalierter Variablen ist.[34] Im SPSS-Output 5.5 ist der Wert des Cramer-V mit 0,764 angegeben. Dies deutet auf einen relativ starken Zusammenhang zwischen den Variablen DKV_G und REGION_1.

SPSS-Output 5.5: Cramer-V zur Überprüfung der Stärke der Beziehung zwischen den Variablen REGION_1 und DKV_G

	Wert	Näherungsweise Signifikanz
Cramer-V	,764	,000
Anzahl der gültigen Fälle	4937	

Nach diesen Ergebnissen ist zu überlegen, wie diese Wechselwirkungen vermieden werden können. Eine Möglichkeit hierzu ist die Eliminierung einer der beiden Variablen. Diese Vorgehensweise ist einfach, robust und wird daher hier gewählt. Allerdings ist sie nicht ganz problemlos, da die Entscheidung darüber, welche Variable entfernt werden soll, nicht einfach ist. Wird die falsche Variable herausgenommen, so kann dies zu Fehlspezifikationen und damit zu verzerrten Schätzwerten für die übrigen Regressionskoeffizienten führen.[35] Um dies möglichst zu vermeiden, ist sachlogisch zu entscheiden, welche Variable (REGION_1 oder DKV_G) nicht in das Modell einzubeziehen ist. Folgende Überlegungen führen dazu, daß die Dauer der Kundenverbindung in die weiteren logistischen Regressionen nicht mit einbezogen wird: Zum einen kann die Wohnregion die Dauer der Kundenverbindung zum Teil erklären und nicht umgekehrt, da nach der Wiedervereinigung der vermutlich wesentlich größere Teil der Kunden aus dem ehemaligen Osten vorwiegend in den neuen Bundesländern und diejenigen aus dem Westen überwiegend in den alten Bundesländern geblieben sind. Zum anderen enthält die Variable DKV_G synthetische Werte, die eigentlich nicht der Realität entsprechen bzw. durch besondere Ereignisse entstanden und damit nicht ohne weiteres mit den übrigen Ausprägungen vergleichbar sind. Ein solcher Wert ist "206", unter dem alle Personen zusammengefaßt sind, die zum

[34] Weitere Erläuterungen zu diesem Maß und seiner Berechnung sind im Anhang B enthalten.
[35] Vgl. Rönz, B., Förster, E., S. 210.

Zeitpunkt der Systemeinführung oder früher Kunden geworden sind. Selbst wenn sie viel länger Kunden waren, werden sie mit einer Dauer der Kundenverbindung von 206 Monaten geführt. Außerdem sind hier auch die Werte "116" und "117" zu nennen, die wegen der technischen Zusammenführung aufgrund der Währungsunion besonders hohe absolute Häufigkeiten aufweisen. Diese Unregelmäßigkeiten können nicht durch eine neue Gruppierung der Variablen DKV_G vermieden werden. Aus den hier aufgeführten Gründen erscheint es sinnvoller, von den beiden Variablen nur REGION_1 in die weiteren Untersuchungen einzubeziehen.

Die nachfolgende logistische Regression im SPSS-Output 5.6 unterscheidet sich von dem vorhergehenden Modell im SPSS-Output 5.3 nur darin, daß jetzt die Variable "DKV_G" nicht mehr als erklärende Variable gewählt wird. Durch den Vergleich der beiden Regressionen wird hier untersucht, welche Veränderungen diese Entscheidung verursacht und ob die Veränderungen in die gewünschte Richtung gehen. Somit konzentrieren sich die nachfolgenden Interpretationen bis auf weiteres auf diesen Vergleich.

SPSS-Output 5.6: Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die letzte Kategorie), bekanntes Einkommen und Vermögen bei der Bank; Methode Einschluß

```
Number of selected cases:                5684
Number rejected because of missing data:  832
Number of cases included in the analysis: 4852

                                        Parameter
                          Value   Freq  Coding
                                          (1)    (2)    (3)    (4)    (5)
ALTER_G
 0-29 Lebensjahre             1   1529  1,000   ,000   ,000   ,000   ,000
 30-39 Lebensjahre            2   1374   ,000  1,000   ,000   ,000   ,000
 40-49 Lebensjahre            3    770   ,000   ,000  1,000   ,000   ,000
 50-59 Lebensjahre            4    504   ,000   ,000   ,000  1,000   ,000
 60-69 Lebensjahre            5    336   ,000   ,000   ,000   ,000  1,000
 70 Lebensjahre und älter     6    339   ,000   ,000   ,000   ,000   ,000

-2 Log Likelihood   6726,1609
* Constant is included in the model.

Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..       SEX       Geschl_umkodiert
          REGION_1  Region (aus Risiko)
          ALTER_G   Alter gruppiert
          EINK      bekanntes Einkommen
          VERMOEGE  Vermögen bei der Bank

Estimation terminated at iteration number 7 because
parameter estimates changed by less than ,001
```

-2 Log Likelihood	4805,608
Cox & Snell - R^2	,327
Nagelkerke - R^2	,436

	Chi-Square	df	Significance
Model	1920,553	9	,0000

Classification Table for WARN_1
The Cut Value is ,50

Observed		Predicted: keine Warnung 1	Predicted: Warnung 1 vorhan	Percent Correct
		k	W	
keine Warnung 1	k	1554	859	64,40%
Warnung 1 vorhan	W	247	2192	89,87%
			Overall	77,21%

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
SEX	-,7877	,0712	122,3705	1	,0000	-,1338	,4549
REGION_1	,5927	,0718	68,1268	1	,0000	,0992	1,8089
ALTER_G			217,8726	5	,0000	,1758	
ALTER_G(1)	1,3082	,2140	37,3570	1	,0000	,0725	3,6997
ALTER_G(2)	2,2406	,2175	106,1287	1	,0000	,1244	9,3986
ALTER_G(3)	2,0374	,2251	81,9211	1	,0000	,1090	7,6710
ALTER_G(4)	1,5505	,2336	44,0449	1	,0000	,0791	4,7138
ALTER_G(5)	,9150	,2628	12,1240	1	,0005	,0388	2,4967
EINK	-,1816	,0132	188,3447	1	,0000	-,1664	,8340
VERMOEGE	-,0001	1,172E-05	92,3129	1	,0000	-,1159	,9999
Constant	-1,0591	,2166	23,9058	1	,0000		

Es ist festzustellen, daß der Wert des Nagelkerke $\tilde{R}^2$ von 0,454 auf 0,436 gesunken ist. Somit erklären die X-Variablen gemeinsam 43,6 % der Variation von Y. Dieser Rückgang ist relativ gering, vor allem wenn man bedenkt, daß diese multiple Regression nur 9 erklärende Variablen, einschließlich der Dummy-Variablen, und damit 8 weniger als das vorhergehende Modell enthält. Trotz der geringfügigen Verschlechterung des Nagelkerke $\tilde{R}^2$ wurden hier gemäß der Klassifikationstabelle 9 Fälle mehr richtig zugeordnet. Dies ist im Vergleich zu den insgesamt 4.852 gültigen Fällen nicht gerade viel, aber dennoch eine positive Entwicklung.

Bedeutender sind hier allerdings die Fragen, ob jetzt alle Regressionskoeffizienten signifikant sind und wie sie sich verändert haben. Die erste Frage ist positiv zu beantworten, womit das Ziel $\beta_j \neq 0$ auch bei dem Regressionskoeffizienten der Variablen REGION_1 erreicht wurde. In bezug auf die zweite Frage zeigt Tabelle 5.4, daß bei den Regressionskoeffizienten der meisten Variablen relativ geringe Veränderungen von bis zu 5,8 % zu beobachten sind. Im Vergleich dazu ist bei dem Regressionskoeffizienten der Variablen REGION_1 ein enormer Anstieg des Wertes um über 360 % (in der Tabelle grau unterlegt)

zu verzeichnen. Die Bedeutung bzw. der partielle Beitrag dieser Variablen zum Modell ist im Verhältnis zu einigen anderen Variablen immer noch relativ gering, aber im Vergleich zur Regression im SPSS-Output 5.3 gestiegen. Dies ist der R-Statistik in der Tabelle 5.5 zu entnehmen.

Tabelle 5.4: Vergleich der geschätzten Regressionskoeffizienten der zwei multiplen Modelle der SPSS-Outputs 5.3 und 5.6

	Regressions-koeffizienten im multiplen Modell unter SPSS-Output 5.3 $b_{M(5.3)}$	Regressions-koeffizienten im multiplen Modell unter SPSS-Output 5.6 $b_{M(5.6)}$	Absolute Differenz zwischen $b_{M(5.6)}$ und $b_{M(5.3)}$	Differenz in Prozent
SEX	- 0,7693	- 0,7877	- 0,0184	- 2,39
REGION_1	0,1281	0,5927	0,4646	362,69
ALTER_G(1)	1,3888	1,3082	- 0,0806	- 5,80
ALTER_G(2)	2,2295	2,2406	0,0111	0,50
ALTER_G(3)	2,0636	2,0374	- 0,0262	- 1,27
ALTER_G(4)	1,6024	1,5505	- 0,0519	- 3,24
ALTER_G(5)	0,9085	0,9150	0,0065	0,72
EINK	- 0,1760	- 0,1816	- 0,0056	- 3,18
VERMOEGE	- 0,0001	- 0,0001	0,0000	0,00

Tabelle 5.5: Variablen und dazugehörige Werte der R-Statistik aus dem SPSS-Output 5.6, absteigend sortiert nach der Höhe des Absolutwertes der R-Statistik, und die Werte der R-Statistik für die gleichen Variablen aus dem SPSS-Output 5.3

	R-Statistik SPSS-Output 5.6	R-Statistik SPSS-Output 5.3
EINK	- 0,1664	- 0,1635
SEX	- 0,1338	- 0,1287
ALTER_G(2)	0,1244	0,1223
VERMOEGE	- 0,1159	- 0,1147
ALTER_G(3)	0,1090	0,1094
REGION_1	0,0992	0,0000
ALTER_G(4)	0,0791	0,0813
ALTER_G(1)	0,0725	0,0754
ALTER_G(5)	0,0388	0,0382

5.5. Identifikation von Ausreißern und das daraus folgende endgültige Modell dieser Untersuchung

5.5.1. Identifikation von Ausreißern

Nachdem das Ziel erreicht wurde, daß möglichst alle Regressionskoeffizienten zum vorgegebenen Signifikanzniveau verschieden von Null sind, wird nun versucht, die Anpassung des Modells an die Daten zu verbessern. Dazu werden die in den multiplen Modellen 4.852 gültigen Fälle nach potentiellen Ausreißern untersucht. Als Ausreißer werden im allgemeinen "... extreme Beobachtungswerte in einer statistischen Reihe [bezeichnet], die qualitativ von der Gesamtheit abweichende statistische Elemente signalisieren ... ".[36] Solche atypischen bzw. extremen Beobachtungen können die statistische Schätzung verzerren. Daher ist deren Identifikation wichtig.

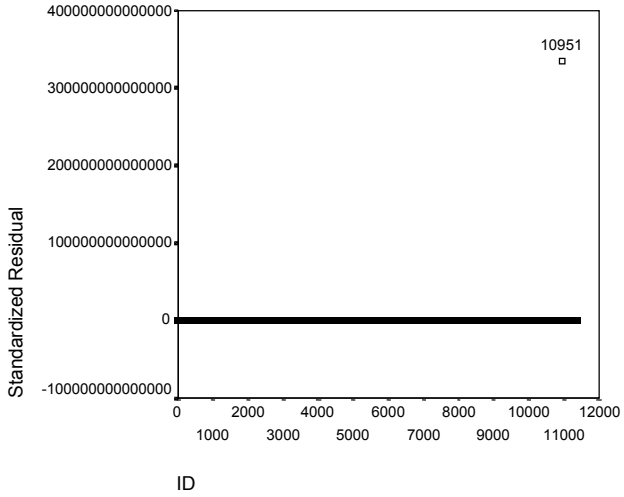
Die SPSS-Software 9.0 bietet die Möglichkeit, verschiedene Kennzahlen zu berechnen, die bei der Entdeckung von Ausreißern helfen können. Aus diesen Kennzahlen werden hier nur diejenigen gewählt, die in der konkreten Untersuchung zur Identifikation besonders beigetragen haben bzw. die atypischen Beobachtungswerte deutlich zeigen. Es werden daher für die als letzte präsentierte multiple logistische Regression (SPSS-Output 5.6) die standardisierten Residuen, die Devianceabweichung, die Cook's Distance und ein ausgewähltes DFBETA gezeigt.

Zunächst werden die standardisierten Residuen $z_i = \hat{u}_i / s_i$ (i = 1,...,4.852) berechnet.[37] Die Standardisierung erfolgt mittels der Division der Residuen durch ihre Standardabweichung s_i und ermöglicht eine bessere Erkennung von größeren Abweichungen als die Residuen selbst. Der nachfolgende Scatterplot der Identifikationsnummern (ID) gegen die standardisierten Residuen im SPSS-Output 5.7 zeigt, daß sich für die Beobachtung mit der ID = 10.951 ein viel größerer Wert ergibt als für die anderen Fälle.

[36] Rönz, B. (1997a), S. 56.

[37] Für alle Beobachtungen der k-ten Gruppe (k = 1,...,q) wird die gleiche Wahrscheinlichkeit $\hat{\pi}_k$ und folglich das gleiche Residuum berechnet. Somit sind z_i und z_k identisch. Es macht daher keinen Unterschied, ob der Laufindex der standardisierten Residuen "i" oder "k" heißt. Letzteres gilt auch für die anderen Kennzahlen in diesem Abschnitt, soweit sie je Beobachtung berechnet werden.

SPSS-Output 5.7: Scatterplot Identifikationsnummern (ID) gegen standardisierte Residuen

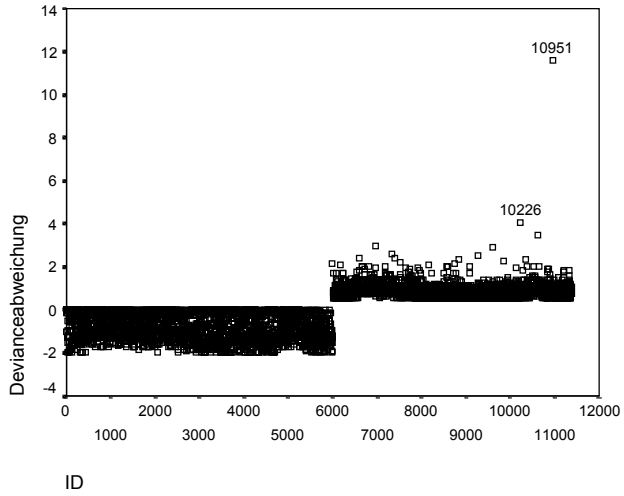


Als nächstes wird die Devianceabweichung für die einzelnen Fälle berechnet als:[38]

$(\text{Devianceabweichung})_k = -\sqrt{-2\log(\hat{\pi}_k)}$.

Diese Kennzahl beinhaltet den Beitrag der i-ten Beobachtung zur Deviance. Ein großer Wert bedeutet, daß für diesen Fall das Modell nicht gut angepaßt ist. Der Scatterplot im SPSS-Output 5.8 zeigt, daß die Werte der Devianceabweichung für die meisten Beobachtungen zwischen -2 und +2 liegen. Für die Person mit der ID = 10.951 wurde dagegen ein deutlich höherer Wert von 11,57 berechnet. Der zweithöchste Wert beträgt 4,02 und gehört zu dem Kunden mit der ID = 10.226.

SPSS-Output 5.8: Scatterplot Identifikationsnummern (ID) gegen Devianceabweichung



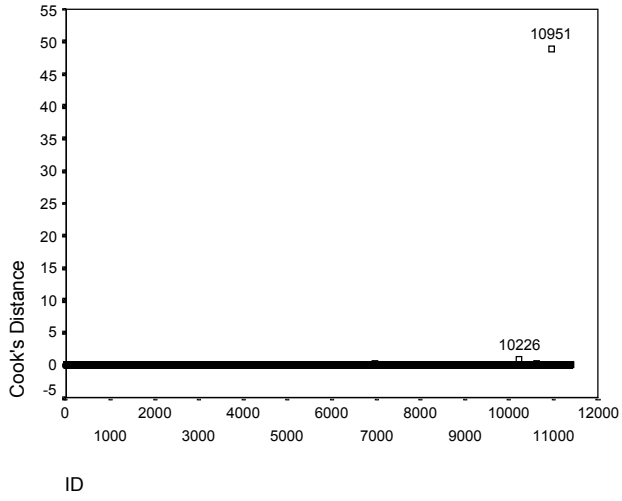
[38] Vgl. SPSS Regression Models™ 9.0, S. 55.

Als dritte Kennzahl wird die Cook's Distance betrachtet. Sie ist ein Maß für den Einfluß der i-ten Beobachtung und gibt die Veränderung in allen Residuen an, wenn die i-te Beobachtung ausgeschlossen wird. Die Berechnung der Cook's Distance basiert auf den standardisierten Residuen z_i sowie auf dem leverage h_i, das die "Hebelwirkung" (Einfluß) der i-ten Beobachtung der erklärenden Variablen auf die Koeffizientenschätzung angibt:[39, 40]

$$d_i = \frac{z_i^2 * h_i}{(1 - h_i)^2} \quad .$$

Nach der Berechnung der Cook's Distance für alle Beobachtungen, wird diese im SPSS-Output 5.9 gegen die jeweilige Identifikationsnummer geplottet.

SPSS-Output 5.9: Scatterplot Identifikationsnummern (ID) gegen Cook's Distance



Die Werte der Cook's Distance sind für die (N-1) Fälle zwischen Null und Eins. Nur für die Beobachtung mit der Identifikationsnummer 10.951 wurde dagegen ein viel höherer Wert von ca. 49 berechnet. Dies deutet darauf hin, daß der Ausschluß dieses Falles eine größere Auswirkung auf die Residuen hat. Die nachfolgende Tabelle zeigt die fünf größten Werte der Cook's Distance und die Identifikationsnummern der dazugehörigen Fälle.

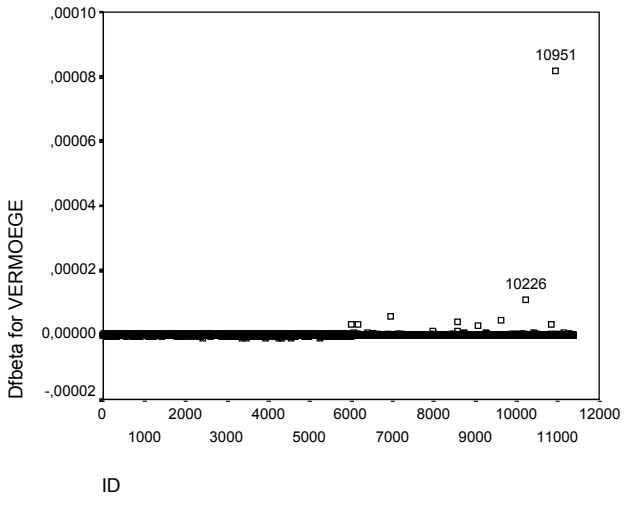
[39] Vgl. SPSS Regression Models™ 9.0, S. 56.
[40] Zur genauen Definition von h_i als i-tes Diagonalelement der sogenannten Projektionsmatrix siehe Hosmer, D. W., Lemeshow, S., S. 151.

SPSS-Output 5.10: Die fünf größten Werte der Cook's Distance und die Identifikationsnummern der dazugehörigen Fälle

	ID	Wert
1	10951	48,96654
2	10226	,87712
3	10633	,26414
4	6969	,26150
5	9610	,16414

Bei allen bis jetzt zur Ausreißeridentifikation verwendeten Kennzahlen ist die Person mit der Identifikationsnummer 10.951 besonders auffällig. Bei zwei der drei Scatterplots fiel auch diejenige mit ID = 10.226 auf, obwohl sich ihre jeweiligen Werte nicht so extrem von der Mehrheit unterscheiden, wie dies bei der ersten Person der Fall ist. Aus diesen Ergebnissen entsteht die Frage nach dem Grund für die Auffälligkeit dieser Fälle. Aufschluß darüber kann der Scatterplot (SPSS-Output 5.11) der Identifikationsnummern gegen DFBETA für die Variable Vermögen bei der Bank geben. Die DFBETAS sind ein weiteres wichtiges Maß zur Identifikation von potentiellen Ausreißern und beinhalten die Differenz zwischen dem Vektor der logistischen Koeffizienten **b**, der unter Einbeziehung aller Beobachtungen geschätzt wird, und dem Vektor **b**(i), der bei Ausschluß der i-ten Beobachtung berechnet wird.[41]

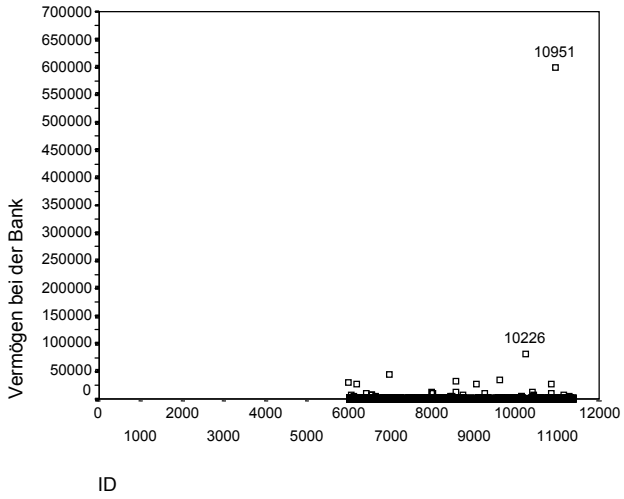
SPSS-Output 5.11: Scatterplot Identifikationsnummern (ID) gegen DFBETA für Vermögen



[41] Vgl. SPSS Regression Models™ 9.0, S. 56.

Die größte Veränderung mit 0,00008 ergibt sich in der Schätzung von β bei Ausschluß der Beobachtung 10.951. Deutlich geringer mit nur 0,00001, aber dennoch an zweiter Stelle, fällt diese Veränderung dagegen bei Ausschluß der Beobachtung 10.226 aus. Es handelt sich also vermutlich um Fälle, die in bezug auf die Variable Vermögen auffällige Ausprägungen aufweisen. Es stellt sich heraus, daß unter allen Kunden, die Warnung 1 erhalten haben, die Person mit ID = 10.951 das höchste Vermögen in Höhe von 598.541,41 DM aufweist und die Person mit ID = 10.226 das zweithöchste in Höhe von 80.520,69 DM. Wie der Scatterplot im SPSS-Output 5.12 zeigt, ist es unter den Kunden mit Warnung 1 eher üblich, kein bzw. kein nennenswertes Vermögen bei der Bank zu haben. Von allen bis jetzt zur Schätzung der multiplen Modelle berücksichtigten Fällen mit Warnung 1 haben ca. 81 % ein Vermögen von unter 100,00 DM bei der Bank. Ein möglicher Grund, weshalb die zwei oben genannten Kunden so hohe Werte ihres Vermögens bei der Bank aufweisen, kann darin liegen, daß sie, bevor sie weitere Darlehen bzw. ein sehr hohes Darlehen erhalten haben, eine gewisse Sicherheit hinterlegen mußten.

SPSS-Output 5.12: Scatterplot Identifikationsnummern (ID) gegen Vermögen bei der Bank, gegeben Warnung 1



Obwohl das Einkommen den größten partiellen Beitrag zur multiplen logistischen Regression im SPSS-Output 5.6 leistet (siehe Tabelle 5.5), sind die Personen mit Warnung 1 und hohem Einkommen nicht so auffällige, atypische Fälle wie die mit hohem Vermögen. In die Beobachtungsreihe passen offensichtlich die Fälle mit hohem Einkommen besser als diejenigen mit hohem Vermögen. Um dies zu veranschaulichen, werden die $\hat{\pi}_k$ für alle Personen mit Warnung 1 geschätzt. Werte von $\hat{\pi}_k \geq 0{,}5$ bedeuten, daß diese Fälle vom Modell richtig zugeordnet werden, also sind sie hier nicht von Interesse. Unter den Kunden

mit Warnung 1 werden diejenigen falsch klassifiziert, für die $\hat{\pi}_k < 0{,}5$ geschätzt wurde. Der SPSS-Output 5.13 beinhaltet diejenigen 5 Fälle von allen zur Schätzung verwendeten 2.439 Kunden mit Warnung 1, für die die niedrigsten Erfolgswahrscheinlichkeiten $\hat{\pi}_k$ geschätzt wurden. Erneut sind die ersten zwei Fälle davon die mit den Identifikationsnummern 10.951 und 10.226, d.h. für sie wurden die niedrigsten Wahrscheinlichkeiten geschätzt, obwohl sie tatsächlich Warnung 1 erhalten haben. Erst an dritter Stelle und mit einem deutlich höheren Wert von $\hat{\pi}_k$ steht der Kunde mit ID = 10.633, der das höchste Einkommen hat, gegeben Warnung 1. Danach kommen diejenigen Kunden (ID = 6.969 bzw. 9.610), die, gegeben Warnung 1, das dritt- und vierthöchste Vermögen bei der Bank aufweisen.

SPSS-Output 5.13: Die kleinsten Werte der geschätzten Wahrscheinlichkeiten, gegeben Warnung 1

	ID	Wert
1	10951	,00000
2	10226	,00031
3	10633	,00223
4	6969	,01112
5	9610	,01593

Als Ergebnis dieser Untersuchungen bezüglich Ausreißern wird entschieden, die Analyse ohne die Fälle mit den Identifikationsnummern 10.951 und 10.226 fortzusetzen. Es erscheint sachgerecht, diese Personen aus der Modellierung herauszulassen, weil bei einem so hohen Vermögen von 598.541,41 DM bzw. 80.520,69 DM das mit den aufgetretenen Zahlungsschwierigkeiten verbundene Risiko aus Sicht der Bank vermutlich nicht sehr hoch ist. Diese Situation ist ungewöhnlich für Kunden mit Zahlungsschwierigkeiten.

Bei der zweiten Person (ID = 10.226) weisen die standardisierten Residuen, Devianceabweichung, Cook's Distance, DFBETA für VERMOEGE und geschätzte Wahrscheinlichkeit nicht so auffällige Werte wie bei der ersten Person auf. Im Vergleich zu den übrigen Personen sind aber diese Werte und vor allem der Betrag des Vermögens außergewöhnlich hoch bzw. atypisch. Deshalb erscheint es auch hier sachgerecht, diesen Fall ebenfalls zu entfernen.

Dagegen hat der Kunde mit ID = 10.633 zwar das höchste Einkommen unter allen Personen mit Warnung 1, aber seine durchschnittlichen monatlichen Einkünfte zwischen 3.851 und 3.950 DM können nach den bereits aufgetretenen Zahlungsschwierigkeiten nicht die Sicherheit der oben genannten Vermögensbeträge bieten. Daher wird dieser Fall nicht als verhältnismäßig atypisch betrachtet und folglich nicht ausgeschlossen.

5.5.2. Die letzte multiple logistische Regression als Ergebnis dieser Untersuchung

Nach der Entfernung der Fälle mit den Identifikationsnummern 10.951 und 10.226 wird erneut das Modell mit den ansonsten gleichen Merkmalen wie die Regression im SPSS-Output 5.6, aber nunmehr mit 4.850 Privatkunden, geschätzt. Als X-Variablen werden Geschlecht, Region, Alter gruppiert, bekanntes Einkommen und Vermögen bei der Bank in einem Schritt (Methode Einschluß) in das Modell aufgenommen.[42] Von den nichtbinären Variablen wird nur ALTER_G als kategorial bestimmt, wobei die Kategorie der 70-jährigen und älteren als Referenzkategorie gewählt wird. Der SPSS-Output 5.14 enthält die wichtigsten Ergebnisse dieser multiplen logistischen Regression, die das Resultat aller in dieser Arbeit durchgeführten Untersuchungen ist und als das adäquate Modell der behandelten Problemstellung empfohlen wird.

SPSS-Output 5.14: Logistische Regression mit Warnung 1 sowie Geschlecht, Region, Alter gruppiert (Bezugskategorie = die letzte Kategorie), bekanntes Einkommen und Vermögen bei der Bank; Methode Einschluß

Ohne Fälle: ID = 10.226 und ID = 10.951

```
Number of selected cases:                  5682
Number rejected because of missing data:    832
Number of cases included in the analysis: 4850
```

	Value	Freq	Parameter Coding (1)	(2)	(3)	(4)	(5)
ALTER_G							
0-29 Lebensjahre	1	1529	1,000	,000	,000	,000	,000
30-39 Lebensjahre	2	1374	,000	1,000	,000	,000	,000
40-49 Lebensjahre	3	769	,000	,000	1,000	,000	,000
50-59 Lebensjahre	4	503	,000	,000	,000	1,000	,000
60-69 Lebensjahre	5	336	,000	,000	,000	,000	1,000
70 Lebensjahre und älter	6	339	,000	,000	,000	,000	,000

```
-2 Log Likelihood   6723,4089
* Constant is included in the model.
```

[42] Ein schrittweises Vorgehen wurde ebenfalls überprüft. Auf die Erläuterung dieses Modells wird hier aber verzichtet, da sich erwartungsgemäß keine Unterschiede zu der Methode Einschluß ergaben.

```
Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..       SEX       Geschl_umkodiert
          REGION_1  Region (aus Risiko)
          ALTER_G   Alter gruppiert
          EINK      bekanntes Einkommen
          VERMOEGE  Vermögen bei der Bank

Estimation terminated at iteration number 8 because
Log Likelihood decreased by less than ,01 percent.

-2 Log Likelihood     4547,370
Cox & Snell - R^2         ,362
Nagelkerke - R^2          ,482

                    Chi-Square   df  Significance
Model                2176,039     9         ,0000

Classification Table for WARN_1
The Cut Value is ,50               Predicted
                       keine Warnung 1  Warnung 1 vorhan  Percent Correct
                                  k           W
Observed                        ┌──────────┬──────────┐
keine Warnung 1    k            │   1605   │   808    │    66,51%
                                ├──────────┼──────────┤
Warnung 1 vorhan   W            │   256    │   2181   │    89,50%
                                └──────────┴──────────┘
                                                Overall  78,06%

---------------------- Variables in the Equation -----------------------
Variable            B      S.E.       Wald    df      Sig       R    Exp(B)
SEX           -,7566     ,0735   106,0710     1    ,0000  -,1244     ,4693
REGION_1       ,5843     ,0740    62,2721     1    ,0000   ,0947    1,7938
ALTER_G                           210,7268     5    ,0000   ,1728
 ALTER_G(1)   1,2425     ,2187    32,2765     1    ,0000   ,0671    3,4644
 ALTER_G(2)   2,2203     ,2227    99,3797     1    ,0000   ,1203    9,2101
 ALTER_G(3)   2,0169     ,2309    76,3317     1    ,0000   ,1051    7,5152
 ALTER_G(4)   1,5530     ,2400    41,8770     1    ,0000   ,0770    4,7255
 ALTER_G(5)    ,9219     ,2717    11,5139     1    ,0007   ,0376    2,5140
EINK          -,1767     ,0134   174,8158     1    ,0000  -,1603     ,8381
VERMOEGE      -,0003 2,657E-05   154,7994     1    ,0000  -,1508     ,9997
Constant      -,8820     ,2213    15,8845     1    ,0001
```

Durch den Vergleich der Ergebnisse dieses Modells mit dem vorletzten Modell im SPSS-Output 5.6 fällt als einer der wichtigsten Unterschiede auf, daß der Wert des Nagelkerke $\widetilde{R}^2$ von 0,436 auf 0,482 spürbar gestiegen ist. Die X-Variablen erklären also gemeinsam 48,2 % der Variation der abhängigen Variablen. Somit ist dieses Modell am besten von allen im Rahmen dieser Arbeit untersuchten multiplen logistischen Regressionen an die Daten angepaßt. Allerdings bleiben über 50 % der Variation von Y unerklärt. Dies läßt vermuten, daß eine Reihe von weiteren wichtigen Einflußfaktoren für das Auftreten von Warnung 1 bzw. von Zahlungsschwierigkeiten existieren. Darunter können Merkmale wie:

- Familienstand,
- bisherige Zahlungsmoral,
- Beruf,
- Beschäftigungsdauer beim jetzigen Arbeitgeber,
- sonstige Zahlungsverpflichtungen

u.ä. eine entscheidende Rolle spielen. Leider bietet die vorliegende Datei keine Informationen darüber und ermöglicht somit nicht, weitere wichtige Faktoren für das Auftreten des Ereignisses zu finden und das Ausmaß ihrer Wirkung zu quantifizieren.

Die Ergebnisse der Kontingenztabelle haben sich im Vergleich zum Modell im SPSS-Output 5.6 auch, allerdings nur geringfügig, verbessert. Insgesamt werden 78,06 % der Fälle vom Modell richtig klassifiziert. Falsch zugeordnet werden dabei 33,49 % oder 808 Personen in der Gruppe der Kunden ohne Warnung 1, in der Gruppe derjenigen mit Warnung 1 dagegen nur 10,5 % oder 256 Personen. Es ist folglich für das Modell verhältnismäßig schwieriger, die Kunden richtig zu identifizieren, die keine Warnung 1 bzw. keine erheblichen Zahlungsschwierigkeiten aufweisen. Oder anders ausgedrückt, werden Personen ohne Warnung 1 eher dem gefährdeten Kreis von Kunden zugeordnet als Personen mit Warnung 1 als ungefährdet eingeschätzt. Damit ist das Modell eher vorsichtig und konservativ.

Das Ergebnis des Tests mittels der Reduktion in der Deviance ΔD ist unverändert geblieben. So üben, wie bei allen bisherigen multiplen Modellen, die X-Variablen gemeinsam einen signifikanten Einfluß auf das Auftreten von Warnung 1 aus. Nach dem Wald-Test sind erneut alle Regressionskoeffizienten signifikant verschieden von Null. Folglich trägt auch jede einzelne X-Variable wesentlich zur Erklärung der Variation von Y bei. Damit nähert sich dieses Modell im Vergleich zu allen im Rahmen dieser Arbeit untersuchten multiplen logistischen Regressionen den beiden Zielen, signifikante Koeffizienten und eine möglichst gute Anpassung des Modells an die Daten, am meisten an.

Das Ausschließen der beiden Fälle mit den Identifikationsnummern 10.951 und 10.226 hat zu Veränderungen in den Koeffizienten geführt. Diese sind in Tabelle 5.6 angegeben. Die Veränderungen sind prozentual gesehen, außer bei der Variablen VERMOEGE, relativ

gering. Der Absolutwert ihres Koeffizienten steigt dagegen um 200 % (in der Tabelle 5.6 grau unterlegt). Damit steigt auch der Absolutwert der R-Statistik von 0,1159 auf 0,1508, was auf einen höheren partiellen Beitrag dieser erklärenden Variablen zum Modell hindeutet.

Tabelle 5.6: Vergleich der geschätzten Regressionskoeffizienten der zwei multiplen Modelle der SPSS-Outputs 5.6 und 5.14

	Regressions-koeffizienten im multiplen Modell unter SPSS-Output 5.6 $\mathbf{b_{M(5.6)}}$	Regressions-koeffizienten im multiplen Modell unter SPSS-Output 5.14 $\mathbf{b_{M(5.14)}}$	Absolute Differenz zwischen $\mathbf{b_{M(5.14)}}$ und $\mathbf{b_{M(5.6)}}$	Differenz in Prozent
SEX	- 0,7877	- 0,7566	0,0311	3,95
REGION_1	0,5927	0,5843	- 0,0084	- 1,42
ALTER_G(1)	1,3082	1,2425	- 0,0657	- 5,02
ALTER_G(2)	2,2406	2,2203	- 0,0203	- 0,91
ALTER_G(3)	2,0374	2,0169	- 0,0205	- 1,01
ALTER_G(4)	1,5505	1,5530	0,0025	0,16
ALTER_G(5)	0,9150	0,9219	0,0069	0,75
EINK	- 0,1816	- 0,1767	0,0049	2,70
VERMOEGE	- 0,0001	- 0,0003	- 0,0002	-200,00

Aus den geschätzten und hier angegebenen Regressionskoeffizienten ergibt sich für jeden Kunden eine Erfolgswahrscheinlichkeit $\hat{\pi}_k$, d.h. die Wahrscheinlichkeit, daß diese Person Warnung 1 erhält, wie folgt:

$$\hat{\pi}_k = \frac{1}{1+e^{-[\,-0{,}8820-0{,}7566*(\text{sex})\ +0{,}5843*(\text{region\_1})+\ \ldots\ -0{,}0003*(\text{vermoege})\,]}}\ .$$

Wird die gleiche Person wie im Abschnitt 5.1. mit den Merkmalen: männlich, wohnhaft in den neuen Bundesländern, 45 Jahre alt (hier ALTER_G(3)), kein bekanntes Einkommen und kein Vermögen bei der Bank, betrachtet, so wird für diese eine Wahrscheinlichkeit in Höhe von 0,76 berechnet, womit das Risiko für das Auftreten von Zahlungsschwierigkeiten bei diesem Kunden höher geschätzt wird als im ersten Modell. Dort waren es 0,65.

Da es sich hier um das endgültige Modell der Untersuchung handelt, ist es wichtig, erneut die ausführliche Interpretation der Regressionskoeffizienten und der R-Statistik vorzunehmen, weil dies Auskunft über die Art und das Ausmaß des Einflusses der erklärenden Variablen auf die abhängige Variable gibt.

Der Regressionskoeffizient der binären Variablen SEX in Höhe von -0,7566 gibt an, daß die log odds bei den Männern (sex = 0) um 0,7566 höher sind als bei den Frauen (sex = 1). Das Auftreten des interessierenden Ereignisses wurde in der Stichprobe bei den Männern ca. (1 / Exp(B)) = 2,1-mal häufiger beobachtet als bei den Frauen. Daher kann bei einem männlichen Kunden im Vergleich zu einem weiblichen, bei sonst gleichen Ausprägungen der X-Variablen, von einem höheren Risiko für das Auftreten von Zahlungsschwierigkeiten ausgegangen werden.

Für die zweite binäre Variable REGION_1 in diesem Modell wurde ein Koeffizient in Höhe von 0,5843 geschätzt, also sind die log odds bei den in den alten Bundesländern wohnenden Kunden (region_1 = 1) um 0,5843 höher als bei denen, die in den neuen Bundesländern wohnen (region_1 = 0). Der Wert von Exp(B) für diese Variable zeigt, daß das Auftreten von Warnung 1 bei den im Westen wohnenden Kunden ca. 1,8-mal wahrscheinlicher ist als bei den im Osten wohnenden.

Der negative Wert des Regressionskoeffizienten für EINK gibt an, daß bei einer Erhöhung dieser X-Variablen um eine Einheit, wobei diese zwischen 1 und 100 DM liegen kann, die log odds um 0,1767 sinken. Daher sinkt auch das Risiko für das Auftreten von Warnung 1 bzw. von erheblichen Zahlungsschwierigkeiten mit steigenden Werten des bekannten Einkommens, vorausgesetzt alle anderen erklärenden Variablen bleiben konstant.

Der Koeffizient für VERMOEGE ist ebenfalls negativ, also gehen die log odds bei einer Erhöhung dieser X-Variablen um eine Einheit, hier 1 Pfennig, um 0,0003 zurück. Mit steigenden Werten des Vermögens bei der Bank sinkt c.p. die Wahrscheinlichkeit für das Auftreten von Warnung 1. Es sei hier noch einmal darauf hingewiesen, daß der geringe Absolutwert des Koeffizienten für VERMOEGE vorwiegend durch die pfenniggenaue Angabe des Vermögens und durch seine hohen Ausprägungen zu erklären ist.

Die Werte aller für die Kategorien der Variablen ALTER_G geschätzten Regressionskoeffizienten sind positiv. Somit sind die log odds bei jeder dieser Kategorien im Vergleich zur Bezugskategorie, der 70-jährigen und älteren, höher. Aus den unter Exp(B) angegebenen Werten ist zu entnehmen, daß das Auftreten von Warnung 1 im Vergleich zur Referenzkategorie bei den Kunden unter 30 Jahren rund 3,5-mal wahrscheinlicher, bei den 30- bis unter 40-jährigen sogar 9,2-mal, bei den 40- bis unter 50-jährigen 7,5-mal, bei den

50- bis unter 60-jährigen 4,7-mal und bei den 60- bis unter 70-jährigen 2,5-mal wahrscheinlicher ist. Die Tatsache, daß gerade bei den 30- bis unter 40-jährigen Kunden im Vergleich zur Bezugskategorie das Auftreten des interessierenden Ereignisses am wahrscheinlichsten ist, kann damit erklärt werden, daß diese Personen im wirtschaftlichen Sinne vermutlich am aktivsten sind. In diesem Alter werden, nachdem die ersten Jahre des Berufslebens vergangen sind, häufig größere Anschaffungen getätigt, die vorfinanziert werden müssen. Hierbei reichen die Ersparnisse nicht unbedingt aus. Je mehr fremdfinanziert wird, um so größer ist das Risiko, daß Zahlungsschwierigkeiten auftreten. Eventuell spielt hier auch das Auftreten einer unerwarteten Arbeitslosigkeit eine Rolle. Dagegen wurde in der letzten Gruppe der 70-jährigen und älteren Kunden das Auftreten von Warnung 1 im Vergleich zu allen anderen Gruppen am wenigsten beobachtet. Vermutlich werden in diesem Alter kaum größere Anschaffungen getätigt. Wenn dies doch der Fall ist, so greifen diese Kunden wahrscheinlich zunächst auf ihre Ersparnisse zurück. Sollten sie sich dennoch bei der Bank verschulden, so bietet die Regelmäßigkeit ihrer Rente eine gewisse Sicherheit. Außerdem dürfte eine Kontoüberziehung bei diesen Kunden eher unpopulär sein.

Abschließend zu diesem multiplen Modell bleibt die Frage über die Wichtigkeit der einzelnen X-Variablen zur Erklärung des Auftretens von Warnung 1 bzw. von erheblichen Zahlungsschwierigkeiten. Auskunft hierzu geben die Absolutwerte der R-Statistik, die außer im SPSS-Output 5.14 auch in der Tabelle 5.7 enthalten sind. Den größten partiellen Beitrag zu der hier empfohlenen multiplen logistischen Regression leistet demnach die Variable EINK, gefolgt von VERMOEGE, SEX usw.

Tabelle 5.7: Variablen und dazugehörige Werte der R-Statistik aus dem SPSS-Output 5.14, absteigend sortiert nach der Höhe des Absolutwertes der R-Statistik

	R-Statistik SPSS-Output 5.14
EINK	- 0,1603
VERMOEGE	- 0,1508
SEX	- 0,1244
ALTER_G(2)	0,1203
ALTER_G(3)	0,1051
REGION_1	0,0947
ALTER_G(4)	0,0770
ALTER_G(1)	0,0671
ALTER_G(5)	0,0376

6. Modelldiagnose

6.1. Prüfung der Linearität der erklärenden Variablen

Die logistische Regression unterstellt Linearität der X-Variablen in der link Funktion:

$g(\pi_k) = \Sigma_j \, x_{kj}\beta_j$.

Es ist folglich zu prüfen, ob die X-Variablen, die in der als adäquat angenommenen multiplen logistischen Regression (SPSS-Output 5.14) enthalten sind, diesen zunächst unterstellten linearen Zusammenhang tatsächlich erfüllen. Gewählt wird hierfür eine grafische Methode, die auf die quasi-stetigen Variablen Alter, bekanntes Einkommen und Vermögen bei der Bank angewandt wird. Dazu wird jede dieser Variablen gruppiert. Für jede der m Gruppen der jeweiligen Variablen wird im Anschluß folgendes berechnet:

$$\log\left(\frac{p_g}{1-p_g}\right), \qquad g = 1,...,m \,, \qquad \text{Gleichung 6.1}$$

worin p_g die beobachtete relative Häufigkeit für das Auftreten von Warnung 1 in der jeweiligen Gruppe und g = 1,...,m der Laufindex für die Gruppen der entsprechenden X-Variablen ist. Anschließend werden für jede Variable die Klassennummern gegen diese Werte geplottet. Wenn die Linearitätsbedingung erfüllt ist, sollten diese Punkte in etwa auf einer Geraden liegen. Anderenfalls ist von einer nichtlinearen Beziehung auszugehen.

Die Variablen SEX und REGION_1 werden hier nicht betrachtet, da sie nominalskaliert sind und daher die Linearitätsüberprüfung keinen Sinn ergibt.

Die Variable ALTER_G wird als erste auf Linearität untersucht. Das Alter des Kunden ist in dem multiplen Modell bereits in Form einer gruppierten Variablen (ALTER_G) eingeschlossen. Daher wird die bisher verwendete Gruppierung in sechs Altersklassen beibehalten. Der nachfolgende SPSS-Output 6.1 enthält für jede dieser Altersklassen die beobachtete relative Häufigkeit für das Auftreten (p_g) sowie für das Nichtauftreten ($1-p_g$) von Warnung 1. Daraus ergibt sich zum Beispiel für die erste Altersklasse nach Gleichung 6.1 ein Wert von:

$$\log\left(\frac{0,475}{0,525}\right) = -0,1 \,.$$

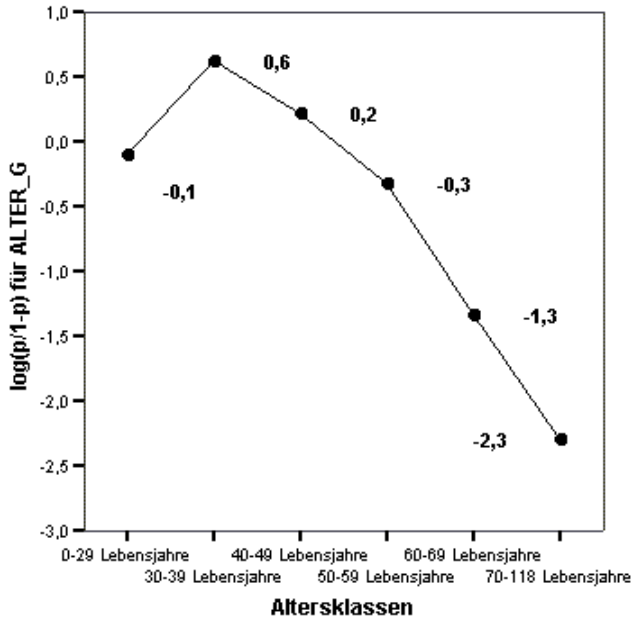
Der SPSS-Output 6.2 zeigt die nach Gleichung 6.1 berechneten Werte, geplottet gegen die Altersklassen, wobei aus dieser Grafik neben jedem Punkt auch der dazugehörige Wert von $\log(p_g/1\text{-}p_g)$ zu entnehmen ist. Zwischen den Punkten ist eine Verbindungslinie eingezeichnet, um den Verlauf deutlich zu machen. Wenn von der ersten Altersklasse der bis unter 30-jährigen Kunden abgesehen wird, kann der Verlauf als linear bezeichnet werden.

SPSS-Output 6.1: Kreuztabelle Warnung 1 und Alter gruppiert mit den beobachteten bedingten relativen Häufigkeiten*

			Warnung 1 (j;n)	
			keine Warnung 1	Warnung 1 vorhanden
Alter gruppiert	0-29 Lebensjahre	% von Alter gruppiert	52,5%	47,5%
	30-39 Lebensjahre	% von Alter gruppiert	34,9%	65,1%
	40-49 Lebensjahre	% von Alter gruppiert	44,7%	55,3%
	50-59 Lebensjahre	% von Alter gruppiert	58,1%	41,9%
	60-69 Lebensjahre	% von Alter gruppiert	79,2%	20,8%
	70 Lebensjahre und älter	% von Alter gruppiert	90,9%	9,1%
Gesamt		% von Alter gruppiert	51,4%	48,6%

* Hierfür wurden nur 5.544 Fälle einbezogen, da bei 140 Kunden in der ersten Teilstichprobe das Alter unbekannt ist.

SPSS-Output 6.2: Scatterplot Altersklassen der Variablen ALTER_G - log(p/1-p)



Eine bessere Erfüllung der Linearitätsbedingung könnte durch eine neue Gruppierung des Alters erreicht werden. Allerdings ist dies in gewisser Weise willkürlich und des weiteren ist nicht sicher, ob nach der Neugruppierung die Regressionskoeffizienten signifikant sein

werden. Daher sollte eventuell mit Hilfe von semiparametrischen Modellen versucht werden, die Schätzung zu verbessern. So kann mit einem generalisierten linearen Modell der Term $\Sigma_j x_{kj}\beta_j$ um eine nichtparametrische Komponente, z.B. für die Variable ALTER_G, erweitert werden.

Als zweite Variable wird das Einkommen auf Linearität untersucht. Die Gruppierung der Variablen EINK in die neue Variable EINK_G ist in der Tabelle 6.1 gezeigt.

Tabelle 6.1: Gruppierung von EINK (Einkommen) in die neue Variable EINK_G

Alter Wert EINK	Neuer Werte EINK_G	Bedeutung
0	0	keine Einkünfte auf dem Privatgirokonto
1 - 5	1	Einkommen* bis 550 DM
6 - 10	2	Einkommen zwischen 551 und 1.050 DM
11 - 15	3	Einkommen zwischen 1.051 und 1.550 DM
16 - 20	4	Einkommen zwischen 1.551 und 2.050DM
21 und mehr	5	Einkommen von 2.051 DM und mehr

* Einkommen - als Durchschnitt der monatlichen Einkünfte auf dem Privatgirokonto in den letzten 12 Monaten.

Der SPSS-Output 6.3 enthält die beobachteten relativen Häufigkeiten für das Ereignis[43], aus denen nach Gleichung 6.1 die für die 6 Gruppen gültigen Werte berechnet und im SPSS-Output 6.4 gegen die Einkommensklassen geplottet werden. Hier kann ein nahezu linearer Verlauf festgestellt werden.

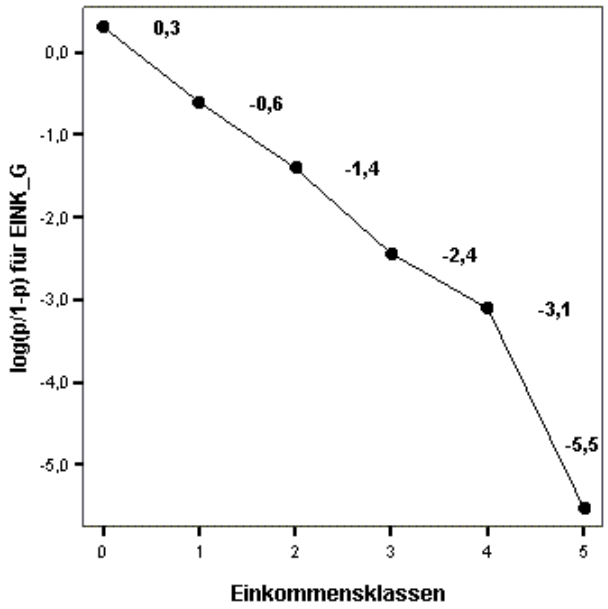
SPSS-Output 6.3: Kreuztabelle Warnung 1 und Einkommen gruppiert mit den beobachteten bedingten relativen Häufigkeiten*

			Warnung 1 (j;n)	
			keine Warnung 1	Warnung 1 vorhanden
Einkommen gruppiert	0	% von Einkommen gruppiert	42,4%	57,6%
	1	% von Einkommen gruppiert	64,7%	35,3%
	2	% von Einkommen gruppiert	80,3%	19,7%
	3	% von Einkommen gruppiert	92,0%	8,0%
	4	% von Einkommen gruppiert	95,7%	4,3%
	5	% von Einkommen gruppiert	99,6%	,4%
Gesamt		% von Einkommen gruppiert	52,6%	47,4%

* Hierfür wurden alle 5.684 Fälle einbezogen.

[43] In der letzten Einkommensklasse, Einkommen von 2.051 DM und mehr, gibt es nur 2 Personen (bzw. 0,4 %) mit Warnung 1. Wegen dieser geringen Häufigkeit werden in dieser Klasse alle höheren Einkommen zusammengefaßt. So wird vermieden, daß die beobachtete relative Häufigkeit einer oder mehrerer Gruppen gleich Null ist, da Logarithmus von Null nicht definiert ist.

SPSS-Output 6.4: Scatterplot Einkommensklassen der Variablen EINK_G - log(p/1-p)



Als dritte und letzte Variable wird VERMOEGE auf Linearität untersucht. Die verwendete Gruppierung ist in der Tabelle 6.2 angegeben. Die beobachteten relativen Häufigkeiten für das Auftreten bzw. Nichtauftreten von Warnung 1 in der jeweiligen Gruppe sind im SPSS-Output 6.5 enthalten. Im SPSS-Output 6.6 sind die Vermögensklassen gegen die Werte von $\log(p_g/1\text{-}p_g)$ geplottet. Der Verlauf ist zwar durch die feinere Gruppierung sprunghafter als bei EINK_G, dennoch zeigt der Scatterplot, daß durchaus von einer Linearität des Vermögens bei der Bank (VERMOEGE) in der link Funktion ausgegangen werden kann.

Tabelle 6.2: Gruppierung von VERMOEGE (Vermögen bei der Bank) in die neue Variable VERM_G

Gruppen-nummer	Bedeutung
0	kein Vermögen*
1	Vermögen zwischen 0,01 und 500,00 DM
2	Vermögen zwischen 500,01 und 1.000,00 DM
3	Vermögen zwischen 1.000,01 und 1.500,00 DM
4	Vermögen zwischen 1.500,01 und 2.000,00 DM
5	Vermögen zwischen 2.000,01 und 2.500,00 DM
6	Vermögen zwischen 2.500,01 und 3.500,00 DM
7	Vermögen zwischen 3.500,01 und 5.000,00 DM
8	Vermögen zwischen 5.000,01 und 10.000,00 DM
9	Vermögen zwischen 10.000,01 und 50.000,00 DM
10	Vermögen von 50.000,01 DM und mehr

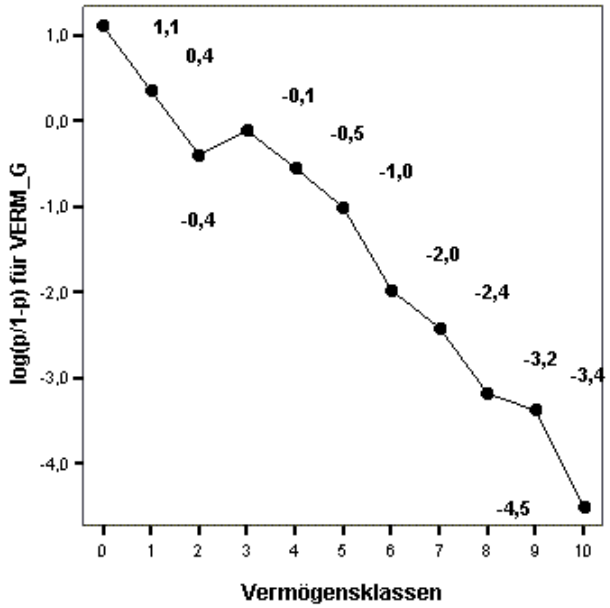
* Vermögen - als Summe aller Guthaben des Kunden auf Konten bei der Bank.

SPSS-Output 6.5: Kreuztabelle Warnung 1 und Vermögen gruppiert mit den beobachteten bedingten relativen Häufigkeiten*

			Warnung 1 (j;n)	
			keine Warnung 1	Warnung 1 vorhanden
VERM_G	0	% von VERM_G	24,8%	75,2%
	1	% von VERM_G	41,3%	58,7%
	2	% von VERM_G	59,8%	40,2%
	3	% von VERM_G	52,9%	47,1%
	4	% von VERM_G	63,4%	36,6%
	5	% von VERM_G	73,3%	26,7%
	6	% von VERM_G	87,8%	12,2%
	7	% von VERM_G	91,8%	8,2%
	8	% von VERM_G	96,0%	4,0%
	9	% von VERM_G	96,7%	3,3%
	10	% von VERM_G	98,9%	1,1%
Gesamt		% von VERM_G	52,6%	47,4%

* Hierfür wurden alle 5.684 Fälle einbezogen.

SPSS-Output 6.6: Scatterplot Vermögensklassen der Variablen VERM_G - log(p/1-p)



6.2. Modellvalidität (Modelvalidity)

Ein Modell ist grundsätzlich besser an die Stichprobe angepaßt, mit deren Hilfe es geschätzt wurde, als an die Grundgesamtheit, aus der diese Stichprobe stammt. Daher ist es wichtig, nachdem die Modellbildung abgeschlossen ist, die Güte der Modellanpassung (goodness-of-fit) zu überprüfen. Zu diesem Zweck wurden vor der Untersuchung die

vorliegenden Daten per Zufallsgenerator in zwei Hälften geteilt, von denen nur der erste Teil zur Modellschätzung verwendet wurde. Der zweite Teil der Datei, der weitere 5.695 einzelne Privatkunden enthält, wird nun zur Überprüfung der Modellvalidität verwendet. Allerdings werden die Kunden, bei denen das Alter und/oder die Region unbekannt sind, hierbei nicht berücksichtigt. Daher verbleiben nur 4.889 Personen für die Modellvalidierung. Diese wird auf folgende Weise durchgeführt: Für jede der 4.889 Personen wird die Wahrscheinlichkeit für das Auftreten von Warnung 1 mittels der im endgültigen Modell (SPSS-Output 5.14) geschätzten Regressionskoeffizienten folgendermaßen bestimmt:

$$\hat{\pi}_k = \frac{1}{1+e^{-[\,-0{,}8820-0{,}7566*(\text{sex})\;+0{,}5843*(\text{region\_1})+\;\ldots\;-0{,}0003*(\text{vermoege})\,]}}\;.$$

Danach werden alle Personen in zwei Gruppen geteilt. Für die Kunden, bei denen $\hat{\pi}_k$ kleiner als eine vorgegebene Schwelle von z.B. 0,5 geschätzt wird, wird das Nichtauftreten von Warnung 1 vorhergesagt. Für alle anderen wird das Auftreten von Warnung 1 erwartet. Anschließend wird das Vorhergesagte mit dem tatsächlich beobachteten Auftreten des Ereignisses verglichen. So kann festgestellt werden, wie oft eine falsche Vorhersage bzw. wie oft aufgrund des Modells eine richtige Vorhersage getroffen wurde.

Diese Vorgehensweise wird hier mit drei verschiedenen Schwellenwerten durchgeführt. Zunächst wird als solcher Wert 0,5 gewählt, der auch in den Klassifikationstabellen der SPSS-Outputs als "Cut Value" benutzt wurde. Damit werden alle Personen mit $\hat{\pi}_k < 0{,}5$ zu derjenigen Gruppe ohne Warnung 1 und die Personen mit $\hat{\pi}_k \geq 0{,}5$ zu der Gruppe mit mindestens einer Warnung 1 zugeordnet. Gleichermaßen wird auch mit den Schwellenwerten in Höhe von 0,25 und 0,75 verfahren. Der Tabelle 6.3 ist zu entnehmen, wie viele Personen bei den verschiedenen Schwellenwerten richtig und wie viele falsch klassifiziert wurden. In der letzten Spalte ist angegeben, in wieviel Prozent aller Fälle das Auftreten bzw. Nichtauftreten der Warnung 1 richtig vorhergesagt wurde.

Das schlechteste Ergebnis liefert die Vorhersage mit einem Schwellenwert von 0,75. Hierbei stimmt die Vorhersage bei 67,23 % aller Fälle mit dem beobachteten Zustand überein. Eine bessere Vorhersage wird bei einem Schwellenwert von 0,25 erreicht. Dennoch ist als Schwellenwert 0,5 zu empfehlen, da bei diesem mit 76,36 % eindeutig die zutreffendste Vorhersage gemacht wurde. In diesem Fall wurde bei 2.152 Kunden das

Auftreten und bei 1.581 das Nichtauftreten von Warnung 1 richtig prognostiziert. Eine falsche Zuordnung der Kunden erfolgte in 23,64 % der Fälle.

Das Ergebnis, daß 76,36 % der Fälle richtig zugeordnet wurden, ist insoweit zufriedenstellend, als daß dieses als adäquat vorgestellte Modell in der Stichprobe, mit deren Hilfe es geschätzt wurde, 78,06 % der Fälle richtig zuordnet[44] und in der zweiten Stichprobe (validation sample) nahezu gleich gut vorhersagt.

Tabelle 6.3: Modellvalidität mit verschiedenen Schwellenwerten

Schwellen-wert	beobachtet		vorhergesagt		falsch zugeordnet pro Kategorie	gesamt falsch	gesamt richtig
			0	**1**			
0,5	**0**	2452	1581	871	35,52 %		
	1	2437	285	2152	11,69 %	23,64 %	**76,36 %**
0,25	**0**	2452	1153	1299	52,98 %		
	1	2437	57	2380	2,34 %	27,74 %	**72,26 %**
0,75	**0**	2452	2201	251	10,24 %		
	1	2437	1351	1086	55,44 %	32,77 %	**67,23 %**

[44] Siehe Klassifikationstabelle im SPSS-Output 5.14.

7. Zusammenfassung und Ausblick

Ziel der vorliegenden Arbeit war es zu untersuchen, welchen Einfluß bestimmte Faktoren auf das Auftreten von Zahlungsschwierigkeiten bei den Privatkunden einer Bank ausüben. Als Methode wurde hierfür die logistische Regression angewandt. Es hat sich gezeigt, daß diese Methode, die im Rahmen des Credit Scoring neben der Diskriminanzanalyse eine wichtige Rolle spielt, auch bei einer Problemstellung außerhalb der unmittelbaren Kreditwürdigkeitsprüfung schnell und einfach zufriedenstellende Resultate liefert.

Als Ergebnis wurde ein multiples Modell präsentiert, in dem sich die Variablen bekanntes Einkommen, Vermögen bei der Bank und Geschlecht als die wichtigsten Einflußfaktoren unter den zur Verfügung stehenden Merkmalen herausstellten, gefolgt von dem Alter des Kunden und der Region, in der er wohnt. Allerdings wurde festgestellt, daß es weitere wichtige Faktoren für das Auftreten von Zahlungsschwierigkeiten bei Privatkunden geben muß, deren Einschluß in die Regressionsschätzung eine vermutlich bessere Anpassung an die Daten ermöglichen wird.

Einige bei der Untersuchung festgestellte Ergebnisse erscheinen selbstverständlich oder trivial. Daher könnte die Frage entstehen, wozu die statistische Analyse notwendig war. Einerseits gibt die multiple logistische Regression die Möglichkeit, für jeden Kunden mit seinen individuellen Ausprägungen der X-Variablen die Wahrscheinlichkeit für das Auftreten einer Warnung 1 zu schätzen und diese individuellen Wahrscheinlichkeiten miteinander zu vergleichen. Dies ist ohne das Modell kaum möglich. Andererseits ist eine statistische Untersuchung auch dann erfolgreich, wenn sie nur die Vermutungen bestätigt und damit keine neuen Erkenntnisse bringt. Dann kann weiter "auf das Gefühl" vertraut werden.

Das aus der Analyse resultierende multiple Modell kann u.a. im Sinne eines Credit Scoring als eine quantitative Entscheidungshilfe für die Mitarbeiter der Bank im Bereich Kundenbetreuung genutzt werden. So kann z.B. die Entscheidung, ob ein Kunde ein Darlehen erhält, mit Hilfe der anhand des Modells berechneten Wahrscheinlichkeiten für das Auftreten von Zahlungsschwierigkeiten getroffen werden. Dieses Verfahren

kann eine unterstützende Rolle spielen, indem die Mitarbeiter auf relativ objektive Kriterien zurückgreifen können, ohne daß sie auf ihre Erfahrung und "Gefühl" verzichten müssen.

Denkbar ist aber auch die Möglichkeit, daß die Entscheidungsfindung nur anhand der geschätzten Wahrscheinlichkeiten erfolgt. Das würde bedeuten, daß keine subjektiven Entscheidungsregeln einbezogen werden. Somit würden z.B. in bezug auf Darlehen nur die Kunden, für die eine Wahrscheinlichkeit unter 0,5 geschätzt wurde, einen Kredit erhalten. Ein solches Vorgehen hat auch den Vorteil, daß es mit einer einfachen Evaluierung verbunden ist. Somit kann das Modell als Entscheidungsgrundlage leicht verbessert werden, da genau bekannt und quantifizierbar ist, was verändert wurde. Allerdings ist es fraglich, inwieweit Entscheidungen im Bereich Kundenservice "automatisch" getroffen werden sollen. Dies kann und wird sich wahrscheinlich negativ auswirken, denn es verleiht den Eindruck eines Automatismus. Außerdem darf nicht vergessen werden, daß die Ergebnisse statistischer Verfahren keine absoluten Wahrheiten sind, sondern immer unter dem Blickwinkel der Stärken und Schwächen des jeweiligen Verfahrens zu sehen, zu interpretieren und zu berücksichtigen sind. Dies kommt einer Mischung aus "Modell + Erfahrung" als Grundlage für die jeweilige Entscheidung entgegen, auch wenn die Evaluierung eines solchen Vorgehens im Vergleich zu einer Entscheidung nur auf Basis des Modells nicht so einfach ist, denn sie enthält unter dem Begriff Erfahrung einen nicht genau quantifizierbaren Teil. Es ist sicherlich von Vorteil, wenn im Einzelfall trotz der statistischen Ergebnisse individuell und nicht unbedingt automatisch entschieden werden kann. So kann z.B. die Möglichkeit bestehen, bei einem negativen Schätzergebnis dennoch ein Darlehen zu gewähren, wenn eine umfangreichere Prüfung und/oder sogenannte weiche Kriterien dafür sprechen.

Unabhängig davon für welche Vorgehensweise man sich entscheidet, kann die Berücksichtigung statistisch überprüfter Einflußgrößen in bezug auf das Auftreten von Zahlungsschwierigkeiten zu einer stärker kundenorientierten Betreuung führen, die jedoch besser die individuellen Risiken berücksichtigt. Außerdem kann eine Standardisierung bei der Vergabe von Warnungen erzielt werden.

Ein weiteres Einsatzfeld des hier vorgestellten Verfahrens kann in der Preisbildung gesehen werden. Wenn höhere Risiken geschätzt werden, können diese mit höheren

Preisen verbunden werden. Das heißt, daß bei einem Kunden mit höher geschätzter Wahrscheinlichkeit für das Auftreten von Zahlungsschwierigkeiten z.B. für ein Darlehen höhere Zinsen berechnet werden und umgekehrt. Somit kann das Modell im Bereich der individuellen Preisgestaltung hilfreich sein, die ihrerseits im harten Wettbewerb entscheidend ist.

Abschließend kann gesagt werden, daß die Vorteile des hier dargestellten Verfahrens eindeutig in den schnell verfügbaren Ergebnissen liegen, die außerdem übersichtlich, verständlich, einfach zu handhaben und zuverlässig sind. Somit kann und wird die logistische Regression in Zukunft in noch mehr Bereichen zur Risikoeinschätzung eingesetzt werden.

Anhang A

Testentscheidungen unter SPSS

SPSS gibt bei der Durchführung von Tests eine sogenannte Überschreitungswahrscheinlichkeit "Sign" oder α_{sign} an. Dies ermöglicht es, in Abhängigkeit von der Art des zu untersuchenden Sachverhaltes selbst zu entscheiden, bei welcher Irrtumswahrscheinlichkeit die Nullhypothese verworfen werden soll. Je größer die Sicherheit vor einer Fehlentscheidung sein soll, desto geringer muß α gewählt werden.[45]

Da das Programm das vom Benutzer vorgegebene Signifikanzniveau α nicht kennt, wird α_{sign} im Output mit angegeben. Dieses α_{sign} gibt die Wahrscheinlichkeit an, "... einen Wert der Teststatistik zu beobachten, der wenigstens so groß ist wie der aus der Stichprobe berechnete Wert, d.h., es ist das exakte Signifikanzniveau, mit der [eigentlich: dem] die Nullhypothese verworfen werden könnte ... ".[46] Anders ausgedrückt: "Wollte man die Nullhypothese verwerfen, würde das mit mehr als ... [α_{sign}] eine Fehlentscheidung sein."[47]

Der Benutzer muß zunächst α für sich festlegen, um eventuelle nachträgliche Manipulationen zu vermeiden. Nach der Testdurchführung muß er α und α_{sign} vergleichen. Dabei gilt, wenn $\alpha_{sign} < \alpha$ ist, dann ist die Nullhypothese (H_0) aufgrund der Stichprobe vom Umfang N und zum vorgegebenen Signifikanzniveau α abzulehnen. Für den Fall, daß $\alpha_{sign} \geq \alpha$ ist, ist H_0 beizubehalten.

Zur Überprüfung von Nullhypothesen wird in dieser Arbeit jeweils ein Signifikanzniveau von 5 % ($\alpha = 0{,}05$) vorgegeben.

[45] Vgl. Bühl, A., Zöfel, P. (1998), S. 101.
[46] Rönz, B. (1997b), S. 155.
[47] Rönz, B. (1997a), S. 140.

Anhang B

Chi-Quadrat-Unabhängigkeitstest nach Pearson[48]

Mit Hilfe des Chi-Quadrat-Unabhängigkeitstests nach Pearson wird überprüft, ob sich Wechselwirkungen zwischen zwei Zufallsvariablen X und Y quantitativ nachweisen lassen. Die Nullhypothese lautet demnach auf Unabhängigkeit beider Variablen. Der Test stellt keine Anforderungen an das Skalenniveau der Variablen, dennoch müssen folgende Voraussetzungen erfüllt sein:

- unabhängige Beobachtungen,
- vollständige Aufteilung der Variablenausprägungen,
- die erwartete absolute Häufigkeit für das Auftreten von (x_j, y_k) darf nicht kleiner als 1 bzw. nicht mehr als 20 % dieser erwarteten absoluten Häufigkeiten dürfen kleiner als 5 sein.[49]

Die Teststatistik wird wie folgt berechnet:

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(h_{jk} - \hat{e}_{jk}\right)^2}{\hat{e}_{jk}} \quad , \qquad \text{Gleichung A1}$$

worin j = 1,...,J der Laufindex für die Ausprägungen der Variablen X und k = 1,...,K der Laufindex für die Ausprägungen der Variablen Y ist. Bei diesem Test findet ein Vergleich der beobachteten absoluten Häufigkeiten (h_{jk}) mit den unter H_0 zu erwartenden absoluten Häufigkeiten ($\hat{e}_{jk}$) für das Auftreten von (x_j, y_k) statt. Die quadrierten Differenzen werden durch den Nenner relativiert. Unter H_0 ist die Teststatistik approximativ χ^2-verteilt, wobei diese Approximation um so besser ist, je größer der Stichprobenumfang ist. Die Anzahl der Freiheitsgrade ist df = (J-1)(K-1). Die Nullhypothese wird abgelehnt, falls $\chi^2 > \chi^2_{1-\alpha;\,df}$ ist. Hierbei ist $\chi^2_{1-\alpha;\,df}$ das Quantil der Ordnung 1-α der χ^2-Verteilung mit df-Freiheitsgraden und α das vorgegebene Signifikanzniveau.

[48] Siehe u.a. Rönz, B. (1998), S. 42 ff.; Rönz, B., Strohe, H. G., S. 68 f.; Bosch, K., S. 384 ff.

[49] Diese Bedingung ist äquivalent zu der, daß die erwartete absolute Häufigkeit einer Zelle nicht kleiner als 1 sein darf bzw. nicht mehr als 20 % der Zellen eine erwartete absolute Häufigkeit kleiner 5 aufweisen dürfen. Mit Zellen sind hier die Zellen einer Kontingenztabelle der beiden Variablen X und Y gemeint. Diese Zellen enthalten u.a. die oben genannten erwarteten absoluten Häufigkeiten $\hat{e}_{jk}$.

Cramer-V[50]

Das Cramer-V ist ein Maß für die Stärke des Zusammenhangs zwischen zwei Variablen mit nominalem oder höherem Skalenniveau. Daher sagt dieses Maß nichts über Art und Richtung der Beziehung aus. Der Wert wird wie folgt ermittelt:

$$V = \sqrt{\frac{\chi^2}{n * (c-1)}} \ ,$$

worin χ^2 die in Gleichung A1 gezeigte χ^2-Statistik, n der Stichprobenumfang und c das Minimum aus der Anzahl der Ausprägungen beider Variablen (c = min(J;K)) ist. Es gilt $0 \leq V \leq 1$, d.h. je näher der Wert an 1 ist, um so stärker ist die Beziehung zwischen den Variablen.

[50] Siehe u.a. Hartung, J., Elpelt, B., Klösener, K.-H., S. 452.

Literaturverzeichnis

Bosch, K. = Bosch, K.: Statistik-Taschenbuch, Oldenbourg, München, Wien 1998

Bühl, A., Zöfel, P. (1998) = Bühl, A., Zöfel, P.: SPSS Version 8: Einführung in die moderne Datenanalyse unter Windows, Addison-Wesley-Longman, Bonn 1998

Bühl, A., Zöfel, P. (2000) = Bühl, A., Zöfel, P.: SPSS Version 10: Einführung in die moderne Datenanalyse unter Windows, Addison-Wesley, München 2000

Collett, D. = Collett, D.: Modelling binary data, Chapman & Hall, London et. al. 1991

Dobson, A. J. = Dobson, A. J.: An introduction to generalized linear models, Chapman & Hall, London et. al. 1991

Fahrmeir, L., Hamerle, A. = Fahrmeir, L., Hamerle, A.: Multivariate statistische Verfahren, de Gruyter, Berlin, New York 1984

Hartung, J., Elpelt, B., Klösener, K.-H. = Hartung, J., Elpelt, B., Klösener, K.-H.: Statistik: Lehr- und Handbuch der angewandten Statistik; mit zahlreichen, vollständig durchgerechneten Beispielen, Oldenbourg, München, Wien 1993

Hosmer, D. W., Lemeshow, S. = Hosmer, D. W., Lemeshow, S.: Applied logistic regression, John Wiley & Sons, New York et. al. 1989

Kleinbaum, D. G. = Kleinbaum, D. G.: Logistic regression: a self-learning text, Springer-Verlag, New York et. al. 1994

Rönz, B. (1997a) = Rönz, B.: Script Computergestützte Statistik I, Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät der Humboldt-Universität zu Berlin, 1997

Rönz, B. (1997b) = Rönz, B.: Script Generalisierte lineare Modelle, Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät der Humboldt-Universität zu Berlin, 1997

Rönz, B. (1998) = Rönz, B.: Script Computergestützte Statistik II, Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät der Humboldt-Universität zu Berlin, 1998

Rönz, B., Förster, E. = Rönz, B., Förster, E.: Regressions- und Korrelationsanalyse: Grundlagen, Methoden, Beispiele, Gabler, Wiesbaden 1992

Rönz, B., Strohe, H. G. = Rönz, B., Strohe, H. G. (Hrsg.): Lexikon Statistik, Gabler, Wiesbaden 1994

SPSS Regression Models™ 9.0 = SPSS Regression Models™ 9.0, SPSS Inc., Chicago 1999

Erklärung zur Urheberschaft

Hiermit erkläre ich, daß ich die vorliegende Arbeit allein und nur unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Nadia Landmann

Berlin, 19. Dezember 2000