Kapitel 25
Diskriminanzanalyse

Ziel der Diskriminanzanalyse ist es, die Werte einer abhängigen (zu erklärenden) Variablen durch die Werte einer oder mehrerer unabhängigen (erklärenden) Variablen zu erklären. Dabei sollen nicht nur Zusammenhänge zwischen den Variablen entdeckt, sondern auch unbekannte Werte der abhängigen Variablen anhand der Werte aus den erklärenden Variablen vorhergesagt werden. Insoweit stimmt das Ziel der Diskriminanzanalyse mit dem einer Regressionsanalyse überein. Der wesentliche Unterschied zwischen den beiden Verfahren besteht in der Art der Werte der abhängigen Variablen. Während mit einer Regressionsanalyse nur abhängige Variablen mit Intervallskalenniveau untersucht werden können, versucht die Diskriminanzanalyse, eine Zuordnung von Fällen zu einer von mehreren alternativen Gruppen vorzunehmen. Die Werte der abhängigen Variablen geben also lediglich eine Gruppenzugehörigkeit an und besitzen damit Nominal- oder Ordinalskalenniveau.

Ein typischer Anwendungsfall der Diskriminanzanalyse ist die Kreditwürdigkeitsprüfung. Um die Kreditwürdigkeit von Unternehmen zu bestimmen, kann eine Geschäftsbank auf der Basis ihrer in bisherigen Kreditgeschäften gesammelten Erfahrungen versuchen, aus dem Umsatz, der Zuwachsrate des Umsatzes, der Umsatzrentabilität, dem Verhältnis von Eigen- und Fremdkapital, der Branche und andere Merkmalen eines Unternehmens Rückschlüsse auf dessen Kreditwürdigkeit zu ziehen. Mit Hilfe der Diskriminanzanalyse kann ein Unternehmen anhand dieser Merkmale dann zum Beispiel einer von verschiedenen Risikogruppen zugeordnet werden.

25.1 Das Verfahren der Diskriminanzanalyse

25.1.1 Diskriminanzfunktion berechnen

Beispiel

Im folgenden wird mit Hilfe der Diskriminanzanalyse versucht, das Wahlverhalten von wahlberechtigten Personen zu prognostizieren. Hierzu werden die Daten aus der Datendatei *allbus.sav* von der Begleit-CD verwendet.[267] Die Datei enthält Angaben darüber, welche Partei eine Person zum Zeitpunkt der Befragung gewählt hätte, wenn zu dem Zeitpunkt Bundestagswahlen stattgefunden hätten. Mit der Diskriminanzanalyse soll nun versucht werden, die von einer Person präferierte Partei anhand verschiedener Personenmerkmale vorherzusagen. Dabei werden als Parteien zunächst lediglich die CDU/CSU und die SPD berücksichtigt. Als Personenmerkmale sollen das Alter, der höchste von einer Person erreichte Schulabschluß, das Einkommen, das Erhebungsgebiet (Ost- oder Westdeutschland) und die Selbsteinstufung der Befragten auf einer politischen Links-Rechts-Skala betrachtet werden.[268]

Die Diskriminanzfunktion

Die Diskriminanzanalyse läßt sich gedanklich in zwei Schritte unterteilen. Im ersten Schritt wird eine Diskriminanzfunktion geschätzt, der zweite Schritt nimmt eine Klassifizierung der Fälle und damit eine Unterteilung in einzelne Gruppen vor. Die Schätzung der Diskriminanzfunktion hat große Ähnlichkeit mit der Schätzung einer Regressionsfunktion in der Regressionsanalyse. Die Eigenart der Diskriminanzanalyse ergibt sich erst aus dem zweiten Schritt. Dieser dient dazu, aus den stetigen Werten der erklärenden Variablen diskrete Werte und damit Gruppenzugehörigkeiten der abhängigen Variablen zu berechnen.

Die zu schätzende Diskriminanzfunktion hat die allgemeine Form:

$$D = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_n \cdot X_n$$

Dabei bezeichnen die X_i die verschiedenen erklärenden Variablen und die b_i die Koeffizienten, mit denen die Variablen in die Diskriminanzfunktion eingehen. Der erste Schritt der Diskriminanzanalyse besteht nun darin, diese Koeffizienten zu schätzen.

---

[267] Zu den Datendateien von der Begleit-CD siehe im einzelnen Kapitel *1, Überblick*.

[268] Es mag zunächst trivial erscheinen, das Wahlverhalten durch die von den Befragten selbst vorgenommene politische Einordnung auf der Links-Rechts-Skala vorherzusagen, allerdings besteht zwischen der Einstufung auf dieser Skala und der Entscheidung für eine politische Partei nur ein begrenzter Zusammenhang. Würde man die von einer Person gewählte Partei ausschließlich anhand der Links-Rechts-Selbsteinstufung vorhersagen, erhielte man keine befriedigenden Ergebnisse.

Die Funktion einer Regressionsanalyse läßt sich in der allgemeinen Form genauso darstellen wie die Diskriminanzfunktion. Die b_i würden in dem Fall die Regressionskoeffizienten darstellen. Diese werden bei der Regressionsanalyse nach der Methode der kleinsten Quadrate bestimmt. Dies bedeutet, daß die Koeffizienten so gewählt werden, daß die Summe der quadrierten Abweichungen der geschätzten Funktionswerte von den tatsächlichen Werten der abhängigen Variablen minimal ist. Dieses Verfahren wird für die Diskriminanzanalyse nicht übernommen. Vielmehr wird bereits bei der Schätzung der Diskriminanzfunktion berücksichtigt, daß im zweiten Schritt der Analyse aus den stetigen Werten Rückschlüsse auf eine diskrete Gruppenzugehörigkeit gezogen werden sollen. Daher wird für die Schätzung der Koeffizienten eine Methode verwendet, die bereits von diskreten Gruppen ausgeht und somit einen Zusammenhang zwischen den Funktionswerten und einzelnen Gruppen wahrscheinlich werden läßt. Die Koeffizienten der Diskriminanzfunktion werden so bestimmt, daß der folgende Quotient maximal wird.

$$\frac{\text{Quadratsumme der Funktionswerte zwischen den Gruppen}}{\text{Quadratsumme der Funktionswerte innerhalb der Gruppen}}$$

Diese Vorgehensweise bewirkt, daß sich die Funktionswerte der Diskriminanzfunktion von Fällen verschiedener Gruppen möglichst deutlich voneinander unterscheiden.

Einstellungen des Beispiels

Abbildung 25.1 zeigt einen geringen Teil der Ergebnisse für das oben beschriebene Beispiel. Die folgenden Einstellungen liefern die in Abbildung 25.1 sowie die in den nachfolgenden Abbildungen gezeigten Ergebnisse.

- **Daten:** Als Datengrundlage dient die Datei *allbus.sav* von der Begleit-CD. Die Daten in der Datei sind gewichtet mit der Variablen *v434*.[269] Da im folgenden nur die Parteien CDU/CSU und SPD betrachtet werden sollen, sind alle Fälle von Personen, die andere Parteien gewählt haben, ausgeschlossen. Die verbleibenden Fälle sind dadurch gekennzeichnet, daß sie in der Variablen *v325* (Wahlabsicht bei der Bundestagswahl) den Wert *1* oder *2* aufweisen.[270]

[269] Diese Gewichtung dient der Korrektur einer überproportionalen Berücksichtigung von Personen aus den neuen Bundesländern, die bei der Datenerhebung bewußt vorgenommen wurde. Um die Gewichtung einzuschalten, wählen Sie den Befehl DATEN, FÄLLE GEWICHTEN. Dieser Befehl öffnet ein Dialogfeld, in dem die Option *Fälle gewichten mit* zu wählen und die Variable *v434* in das Feld *Häufigkeitsvariable* zu verschieben ist. Anschließend können Sie das Dialogfeld mit der Schaltfläche *OK* schließen.

[270] Um alle anderen Fälle auszuschließen, verwenden Sie den Befehl DATEN, FÄLLE AUSWÄHLEN. Wählen Sie in dem damit geöffneten Dialogfeld die Option *Falls Bedingung zutrifft*, und klicken Sie auf die zu dieser Option gehörenden Schaltfläche *Falls*, die ein weiteres Dialogfeld öffnet. Geben Sie dort in dem großen Eingabefeld die Bedingung *v325 = 1 | v325 = 2* ein. Danach können Sie dieses Dialogfeld mit *Weiter* und das Hauptdialogfeld mit *OK* schließen. Stellen Sie zuvor aber sicher, daß im Hauptdialogfeld in der Gruppe *Nicht ausgewählte Fälle* die Option *Filtern* markiert ist. Möchten Sie nach Durchführung der Diskriminanzanalyse erreichen,

- **Befehl:** Um eine Diskriminanzanalyse durchzuführen, wählen Sie den Befehl

STATISTIK
KLASSIFIZIEREN ▸
DISKRIMINANZANALYSE...

- **Gruppenvariable:** Fügen Sie die Variable *v325* in das Feld *Gruppenvariable* ein, und wählen Sie anschließend die Schaltfläche *Bereich definieren*. Geben Sie in dem damit geöffneten Dialogfeld in das Feld *Minimum* den Wert *1* und in das Feld *Maximum* den Wert *2* ein.
- **Unabhängige Variablen:** Verschieben Sie die folgenden Variablen in das Feld *Unabhängige Variable(n)*:
  - *v3* (Erhebungsgebiet, Ost-West)
  - *v37* (Alter)
  - *v112* (Links-Rechts-Selbsteinstufung)
  - *v142* (Schulabschluß)
  - *v261* (Einkommen)
- **Statistiken:** Wählen Sie in dem Dialogfeld der Schaltfläche *Statistik* die Option *Nicht Standardisiert* aus der Gruppe *Funktionskoeffizienten* sowie die Optionen *Mittelwert* und *Univariate ANOVA* aus der Gruppe *Deskriptive Statistiken*.
- **Klassifizieren:** Öffnen Sie das Dialogfeld der Schaltfläche *Klassifizieren*, und kreuzen Sie dort in der Gruppe *Anzeigen* die Optionen *Fallweise Ergebnisse* und *Zusammenfassende Tabelle* an. Wählen Sie zusätzlich die Option *Fälle beschränken auf die ersten*, und geben Sie in das zugehörige Eingabefeld den Wert *30* ein. Wählen Sie weiterhin in der Gruppe *A-priori-Wahrscheinlichkeit* die Option *Aus der Gruppengröße berechnen*.

Bei allen anderen Optionen werden die Voreinstellungen verwendet. Die Dialogfelder im Abschnitt *25.5, Einstellungen der Diskriminanzanalyse* , S. 626 zeigen die beschriebenen Einstellungen. Abbildung 25.1 gibt die Tabelle *Kanonische Diskriminanzfunktionskoeffizienten* wieder.

Funktionskoeffizienten und Funktionswerte

Die Tabelle gibt die geschätzten Koeffizienten der Diskriminanzfunktion, deren allgemeine Form oben auf S. 592 dargestellt ist, für das hier betrachtete Beispiel an. So geht etwa der Schulabschluß mit einem Koeffizienten von 0,05 in die Funktion ein, während die Links-Rechts-Selbsteinstufung mit dem Koeffizienten 0,586 berücksichtigt wird. Die geschätzte Diskriminanzfunktion läßt sich damit schreiben als:

---

daß wieder alle Fälle der Datendatei aktiviert werden, wählen Sie in dem Dialogfeld des Befehls DATEN, FÄLLE AUSWÄHLEN die Option *Alle Fälle*, und schließen Sie das Dialogfeld mit *OK*.

$$D = -2{,}754 - 0{,}03 \cdot Gebiet - 0{,}017 \cdot Alter + 0{,}586 \cdot Selbsteinstufung + 0{,}05 \cdot Schulabschluß + 0{,}000 \cdot Nettoeinkommen$$[271]

In diese Diskriminanzfunktion lassen sich unmittelbar die Variablenwerte der fünf unabhängigen Variablen einsetzen. Beispielsweise weisen diese Variablen im ersten Fall die folgenden Werten auf:

Gebiet (v3)	Alter (v37)	Selbsteinstufung (v112)	Schulabschluß (v142)	Einkommen (v261)
1	35	3	5	4000

Damit ergibt sich für den ersten Fall der folgende Funktionswert der Diskriminanzfunktion:

$$D = -2{,}754 - 0{,}03 \cdot 1 - 0{,}017 \cdot 35 + 0{,}586 \cdot 3 + 0{,}05 \cdot 5 + 0{,}000066 \cdot 4000$$
$$= -1{,}107$$

Auf diese Weise werden die Funktionswerte bei der Diskriminanzanalyse für jeden Fall der Datendatei berechnet. Per Voreinstellung werden diese Funktionswerte im Output nicht einzeln ausgewiesen, sie können jedoch explizit mit der Option *Ergebnisse für jeden Fall* in dem Dialogfeld der Schaltfläche *Klassifizieren* angefordert werden. Dies ist in den oben beschriebenen Einstellungen geschehen. Abbildung 25.2 zeigt die Tabelle, die man mit dieser Option erhält. Dort wird für den ersten Fall der Funktionswert -1,114 angegeben. Der von uns berechnete Wert von -1,107 weicht geringfügig davon ab, weil wir mit gerundeten Werten gerechnet haben.

Kanonische Diskriminanzfunktionskoeffizienten

	Funktion
	1
ERHEBUNGSGEBIET: WEST - OST	-,030
ALTER: BEFRAGTE<R>	-,017
LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	,586
ALLGEMEINER SCHULABSCHLUSS	,050
BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	,000
(Konstant)	-2,754

Nicht-standardisierte Koeffizienten

Abbildung 25.1: Unstandardisierte Koeffizienten der Diskriminanzfunktion

[271] Der Koeffizient des Einkommens ist sehr gering, er ist aber nicht tatsächlich gleich null, sondern beträgt ungefähr 0,000066. Diesen Wert kann man sich im Ausgabenavigator anzeigen lassen.

Weitere Schritte der Diskriminanzanalyse

Nachdem nun anhand der Diskriminanzfunktion für jeden Fall ein Funktionswert berechnet werden kann, besteht der zweite Schritt der Diskriminanzanalyse darin, die einzelnen Fälle mit Hilfe der Funktionswerte einer der verschiedenen Gruppen aus der abhängigen Variablen, in diesem Fall also entweder den Parteien CDU/CSU oder der Partei SPD, zuzuordnen. Im Anschluß daran lassen sich die mit Hilfe der Funktionswerte vorgenommenen Zuordnungen der Fälle zu den verschiedenen Gruppen mit den tatsächlichen Gruppenzugehörigkeiten vergleichen. Ein solcher Vergleich kann einen ersten Eindruck von der Güte der Diskriminanzschätzung vermitteln.

25.1.2 Klassifizieren

Bei der Diskriminanzanalyse wird angenommen, daß ein Zusammenhang zwischen den erklärenden und der unabhängigen Variablen besteht. Dieser Zusammenhang soll in der Diskriminanzfunktion so ausgenutzt werden, daß sich für Fälle, die unterschiedlichen Gruppen der abhängigen Variablen zuzuordnen sind, möglichst unterschiedliche Funktionswerte ergeben. Auch wenn dies grundsätzlich gelingt, wird sich sehr selten eine Funktion finden, deren Werte eine eindeutige Gruppenzuordnung ermöglichen. Dies würde bedeuten, daß ein perfekter Zusammenhang zwischen den unabhängigen und der abhängigen Variablen besteht. Für die Präferenzen bezüglich der politischen Partei wird sich zum Beispiel zeigen, daß kleine Funktionswerte tendenziell anzeigen, daß eine Person zur SPD neigt, während große Funktionswerte auf eine Präferenz bezüglich der CDU/CSU hinweisen. Es läßt sich allerdings nicht ein eindeutiger Grenzwert bestimmen, der die Befragten eindeutig den beiden Gruppen zuordnet. Es ist durchaus möglich, daß sich für eine Person ein niedriger Funktionswert ergibt und diese dennoch die CDU/CSU der SPD vorziehen würde. Entsprechend kann auch umgekehrt ein hoher Funktionswert für einen SPD-Anhänger berechnet werden. Ein bestimmter Funktionswert kann sich somit für jede der verschiedenen (in diesem Fall der beiden) Gruppen ergeben, allerdings tritt er bei den verschiedenen Gruppen mit unterschiedlicher Wahrscheinlichkeit ein. Für einen SPD-Wähler ergibt sich ein niedriger Funktionswert mit einer größeren Wahrscheinlichkeit als für einen CDU-Anhänger. Diese unterschiedlichen Eintrittswahrscheinlichkeiten der Funktionswerte bei den verschiedenen Gruppen werden bei der Zuordnung der Fälle zu den Gruppen ausgenutzt. Die Wahrscheinlichkeit, mit der ein Fall, für den sich der Funktionswert D ergibt, der Gruppe G_i angehört, kann als

$$P(G_i \mid D)$$

geschrieben werden. Diese Wahrscheinlichkeiten werden für jeden Fall jeweils für jede der potentiellen Gruppen berechnet. Die verschiedenen Wahrscheinlichkeiten eines Falles addieren sich stets zu 1, da jeder Fall mit Sicherheit genau einer Gruppe entstammt.

Bei der konkreten Berechnung der Wahrscheinlichkeiten eines Falles wird der *Satz von Bayes* ausgenutzt, der es erlaubt, die Wahrscheinlichkeit für die Grup-

penzugehörigkeit eines Funktionswertes anhand zweier bekannter Wahrscheinlichkeiten zu ermitteln:

$$P(G_i \mid D) = \frac{P(D \mid G_i) \cdot P(G_i)}{\sum_{i=1}^{g} P(D \mid G_i) \cdot P(G_i)}$$

P(G$_i$ | D) ist die Wahrscheinlichkeit für die Zugehörigkeit eines Falles zur Gruppe *G$_i$* bei gegebenem Funktionswert *D*. Diese Wahrscheinlichkeit wird als A-posteriori-Wahrscheinlichkeit bezeichnet. Sie läßt sich aus den Wahrscheinlichkeiten *P(G$_i$)* (A-priori-Wahrscheinlichkeit) und *P(D | G$_i$)* (bedingte Wahrscheinlichkeit) berechnen. Diese beiden Wahrscheinlichkeiten können folgendermaßen interpretiert werden:

- **A-priori-Wahrscheinlichkeit P(G$_i$):** Dies ist die Wahrscheinlichkeit für eine Gruppenzugehörigkeit, von der man ausgehen muß, wenn keinerlei weitere Informationen zur Verfügung stehen. Nimmt man zum Beispiel eine Zuordnung zu drei Gruppen vor, so beträgt die Wahrscheinlichkeit für die Zugehörigkeit eines Falles zu einer bestimmten Gruppe 1/3, also ungefähr 33%, sofern man über keinerlei zusätzliche Informationen verfügt. Alternativ kann man davon ausgehen, die relativen Häufigkeiten, mit denen die einzelnen Gruppen in der vorliegenden Stichprobe vertreten sind, seien repräsentativ für die Grundgesamtheit. Dann liegt es nahe, diese relativen Häufigkeiten als A-priori-Wahrscheinlichkeiten zu verwenden. Beispielsweise haben 57,4% der Befragten angegeben, sie würden bei der Wahl die CDU/CSU wählen, während 42,6% ihre Stimme der SPD geben würden. (Diese Zahlen beziehen sich auf die in diesem Beispiel berücksichtigen Personen, unter denen sich ausschließlich CDU/CSU- und SPD-Wähler befinden.) Greift man nun eine beliebige Person aus der Datendatei heraus, so gehört diese mit einer Wahrscheinlichkeit von 42,6% der Gruppe der SPD-Wähler und mit einer Wahrscheinlichkeit von 57,4% den CDU/CSU-Wählern an.[272]
- **Bedingte Wahrscheinlichkeit P(D | G$_i$):** Die bedingte Wahrscheinlichkeit ist die Wahrscheinlichkeit, mit der sich ein bestimmter Funktionswert *D* ergibt, wenn der jeweilige Fall der Gruppe *G$_i$* entstammt. Die Berechnung dieser Wahrscheinlichkeiten erfolgt unter der Annahme, daß die Funktionswerte innerhalb jeder Gruppe normalverteilt sind und die Parameter der Verteilung geschätzt werden können.

Diese Wahrscheinlichkeiten sowie die jeweiligen Funktionswerte der Diskriminanzfunktion werden in der Tabelle aus Abbildung 25.2 für die ersten 30 in die Analyse aufgenommenen Fälle aus der Datei *allbus.sav* aufgelistet. Eine solche Liste wurde durch die Option *Ergebnisse für jeden Fall* aus dem Dialogfeld der

[272] Wenn Sie die Diskriminanzanalyse über die Befehlssyntax und nicht mit Hilfe der Dialogfelder durchführen, besteht alternativ die Möglichkeit, die A-priori-Wahrscheinlichkeiten für die einzelnen Gruppen explizit vorzugeben. Dies ist sinnvoll, wenn aus anderen Quellen Informationen über die relativen Häufigkeiten der einzelnen Gruppen in der Grundgesamtheit vorliegen.

Schaltfläche *Klassifizieren* angefordert. In diesem Dialogfeld wurde auch festgelegt, daß für die A-priori-Wahrscheinlichkeiten abweichend von der Voreinstellung angenommen wird, sie entsprächen der in der Stichprobe beobachteten relativen Häufigkeiten.

Fallweise Statistiken

			Höchste Gruppe					Zweithöchste Gruppe			Diskriminanzwerte
	Fall-nummer	Tatsächliche Gruppe	Vorhergesagte Gruppe	P(D>d | G=g) p	P(D>d | G=g) df	P(G=g | D=d)	Quadrierter Mahalanobis-Abstand zum Zentroid	Gruppe	P(G=g | D=d)	Quadrierter Mahalanobis-Abstand zum Zentroid	Funktion 1
Original	1	1	2**	,719	1	,712	,129	1	,288	2,678	-1,114
	2	1	1	,822	1	,710	,050	2	,290	1,108	,298
	3	2	1**	,545	1	,601	,366	2	,399	,451	-,082
	4	1	1	1,000	1	,765	,000	2	,235	1,628	,522
	10	1	1	,843	1	,717	,039	2	,283	1,164	,325
	17	1	1	,932	1	,784	,007	2	,216	1,855	,608
	23	1	2**	,959	1	,594	,003	1	,406	1,501	-,703
	25	1	1	,637	1	,856	,222	2	,144	3,057	,994
	35	1	1	,606	1	,628	,266	2	,372	,579	,007
	36	2	1**	,966	1	,775	,002	2	,225	1,741	,565
	44	1	2**	,835	1	,545	,043	1	,455	1,142	-,546
	45	2	2	,826	1	,675	,048	1	,325	2,241	-,974
	46	2	2	,184	1	,895	1,766	1	,105	6,791	-2,083
	48	1	1	,882	1	,798	,022	2	,202	2,031	,671
	49	2	2	,700	1	,719	,149	1	,281	2,763	-1,140
	50	1	1	,369	1	,911	,808	2	,089	4,734	1,421
	59	2	2	,650	1	,737	,206	1	,263	2,996	-1,208
	64	2	2	,352	1	,837	,865	1	,163	4,872	-1,685
	66	2	2	,966	1	,597	,002	1	,403	1,525	-,712
	73	2	2	,887	1	,566	,020	1	,434	1,287	-,612
	76	2	2	,940	1	,633	,006	1	,367	1,830	-,830
	77	1	1	,064	1	,972	3,434	2	,028	9,798	2,376
	82	2	2	,210	1	,886	1,572	1	,114	6,406	-2,008
	86	1	1	,434	1	,899	,612	2	,101	4,240	1,305
	95	1	1	,820	1	,709	,052	2	,291	1,102	,296
	96	1	1	,891	1	,795	,019	2	,205	1,999	,659
	97	1	1	,409	1	,903	,682	2	,097	4,421	1,348
	98	1	1	,285	1	,927	1,144	2	,073	5,506	1,592
	102	1	2**	,769	1	,518	,086	1	,482	,967	-,461
	103	2	1**	,865	1	,724	,029	2	,276	1,225	,353

**. Falsch klassifizierter Fall

Abbildung 25.2: Klassifizierungsstatistiken der Diskriminanzanalyse

Die letzte Spalte der Tabelle (*Diskriminanzwerte*) gibt für die einzelnen Fälle jeweils den Funktionswert der Diskriminanzanalyse wieder. Für Fall 1 wird zum Beispiel der oben manuell berechnete Wert von -1,114 ausgewiesen. Bei einem solchen Funktionswert entstammt der Fall mit der größten Wahrscheinlichkeit aus der Gruppe 2 (*Vorhergesagte Gruppe*), es handelt sich also wahrscheinlich um einen SPD-Wähler. Die geschätzte Wahrscheinlichkeit dafür, daß der Fall zur zweiten Gruppe gehört, beträgt 71,2% (Spalte *P(G=g | D=d)* unter der Überschrift *Höchste Gruppe*). Tatsächlich würde die zu diesem Fall gehörende Person allerdings die CDU wählen und gehört damit der Gruppe 1 an (*Tatsächliche Gruppe*). In diesem Fall wurde also anhand der geschätzten Wahrscheinlichkeiten eine falsche Zuordnung vorgenommen. Neben der A-posteriori-Wahrscheinlichkeit wird in der Spalte *p* unter der Überschrift *Höchste Gruppe* auch die bedingte Wahrscheinlichkeit dieses Falles für die Zugehörigkeit zur Gruppe 2 mitgeteilt. Unter der Voraussetzung, daß der Fall der zweiten Gruppe entstammt, ergibt sich ein Funktionswert der Größe -1,114 mit einer Wahrscheinlichkeit von 71,9%.

Zusätzlich enthält die Tabelle auch Angaben für die Gruppe, der ein Fall mit der zweitgrößten Wahrscheinlichkeit angehört. In diesem Beispiel ist dies trivial, da ohnehin nur zwei Gruppen unterschieden werden. Dementsprechend stellt auch die Angabe der Wahrscheinlichkeit, mit der ein Fall aus der zweitwahrscheinlichsten Gruppe stammt, eine Redundanz dar, denn jeder Fall ist genau einer Gruppe zuzuordnen, und die einzelnen Wahrscheinlichkeiten, mit denen die Fälle den verschiedenen Gruppen angehören, summieren sich stets zu 1. Wenn lediglich zwei Gruppen in Frage kommen, ergibt sich die geringere A-posteriori-Wahrscheinlichkeit damit unmittelbar aus der größeren und umgekehrt. Die Betrachtung der zweitgrößten A-posteriori-Wahrscheinlichkeit kann Aufschluß über Unsicherheiten bei der Zuordnung der Fälle zu den Gruppen geben. Liegen die beiden größten Wahrscheinlichkeiten sehr nahe beieinander, wie zum Beispiel bei dem Fall mit der Nummer 44, ist die Gruppenzuordnung besonders unsicher. Entsprechend wurde auch Fall 44 der falschen Gruppe zugeordnet.

Durch einen Vergleich der Gruppen, denen die einzelnen Fälle aufgrund der A-posteriori-Wahrscheinlichkeiten zugeordnet werden (*Höchste Gruppe*), mit den Gruppen, aus denen sie tatsächlich stammen (*Tatsächliche Gruppe*), läßt sich ein Eindruck von der Güte der Diskriminanzfunktion gewinnen. Im Idealfall würden alle Gruppenzuordnungen mit den tatsächlichen Gruppen übereinstimmen, Abweichungen zwischen den tatsächlichen und den zugeordneten Gruppen zeigen dagegen tendenziell Modellfehler an. In der dargestellten Tabelle werden falsch zugeordnete Fälle durch zwei Sternchen gekennzeichnet. Auf diese Weise ist sehr schnell zu erkennen, daß von den ersten 30 Fällen insgesamt 7 falsch zugeordnet wurden. Dies ist zwar kein sehr geringer Fehleranteil, jedoch vor dem Hintergrund, daß sich das Wahlverhalten im allgemeinen nicht einfach durch wenige Merkmale der Personen erklären läßt, durchaus befriedigend.[273]

Fehler der Diskriminanzanalyse, die in falschen Gruppenzuordnungen zum Ausdruck kommen, müssen nicht in falschen Parametern der Funktion begründet sein. Vielmehr ist anzunehmen, daß das Erklärungsmodell fehlerhaft ist. In den wenigsten Fällen, insbesondere bei wirtschafts- und sozialwissenschaftlichen Untersuchungen, wird es möglich sein, eine abhängige Variable perfekt durch eine oder mehrere unabhängige Variablen zu beschreiben, da häufig eine sehr große Anzahl an oftmals nicht erfaßbaren oder quantifizierbaren Faktoren Einfluß auf die abhängige Variable ausüben. Sind alle Fälle durch die Analyse den einzelnen Gruppen richtig zugeordnet worden, besagt dies lediglich, daß das Modell und die Funktion gut geeignet sind, Zusammenhänge zwischen den unabhängigen und der abhängigen Variablen in der Stichprobe aufzuzeigen. Daraus folgt jedoch noch nicht automatisch, daß die Analyse auch für Prognosezwecke geeignet ist. Dies ist

[273] In diesem Beispiel wurden nur Personen betrachtet, die entweder die CDU/CSU oder die SPD wählen würden. Anhänger anderer Parteien wurden aus der Analyse ausgeschlossen. Wären diese Fälle nicht schon in der Datendatei deaktiviert worden, während bei der Diskriminanzanalyse dennoch lediglich zwischen CDU/CSU und SPD unterschieden wird, würden in der Tabelle auch die Personen aufgeführt, die eine andere Partei als CDU/CSU oder SPD gewählt hätten. In der Spalte *Tatsächliche Gruppe* würden die entsprechenden Fälle den Eintrag *Ungruppiert* aufweisen, durch die Diskriminanzanalyse würde aber dennoch eine Gruppenzuordnung vorgenommen werden.

lediglich dann der Fall, wenn die Stichprobe die Grundgesamtheit sehr gut repräsentiert und die zu prognostizierenden Fälle ebenfalls dieser Grundgesamtheit entstammen. Bestehen jedoch Unterschiede zwischen der Datenstruktur der Grundgesamtheit und der Struktur der Stichprobe, ergeben sich systematische Verzerrungen bei den Prognosen.

25.2 Ergebnisse der Diskriminanzanalyse

Der Standard-Output der Diskriminanzanalyse enthält die wesentlichen Informationen über das aufgestellte diskriminanzanalytische Modell. Zusätzlich werden einige Maßzahlen ausgewiesen, die der Überprüfung der Modellgüte dienen. Ergänzend zu diesem Standard-Output können Sie zahlreiche Zusatzinformationen für die Beschreibung des Modells und der Ergebnisse anfordern.

25.2.1 Vergleiche der Gruppenmittelwerte

Neben einem Vergleich der von der Diskriminanzanalyse vorgenommenen Gruppenzuordnungen mit den tatsächlichen Gruppen gibt es noch einige weitere Möglichkeiten, das Modell zur Erklärung der abhängigen Variablen auf seine Güte hin zu überprüfen. Einige dafür geeignete Angaben sind automatisch im Standard-Output enthalten, andere können zusätzlich durch entsprechende Optionen angefordert werden.

Einer dieser Methoden zur Überprüfung der Analyse liegt der Gedanke zugrunde, daß sich eine gute Diskriminanzfunktion unter anderem dadurch auszeichnet, daß sich die durchschnittlichen Funktionswerte der einzelnen Gruppen (also die Mittelwerte der Diskriminanzfunktion in den einzelnen Gruppen) deutlich voneinander unterscheiden. Diese Überlegung bildet die Grundlage für den Output aus Abbildung 25.3, der Maßzahlen zur Überprüfung der Modellgüte enthält.

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	,398[a]	100,0	100,0	,533

a. Die ersten 1 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

Wilks' Lambda

Test der Funktion(en)	Wilks-Lambda	Chi-Quadrat	df	Signifikanz
1	,716	79,270	5	,000

Abbildung 25.3: Ergebnisse der Diskriminanzfunktion zur Bewertung der Modellgüte

Eigenwert

Die ausgewiesenen Maßzahlen ähneln denen einer Varianzanalyse, und auch das Konzept dieser Untersuchung ist dem der Varianzanalyse sehr ähnlich. Als Testgröße wird der Eigenwert betrachtet. Dieser entspricht nicht nur in seiner Funktion dem F-Wert einer Varianzanalyse, sondern wird auch auf sehr ähnliche Weise berechnet. Der Eigenwert ergibt sich aus dem Quotienten der Quadratsumme zwischen den Gruppen (QSZ) und der Quadratsumme innerhalb der Gruppen (QSI). Der Unterschied zur Berechnung des F-Wertes besteht lediglich darin, daß beim F-Wert die jeweiligen Freiheitsgrade berücksichtigt werden.

$$\text{Eigenwert} = \frac{\text{QSZ}}{\text{QSI}}$$

Ein großer Eigenwert ergibt sich, wenn die Streuung zwischen den Gruppen im Verhältnis zur Streuung innerhalb der Gruppen sehr groß ist. Dies ist die von einer Diskriminanzanalyse angestrebte Situation. Wird dies erreicht, ist gewährleistet, daß sich die Funktionswerte der einzelnen Gruppen deutlich voneinander unterscheiden, während die Werte innerhalb einer Gruppe sehr ähnlich sind. Dadurch ist es anschließend leicht möglich, anhand eines Funktionswertes auf die Gruppenzugehörigkeit zu schließen. In Abbildung 25.3 wird ein Eigenwert von 0,398 ausgewiesen. Dieser Wert zeigt an, daß die Streuung zwischen den Gruppen nur das 0,4fache der Streuung innerhalb der Gruppen beträgt. Die Tatsache, daß überhaupt eine Streuung zwischen den Gruppen vorliegt, deutet darauf hin, daß das zugrundeliegende Modell durchaus einen gewissen Erklärungswert besitzt, der relativ geringe Eigenwert läßt jedoch vermuten, daß es noch verbesserungsfähig ist.

Die beiden Spalten *% der Varianz* und *Kumulierte %* in der Tabelle *Eigenwerte* haben im vorliegenden Fall keinen Aussagegehalt, da die Werte bei einer Unterscheidung von nur zwei Gruppen stets 100% betragen.

Kanonischer Korrelationskoeffizient

Der kanonische Korrelationskoeffizient mißt die Strenge des Zusammenhangs zwischen den Funktionswerten der Diskriminanzfunktion und den Gruppen der abhängigen Variablen. Dieser Koeffizient ergibt sich als Quadratwurzel des Quotienten aus der Quadratsumme zwischen den Gruppen und der gesamten Quadratsumme (QSZ + QSI). Damit ist er genauso definiert wie der Wert *eta*, der bei einer Varianzanalyse betrachtet wird. Für den Fall von lediglich zwei Gruppen ist dieser Wert mit dem Pearson'schen Korrelationskoeffizienten identisch.

$$\text{Kanonischer Korrelationskoeffizient} = \sqrt{\frac{\text{QSZ}}{\text{QSZ} + \text{QSI}}}$$

Da der Eigenwert als Quotient aus QSZ und QSI definiert ist und zudem die Summe dieser beiden Quadratsummen die gesamte Quadratsumme ergibt, läßt

sich der kanonische Korrelationskoeffizient auch mit Hilfe des Eigenwertes berechnen:

$$\text{Kanonischer Korrelationskoeffizient} = \sqrt{\frac{\text{Eigenwert}}{1+\text{Eigenwert}}}$$

In diesem Fall beträgt der Eigenwert 0,398, so daß sich folgender kanonischer Korrelationskoeffizient ergibt:

$$\text{Kanonischer Korrelationskoeffizient} = \sqrt{\frac{0{,}398}{1{,}398}} = 0{,}53$$

Dieser Wert wird auch in der Tabelle *Eigenwerte* in der Spalte *Kanonische Korrelation* ausgewiesen.

Die Formel zur Berechnung des kanonischen Korrelationskoeffizienten zeigt deutlich, daß dieser ebenso wie der Eigenwert durch die Verteilung der gesamten Streuung auf die Streuungen innerhalb und die Streuung zwischen den Gruppen bestimmt ist. Der kanonische Korrelationskoeffizient mißt den Anteil der Streuung zwischen den Gruppen an der gesamten Streuung. Die Werte des Koeffizienten liegen zwischen 0 und 1. Je größer der Wert ist, desto größer ist die Streuung zwischen den Gruppen im Verhältnis zur Streuung innerhalb der Gruppen, so daß ein großer kanonischer Korrelationskoeffizient auf eine gute Trennung zwischen den Gruppen und damit auf einen hohen Erklärungsgehalt des Modells hinweist.

Wilks' Lambda

Wilks' Lambda ist der Quotient aus der Quadratsumme innerhalb der Gruppen und der gesamten Quadratsumme:

$$\text{Wilks' Lambda} = \frac{\text{QSI}}{\text{QSZ} + \text{QSI}}$$

Ein Vergleich mit der Formel zur Berechnung des kanonischen Korrelationskoeffizienten zeigt, daß Wilks' Lambda und der Korrelationskoeffizient in der Summe stets 1 ergeben. Zwischen den beiden Größen besteht damit der Zusammenhang

$$\text{Wilks' Lambda} = 1 - \text{Kanonischer Korrelationskoeffizient}$$

Damit ist die Angabe beider Werte in gewisser Weise redundant, und man kann sich bei der Interpretation der Testergebnisse je nach Vorliebe auf einen der beiden Werte konzentrieren. Im Zusammenhang mit der Diskriminanzanalyse stellt Wilks' Lambda den häufiger betrachteten Wert dar.

Hypothesentest mit Chi-Quadrat

Wilks' Lambda läßt sich in ein annähernd Chi-Quadrat-verteiltes Maß transformieren. Kleine Lambda-Werte werden dabei in große Werte des Chi-Quadrat-

Maßes überführt. Durch diese Transformation läßt sich ein Hypothesentest durchführen. Er testet die Hypothese, die Gruppenmittelwerte der Funktionswerte aus der Diskriminanzfunktion seien in der Grundgesamtheit identisch. Der Chi-Quadrat-Wert wird in Abbildung 25.3 in der Tabelle Wilks' Lambda mit 79,270 angegeben. Bei den vorliegenden fünf Freiheitsgraden ergibt sich damit ein Signifikanzwert von 0,000.[274] Damit kann die Hypothese, derzufolge in der Grundgesamtheit kein Unterschied zwischen den gruppenweisen Funktionsmittelwerten besteht, zurückgewiesen werden. Man kann also davon ausgehen, daß zumindest nicht alle Gruppenmittelwerte der Funktion identisch sind. Dies besagt jedoch noch nicht, daß das gesamte Modell der Diskriminanzanalyse gut geeignet ist, um von den Funktionswerten auf die Gruppen zu schließen. Auch ein signifikanter Unterschied der Gruppenmittelwerte kann sehr klein sein, so daß es schwierig ist, auf der Basis der Funktionswerte die richtige Gruppenzuordnung vorzunehmen. Der hohe Signifikanzwert kann lediglich so interpretiert werden, daß das Modell nicht vollkommen ungeeignet zur Erklärung der abhängigen Variablen ist.

25.2.2 Standardisierte Koeffizienten

Ebenfalls Bestandteil des Standard-Output ist eine Tabelle mit den standardisierten Koeffizienten. Diese wird in Abbildung 25.4 wiedergegeben.

Standardisierte kanonische Diskriminanzfunktionskoeffizienten

	Funktion
	1
ERHEBUNGSGEBIET: WEST - OST	-,012
ALTER: BEFRAGTE<R>	-,306
LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	1,027
ALLGEMEINER SCHULABSCHLUSS	,051
BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	,114

Abbildung 25.4: Standardisierte Koeffizienten der Diskriminanzfunktion

Die standardisierten Koeffizienten ergeben sich, indem die Ausgangswerte der erklärenden Variablen vor der Berechnung der Diskriminanzfunktion standardisiert werden. Dabei werden die Werte jeder Variablen so transformiert, daß sie anschließend einen Mittelwert von 0 sowie eine Standardabweichung von 1 aufweisen. Der Vorteil der Betrachtung standardisierter Werte besteht darin, daß Einflüsse unterschiedlicher Dimensionen in verschiedenen Variablen eliminiert werden.

[274] 0,000 ist nicht gleich null, denn es handelt sich um einen auf drei Dezimalstellen gerundeten Wert.

Die nicht standardisierten Koeffizienten wurden oben in Abbildung 25.1, S. 595 wiedergegeben. Demnach beträgt der Koeffizient für das Nettoeinkommen 0,000. Das Nettoeinkommen schien damit kaum in die Diskriminanzfunktion einzugehen, und es hätte nahegelegen, diese unabhängige Variable aus dem Modell zu entfernen. Der standardisierte Koeffizient des Nettoeinkommens beträgt nun 0,114 und ist damit nicht nur deutlich von null verschieden, sondern auch größer als der standardisierte Koeffizient des Schulabschlusses, der - ebenso wie in der nicht standardisierte Form - 0,05 beträgt. Der sehr kleine Wert des nicht standardisierten Einkommenskoeffizienten erklärt sich damit, daß die Einkommensvariable systematisch wesentlich größere Werte aufweist als beispielsweise die Variable *Schulabschluß*, die lediglich zwischen den Codierungen von *1* bis *5* unterscheidet. Geht ein Einkommenswert von *4.000* mit dem Koeffizienten 0,0003 in die Diskriminanzfunktion ein, hat dies somit den gleichen Effekt wie eine mit einem Koeffizienten von 3 berücksichtigte Schulabschluß-Codierung *4*.

Auch aus den Beträgen der standardisierten Koeffizienten kann man jedoch nicht ohne weiteres auf die Stärke des Zusammenhangs zwischen der betreffenden unabhängigen und der abhängigen Variablen schließen. Die Koeffizienten können durch Wechselwirkungen zwischen den unabhängigen Variablen verzerrt werden. Sind die unabhängigen Variablen untereinander korreliert, wird der Einfluß einer unabhängigen Variablen auf die abhängige Variablen in dem Modell möglicherweise einer anderen unabhängigen Variablen zugeschrieben.

25.2.3 Korrelationen zwischen unabhängigen Variablen und Diskriminanzfunktionen

Der Erklärungswert, den die einzelnen Variablen für die Diskriminanzfunktion besitzen, läßt sich ebenfalls durch Korrelationen ausdrücken. Hierzu werden Korrelationskoeffizienten zwischen den einzelnen unabhängigen Variablen und der Diskriminanzfunktion berechnet.[275] Die Korrelationen zwischen den unabhängigen Variablen und der Diskriminanzfunktion werden im Standard-Output mitgeteilt und sind in Abbildung 25.5 wiedergegeben.

Die Koeffizienten zeigen für die fünf erklärenden Variablen die Korrelationen zwischen den Variablenwerten und den Funktionswerten der Diskriminanzfunktion an. Üblicherweise werden bei der Berechnung von Korrelationskoeffizienten alle Fälle aus der Datendatei gleichzeitig und gleichberechtigt einbezogen. Die Berechnung dieser Koeffizienten weicht jedoch von der üblichen Vorgehensweise ab. Hier werden zunächst die Korrelationskoeffizienten der einzelnen Gruppen berechnet. In diesem Fall werden also für jede Variable zunächst zwei Koeffizienten für die Korrelation zwischen den Variablen- und den Funktionswerten berechnet, jeweils einer für die Fälle der Gruppe 1 (CDU/CSU-Wähler) und einer

[275] Werden in der abhängigen Variablen mehr als zwei Gruppen unterschieden, liefert die Diskriminanzanalyse mehrere Diskriminanzfunktionen. In dem Fall werden Korrelationskoeffizienten zwischen den einzelnen unabhängigen Variablen und den verschiedenen Diskriminanzfunktionen berechnet.

entsprechend für Gruppe 2 (SPD-Wähler). Der ausgewiesene Koeffizient für die gesamte Variable ergibt sich anschließend als Mittelwert der einzelnen Koeffizienten (gepoolte Koeffizienten)

Struktur-Matrix

	Funktion
	1
LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	,940
ERHEBUNGSGEBIET: WEST - OST	-,170
ALLGEMEINER SCHULABSCHLUSS	,110
BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	,078
ALTER: BEFRAGTE<R>	-,057

Gemeinsame Korrelationen innerhalb der Gruppen zwischen Diskriminanzvariablen und standardisierten kanonischen Diskriminanzfunktionen
Variablen sind nach ihrer absoluten Korrelationsgröße innerhalb der Funktion geordnet.

Abbildung 25.5: Korrelationskoeffizienten der unabhängigen Variablen mit der Diskriminanzfunktion

Die Werte dieser gepoolten Korrelationskoeffizienten können sich sehr stark von den Koeffizienten bei einfacher Berechnungsweise unterscheiden. Ergibt sich zum Beispiel für die beiden einzelnen Gruppen jeweils ein Koeffizient von null, so daß auch der ausgewiesene gepoolte Koeffizient den Wert null hat, ist es durchaus möglich, daß bei der gleichzeitigen Betrachtung aller Fälle eine Korrelation beobachtet werden kann, da die Variablen- und Funktionswerte zwischen den Gruppen in dieselbe oder in unterschiedliche Richtungen streuen, so daß entsprechend eine positive oder negative Korrelation berechnet wird.

Der höchste Korrelationskoeffizient wird für die Variable *Links-Rechts-Selbsteinstufung* ausgewiesen. Dieser deutet mit 0,94 auf eine fast perfekte Korrelation zwischen den Funktionswerten und den Variablenwerten hin. Wesentlich niedriger fallen dagegen die Koeffizienten der anderen unabhängigen Variablen aus. Die beiden Variablen *Erhebungsgebiet* und *Alter* haben einen negativen Korrelationskoeffizienten. Ein höheres Alter wirkt damit hinsichtlich der Funktionswerte in die gleiche Richtung wie ein niedrigerer Schulabschluß oder wie ein niedrigerer Wert bei der Links-Rechts-Selbsteinstufung (diese wird auf einer Skala von 1 (*Links*) bis 10 (*Rechts*) gemessen). Für Rückschlüsse von den Beträgen der Koeffizienten auf den Erklärungsgehalt der Variablen gelten allerdings die gleichen Einschränkungen wie für die standardisieren Koeffizienten. Sobald Korrelationen zwischen den Variablen vorliegen, sind solche Rückschlüsse nicht mehr uneingeschränkt möglich, denn die Koeffizienten können verzerrt sein.

25.2.4 Funktionsmittelwerte in den einzelnen Gruppen

Ebenfalls als Bestandteil des Standard-Output der Diskriminanzanalyse werden die durchschnittlichen Funktionswerte für die einzelnen Gruppen ausgegeben. Diese Angaben sind in Abbildung 25.6 dargestellt.

Funktionen bei den Gruppen-Zentroiden

WAHLABSICHT, BUNDESTAGSWAHL; ...	Funktion 1
CDU-CSU	,523
SPD	-,754

Nicht-standardisierte kanonische Diskriminanzfunktionen, die bezüglich des Gruppen-Mittelwertes bewertet werden

Abbildung 25.6: Funktionsmittelwerte in den beiden Gruppen der abhängigen Variablen

Für die erste Gruppe (CDU/CSU) ergab sich ein durchschnittlicher Funktionswert von 0,523. Der entsprechende Mittelwert der zweiten Gruppe (SPD) beträgt -0,754. Diese Mittelwerte werden auch Gruppenzentroide genannt. Je näher diese beieinanderliegen, desto schwieriger ist es, einen Fall anhand seines Funktionswertes einer der Gruppen zuzuordnen. Entscheidend für die Bewertung der Analyseergebnisse ist auch die Frage, ob die Mittelwerte in der Grundgesamtheit signifikant voneinander verschieden sind oder ob möglicherweise gar kein Unterschied zwischen ihnen besteht. Ist letzteres der Fall, ist der betrachtete Modellansatz nicht geeignet, die Werte der abhängigen Variablen vorherzusagen. In diesem Beispiel sind die Gruppenmittelwerte allerdings signifikant verschieden voneinander. Dies wurde oben im Abschnitt *Hypothesentest mit Chi-Quadrat*, S. 602 festgestellt.

25.2.5 Tabelle der Treffsicherheit

Anhand eines Vergleichs der durch die Diskriminanzanalyse vorgenommenen Gruppenzuordnungen mit den tatsächlichen Gruppenzugehörigkeiten (siehe S. 598) ließ sich bereits ein Eindruck von der Güte der Diskriminanzanalyse gewinnen. Angaben über die Treffsicherheit der Zuordnungen können jedoch auch in kompakter Form angefordert werden. Hierzu können Sie sich eine Tabelle mit der Anzahl der richtig und falsch zugeordneten Fälle ausgeben lassen. Diese wurde in den Einstellungen der Diskriminanzanalyse (siehe S. 593) mit der Option *Zusammenfassende Tabelle* aus dem Dialogfeld der Schaltfläche *Klassifizieren* angefordert und ist in Abbildung 25.7 wiedergegeben.

Klassifizierungsergebnisse[a]

		WAHLABSICHT, BUNDESTAGSWAHL; BEFR.	Vorhergesagte Gruppenzugehörigkeit		
			CDU-CSU	SPD	Gesamt
Original	Anzahl	CDU-CSU	120	23	143
		SPD	36	63	99
	%	CDU-CSU	84,1	15,9	100,0
		SPD	36,1	63,9	100,0

a. 75,9% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

Abbildung 25.7: Übersicht über die richtigen und falschen Gruppenzuordnungen

Der obere Tabellenbereich mit der Bereichsüberschrift *Anzahl* enthält Angaben in absoluten Werten, der untere Bereich mit der Beschriftung *%* gibt die entsprechenden Informationen in relativen Werten wieder. Im oberen Tabellenbereich ist folgendes zu erkennen: Die letzte Spalte gibt an, daß sich 143 der betrachteten Personen als CDU/CSU-Wähler ausgegeben haben, während 99 Personen SPD-Wähler sind. Von den 143 CDU/CSU-Wählern wurden 120 von der Diskriminanzanalyse auch tatsächlich in die CDU/CSU-Gruppe eingeordnet. 23 CDU/CSU-Wähler wurden dagegen fälschlicherweise als SPD-Wähler klassifiziert. Bei der Gruppe der SPD-Wähler (die tatsächlich SPD wählen) ist der Fehler noch größer. Von den insgesamt 99 SPD-Wählern wurden mit 63 ungefähr 2/3 richtig als SPD-Wähler erkannt, während 36 Personen fälschlich in die Gruppe der CDU/CSU-Wähler eingeordnet wurden.

Auf die gleiche Weise ist der untere Tabellenbereich mit den Prozentangaben zu lesen: 84,1% der CDU/CSU-Wähler wurden von der Diskriminanzanalyse korrekt als solche erkannt, während 15,9% fehlerhaft der SPD zugeordnet wurden. Von den SPD-Wählern wurden dagegen nur 63,9% in die richtige Gruppe eingeordnet, während 36,1% dem falschen Lager zugeschrieben wurden.

Insgesamt wurden 23 + 36 = 59 der insgesamt 143 + 99 = 242 Personen falsch zugeordnet. Die Fehlerquote ist also mit ungefähr 24% recht hoch. Bereits oben wurde jedoch darauf hingewiesen, daß ein so einfaches Modell wie dieses niemals geeignet sein kann, ein derart komplexes Phänomen wie das Wahlverhalten auch nur annähernd perfekt zu prognostizieren.

25.2.6 Gruppenmittelwerte der Variablen

Durch die Betrachtung der Mittelwerte der einzelnen unabhängigen Variablen in den verschiedenen Gruppen der abhängigen Variablen läßt sich ein Eindruck von dem Erklärungswert der unabhängigen Variablen gewinnen.

Angabe der Mittelwerte

Die Angabe der Gruppenmittelwerte wurde in den Dialogfeldeinstellungen der Diskriminanzanalyse mit der Option *Mittelwert* aus dem Dialogfeld der Schaltfläche *Statistiken* angefordert (siehe S. 593). Abbildung 25.8 zeigt die Tabelle mit den entsprechenden Ergebnissen.

Gruppenstatistik

WAHLABSICHT, BUNDESTAGSWAHL; BEFR.		Mittelwert	Standardabweichung	Gültige Werte (listenweise)	
				Ungewichtet	Gewichtet
CDU-CSU	ERHEBUNGSGEBIET: WEST - OST	1,17	,38	140	142,563
	ALTER: BEFRAGTE<R>	47,09	17,92	140	142,563
	LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	6,52	1,84	140	142,563
	ALLGEMEINER SCHULABSCHLUSS	2,79	1,12	140	142,563
	BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	2451,88	1848,94	140	142,563
SPD	ERHEBUNGSGEBIET: WEST - OST	1,26	,44	104	98,768
	ALTER: BEFRAGTE<R>	48,40	17,54	104	98,768
	LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	4,41	1,62	104	98,768
	ALLGEMEINER SCHULABSCHLUSS	2,64	,85	104	98,768
	BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	2280,46	1525,69	104	98,768
Gesamt	ERHEBUNGSGEBIET: WEST - OST	1,21	,40	244	241,331
	ALTER: BEFRAGTE<R>	47,63	17,74	244	241,331
	LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	5,66	2,03	244	241,331
	ALLGEMEINER SCHULABSCHLUSS	2,73	1,02	244	241,331
	BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	2381,72	1722,71	244	241,331

Abbildung 25.8: Tabelle mit Gruppenmittelwerten der einzelnen unabhängigen Variablen

Der Spalte *Mittelwert* gibt die Mittelwerte der unabhängigen Variablen für die beiden Gruppen *CDU/CSU* und *SPD* sowie im untersten Tabellenbereich für die Gesamtheit der betrachteten Fälle wieder. Ein Vergleich der Mittelwerte in den beiden Gruppen läßt darauf schließen, daß die unabhängigen Variablen durchaus zur Erklärung der abhängigen Variablen geeignet sind. Am deutlichsten wird dies wiederum bei der Links-Rechts-Selbsteinstufung. Diese hat in der Gruppe der CDU/CSU-Wähler einen durchschnittlichen Skalenwert von 6,52 und in der Gruppe der SPD-Wähler den wesentlich niedrigeren Wert von 4,41. Deutlich geringer sind dagegen die Unterschiede bei den anderen Variablen. Die CDU/CSU-Wähler weisen in der zugrundeliegenden Stichprobe im Durchschnitt ein höheres Einkommen, ein geringeres Alter und einen höheren Schulabschluß auf[276]. Zudem

[276] Die Variable *Schulabschluß* hat lediglich Ordinalskalenniveau, so daß eine Mittelwertbetrachtung eigentlich nicht zulässig ist.

kommen sie zu einem größeren Anteil aus den alten Bundesländern. Je näher die Mittelwerte der unabhängigen Variablen in den beiden Gruppen beieinanderliegen, desto schwieriger wird es sein, aus den entsprechenden Variablenwerten Rückschlüsse auf die Gruppenzugehörigkeit zu ziehen. Allerdings sollte auch hier nicht der Fehler gemacht werden, aus der Höhe der Differenz zwischen den beiden Gruppenmittelwerten unmittelbar Rückschlüsse auf den Erklärungsgehalt der betreffenden Variablen zu ziehen, da wiederum die Dimension zu berücksichtigen ist, in der eine Variable gemessen wird. So ist es nicht weiter verwunderlich, daß sich der größte Mittelwertunterschied beim Einkommen ergibt, da diese Variable mit Abstand die größten Werte aufweist.

Neben dem Mittelwert wird die Standardabweichung als Maß für die Streuung der Werte in den einzelnen Gruppen ausgewiesen. Auch diese Größe ist für den Erklärungswert der Variablen von Relevanz. Im Idealfall bestehen große Unterschiede zwischen den Gruppenmittelwerten und nur sehr geringe Streuungen innerhalb der einzelnen Gruppen. Die Standardabweichung wird ebenso wie die Mittelwertdifferenz durch die Dimension der jeweiligen Variablen beeinflußt. Die größere Streuung der Einkommensvariablen innerhalb der beiden Gruppen läßt somit nicht den Schluß zu, diese Variable sei schlechter zur Erklärung des Wahlverhaltens geeignet als beispielsweise das Alter.

Mit Ausnahme der Links-Rechts-Selbsteinstufung weisen die unabhängigen Variablen nur geringe Mittelwertunterschiede auf, während gleichzeitig die Streuung in einzelnen Variablen wie etwa beim Einkommen oder auch beim Alter vergleichsweise hoch ist. Am schlechtesten scheint insgesamt das Einkommen zur Erklärung der Gruppenzugehörigkeit geeignet zu sein. Die Mittelwertdifferenz ist mit unter 200 bei einem durchschnittlichen Einkommen von ungefähr 2.300 recht gering, während gleichzeitig die Streuung in jeder der beiden Gruppen einen Umfang von 75% bzw. 65% des Mittelwertes hat.

Gleichheitstest der Gruppenmittelwerte

In Abbildung 25.8 wurde die absolute Größe der Mittelwertdifferenzen deutlich. Selbst wenn die unabhängigen Variablen keinen Erklärungswert für die abhängige Variable besitzen, wird es jedoch nur in seltenen Ausnahmen vorkommen, daß die Mittelwerte der unabhängigen Variablen in den verschiedenen Gruppen der abhängigen Variablen vollkommen identisch sind. Vielmehr werden sich in aller Regel schon durch zufällige Einflüsse Unterschiede zwischen den Mittelwerten ergeben. Um beurteilen zu können, ob Differenzen zwischen den Mittelwerten auf einen Zufall bei der Ziehung der Stichprobe oder vielleicht doch auf einen systematischen Unterschied der Werte in der Grundgesamtheit zurückzuführen sind, können Sie für die Gruppenmittelwerte der einzelnen Variablen Signifikanztests durchführen. Das Ergebnis dieser Tests ist in Abbildung 25.9 wiedergegeben. In den Dialogfeldeinstellungen (siehe S. 593) wurde er mit der Option *Univariate ANOVA* aus der Gruppe *Deskriptive Statistiken* angefordert.

Gleichheitstest der Gruppenmittelwerte

	Wilks-Lambda	F	df1	df2	Signifikanz
ERHEBUNGSGEBIET: WEST - OST	,989	2,736	1	239	,099
ALTER: BEFRAGTE<R>	,999	,313	1	239	,576
LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	,740	84,140	1	239	,000
ALLGEMEINER SCHULABSCHLUSS	,995	1,150	1	239	,285
BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	,998	,577	1	239	,448

Abbildung 25.9: Signifikanztests für die Mittelwertunterschiede zwischen den Gruppen

Lediglich die Links-Rechts-Selbsteinstufung weist hochsignifikante Mittelwertunterschiede auf.[277] Dies bestätigt noch einmal den Eindruck, daß diese Variable unter der ausgewählten unabhängigen Variablen wohl den größten Erklärungswert besitzt. Bei den übrigen Variablen kann nicht davon ausgegangen werden, sie würden in der Grundgesamtheit unterschiedliche Mittelwerte aufweisen. Bei der Interpretation des Signifikanztests ist zu beachten, daß das Erhebungsgebiet eine dichotome Variable ist und der Schulabschluß lediglich Ordinalskalenniveau besitzt. Für beide Skalen sind die Signifikanztests nicht konzipiert und auch nicht geeignet. Aus den Angaben für diese beiden Variablen sollten daher keine Rückschlüsse auf die Grundgesamtheit gezogen werden.

25.3 Diskriminanzanalyse mit mehr als zwei Gruppen

Unterscheidet man bei der abhängigen Variablen zwischen mehr als zwei Gruppen, ergeben sich gegenüber einer Analyse mit nur zwei Gruppen leichte Veränderungen bei der Berechnung der Gruppenzuordnungen sowie in dem Output der Analyse. Für die Darstellung der Gruppenzuordnungen in Abhängigkeit von den Funktionswerten stehen zudem zusätzliche Grafiken zur Verfügung.

25.3.1 Standard-Output

Beispiel

In dem bisherigen Beispiel wurde bei der Prognose des Wahlverhaltens lediglich zwischen der CDU/CSU und der SPD unterschieden. Im folgenden soll die Betrachtung auf die FDP sowie auf Bündnis90/Grüne ausgeweitet werden. Um eine Diskriminanzanalyse unter Einbeziehung dieser insgesamt vier Parteien durchzuführen, sind die folgenden Dialogfeldeinstellungen erforderlich:

[277] Zu den Werten *Wilks' Lambda* und *F* siehe auch Abschnitt *25.2.1, Vergleiche der Gruppenmittelwerte*, S. 600.

- **Daten:** Die Datengrundlage bildet wieder die Datei *allbus.sav* von der Begleit-CD. Die Daten in der Datei werden weiterhin mit der Variablen *v434* gewichtet.[278] Da im folgenden auch die Parteien FDP und Bündnis90/Grüne einbezogen werden sollen, ist auch die Auswahl der Fälle in der Datendatei zu ändern. Die nun verbleibenden Fälle sind dadurch gekennzeichnet, daß sie in der Variablen *v325* (Wahlabsicht bei der Bundestagswahl) einen der Werte *1*, *2*, *3* oder *6* aufweisen.[279]
- **Befehl:** Zum Aufrufen der Diskriminanzanalyse wählen Sie wieder den Befehl

STATISTIK
KLASSIFIZIEREN ▸
DISKRIMINANZANALYSE...

- **Gruppenvariable:** Fügen Sie die Variable *v325* in das Feld *Gruppenvariable* ein, und klicken Sie anschließend auf die Schaltfläche *Bereich definieren*. Geben Sie in dem damit geöffneten Dialogfeld in das Feld *Minimum* den Wert *1* und in das Feld *Maximum* den Wert *6* ein.
- **Unabhängige Variablen:** Verschieben Sie die folgenden Variablen in das Feld *Unabhängige Variable(n)*:
  - *v3* (Erhebungsgebiet, Ost-West)
  - *v37* (Alter)
  - *v112* (Links-Rechts-Selbsteinstufung)
  - *v142* (Schulabschluß)
  - *v261* (Einkommen)
- **Klassifizieren:** Wählen Sie in dem Dialogfeld *Klassifizieren* die Optionen *Kombinierte Gruppen* und *Territorien*. Ferner wird in der Gruppe *A-priori-Wahrscheinlichkeit* die Option *Alle Gruppen gleich* verwendet.

Abbildung 25.10 gibt verschiedene Elemente des Standard-Output der Diskriminanzanalyse wieder, die grundsätzlich bereits aus dem vorhergehenden Beispiel bekannt sind.

[278] Siehe hierzu Fn. 269, S. 593.

[279] Um alle anderen Fälle auszuschließen, verwenden Sie bei dem Befehl DATEN, FÄLLE AUSWÄHLEN die Option *v325 <= 6 & v325 > 0*. Zur genauen Vorgehensweise siehe Fn. 270, S. 593.

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	,396[a]	77,4	77,4	,532
2	,103[a]	20,2	97,6	,306
3	,012[a]	2,4	100,0	,110

a. Die ersten 3 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

Wilks' Lambda

Test der Funktion(en)	Wilks-Lambda	Chi-Quadrat	df	Signifikanz
1 bis 3	,642	142,824	15	,000
2 bis 3	,896	35,530	8	,000
3	,988	3,895	3	,273

Standardisierte kanonische Diskriminanzfunktionskoeffizienten

	Funktion		
	1	2	3
ERHEBUNGSGEBIET: WEST - OST	,043	,242	,063
ALTER: BEFRAGTE<R>	-,081	,541	,869
LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	1,015	-,141	-,045
ALLGEMEINER SCHULABSCHLUSS	-,058	-,696	,856
BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	,170	,197	-,419

(*wird fortgesetzt*)

Struktur-Matrix

	Funktion		
	1	2	3
LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	,987*	,007	,087
ALLGEMEINER SCHULABSCHLUSS	-,080	-,777*	,490
ALTER: BEFRAGTE<R>	,167	,724*	,617
ERHEBUNGSGEBIET: WEST - OST	-,082	,315*	,156
BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	,065	-,040	-,093*

Gemeinsame Korrelationen innerhalb der Gruppen zwischen Diskriminanzvariablen und standardisierten kanonischen Diskriminanzfunktionen
Variablen sind nach ihrer absoluten Korrelationsgröße innerhalb der Funktion geordnet.

*. Größte absolute Korrelation zwischen jeder Variablen und einer Diskriminanzfunktion

Funktionen bei den Gruppen-Zentroiden

WAHLABSICHT, BUNDESTAGSWAHL; BEFR.	Funktion		
	1	2	3
CDU-CSU	,685	,029	-,033
SPD	-,579	,373	,034
F.D.P.	,103	-,523	,381
BUENDNIS90-GRUENE	-,652	-,461	-,101

Nicht-standardisierte kanonische Diskriminanzfunktionen, die bezüglich des Gruppen-Mittelwertes bewertet werden

Abbildung 25.10: Teile des Standard-Output der Diskriminanzanalyse bei einer Unterscheidung zwischen den vier Parteien CDU/CSU, SPD, FDP und Bündnis90/Grüne

Der wesentliche Unterschied dieser Analyse zur vorhergehenden besteht darin, daß bei der Unterscheidung zwischen vier Gruppen drei Diskriminanzfunktionen berechnet werden. (Die Anzahl der Diskriminanzfunktionen ist immer um eins niedriger als die Anzahl der Gruppen in der abhängigen Variablen.) Daraus ergeben sich für den Output weitere Veränderungen, da bei allen sich auf die Diskriminanzfunktion beziehenden Angaben nun drei Funktionen zu berücksichtigen sind. Dabei beschränken sich die Änderungen nicht auf die Darstellung der Ergebnisse, sondern wirken sich auch auf die Inhalte der Angaben aus.

Zunächst werden die Eigenwerte sowie die weiteren damit im Zusammenhang stehenden Angaben nicht mehr für eine, sondern für drei Funktionen berechnet. Der Eigenwert der ersten Funktion beträgt 0,396, der für die zweite Funktion ist mit 0,103 deutlich geringer, und der Eigenwert der dritten Funktion ist mit 0,012 noch einmal wesentlich niedriger. Die Streuung der Funktionswerte zwischen den Gruppen ist damit im Verhältnis zur Streuung innerhalb der Gruppen bei den nach der ersten Diskriminanzfunktion berechneten Funktionswerten wesentlich größer als bei den Werten der zweiten und insbesondere denen der dritten Funktion.

Dieser Sachverhalt kommt auch in der Spalte *% der Varianz* zum Ausdruck. Den Angaben dieser Spalte liegt folgende Überlegung zugrunde: Für jede Diskriminanzfunktion kann die Streuung der Funktionswerte zwischen den Gruppen durch die *Quadratsumme zwischen den Gruppen* (QSZ) gemessen werden. Die Summe dieser Streuungen über alle drei Diskriminanzfunktionen wird als gesamte Varianz angesehen. In der Spalte *% der Varianz* wird nun angegeben, welcher Anteil der gesamten Streuung auf die einzelnen Funktionen entfällt. Die Angaben dieser Spalte müssen sich also zu 1 bzw. 100% addieren. Der auf die erste Funktion entfallende Anteil der Gesamtstreuung kann berechnet werden als:

$$\%\ \text{der Varianz}_{\text{Funktion1}} = \frac{\text{QSZ}_{\text{Funktion 1}}}{\text{QSZ}_{\text{Funktion 1}} + \text{QSZ}_{\text{Funktion 2}} + \text{QSZ}_{\text{Funktion 3}}} = 77{,}4\%$$

Entsprechend ergibt sich für die zweite Funktion ein Streuungsanteil von 20,2% und für die dritte Funktion ein Anteil von 2,4%. Damit leistet die erste Funktion offenbar einen sehr viel größeren Beitrag zur Unterscheidung zwischen den Gruppen, als es der zweiten und der dritten Funktion gelingt, bei denen sich die Funktionswerte in den unterschiedlichen Gruppen wesentlich stärker ähneln.

Auch der kanonische Korrelationskoeffizient zeigt an, daß die zweite und insbesondere die dritte Funktion einen wesentlich geringeren Erklärungsbeitrag leisten. Mit 0,306 bzw. 0,110 liegen die Koeffizienten dieser beiden Funktionen deutlich unter dem der ersten Funktion, der mit 0,532 schon einigermaßen befriedigend ist, aber bei weitem noch nicht als gut bezeichnet werden kann. Je besser der Erklärungsgehalt einer Funktion ist, desto stärker nähert sich der Korrelationskoeffizient dem Wert 1 an, wovon dieser Koeffizient noch weit entfernt ist.

Die weiteren Angaben über Wilks' Lambda, den Chi-Quadrat-Wert und das Signifikanzniveau beziehen sich nicht unmittelbar auf die einzelnen Funktionen, sondern auf unterschiedliche Kombinationen der Funktionen. Die erste Zeile enthält Angaben für den Fall, daß alle Funktionen (Funktion 1 bis 3) berücksichtigt werden, die zweite Zeile bezieht sich auf die Situation, daß alle Funktionen außer der ersten (also die Funktionen 2 und 3) Berücksichtigung finden, und in der dritten Zeile verbleibt schließlich nur noch die dritte Funktion. Wilks' Lambda ist der Quotient aus der Quadratsumme innerhalb der Gruppen und der gesamten Quadratsumme, kennzeichnet also den Anteil der Streuung innerhalb der Gruppen an der gesamten Streuung. Ergeben sich für dieses Maß große Werte, also Werte nahe bei 1, existieren kaum Unterschiede zwischen den Werten der einzelnen Gruppen, bei kleinen Werten lassen sich die einzelnen Gruppen dagegen gut voneinander trennen. Wilks' Lambda kann nun für verschiedene Kombinationen von Diskriminanzfunktionen berechnet werden. In der ersten Zeile wird Wilks' Lambda für die Situation angegeben, in der noch alle Funktionen zu berücksichtigen sind. Dieser Wert ist mit 0,642 bereits recht hoch, und der für diesen Wert durchgeführte Signifikanztest, der auf einer Chi-Quadrat-Transformation von Wilks' Lambda basiert, zeigt an, daß man mit einer Irrtumswahrscheinlichkeit von 0,000 davon ausgehen kann, daß sich die Mittelwerte der verschiedenen Gruppen auch in der Grundgesamtheit voneinander unterscheiden. Die Angaben der zweiten Zeile beziehen sich auf die Situation ohne Funktion 1. Es werden also alle Funk-

tionen außer der ersten berücksichtigt, in diesem Fall also die zweite und die dritte. Wilks' Lambda ist hier noch größer als bei der Berücksichtigung aller drei Funktionen, so daß sich ein kleinerer Chi-Quadrat-Wert ergibt. Die Signifikanz wird weiterhin mit 0,000 angegeben, ist jedoch bei den nicht mehr ausgewiesenen Dezimalstellen tatsächlich größer als der Signifikanzwert aus der ersten Zeile. Die unterste Zeile schließlich, in der nur noch die dritte Funktion berücksichtigt ist, weist den größten Wert für Wilks' Lambda und zugleich die höchste Irrtumswahrscheinlichkeit auf. Bei einer Irrtumswahrscheinlichkeit von 27,3% kann die Nullhypothese, die Gruppenmittelwerte in der Grundgesamtheit seien identisch, nicht mehr zurückgewiesen werden.

Alle bisher betrachteten Ergebnisse zeigen damit, daß die zweite und insbesondere die dritte Funktion einen sehr geringen Erklärungsanteil besitzen. In einem solchen Fall ist in Betracht zu ziehen, ob vor allem die dritte Funktion überhaupt in das Modell einbezogen werden soll. Allgemein gilt die Empfehlung, das Modell so einfach wie möglich zu halten, um damit zufällige Einflüsse, die mit der Stichprobenbetrachtung verbunden sind, zu minimieren. Der Übergang von einem einfachen zu einem komplizierteren Modell ist im allgemeinen nur dann gerechtfertigt, wenn dadurch eine deutliche Verbesserung des Erklärungsgehalts erreicht werden kann. Da dies hier in bezug auf die dritte Funktion offenbar nicht der Fall ist, sollte sie möglicherweise unberücksichtigt bleiben.

Wenn Sie anstatt des Dialogfeldes die Befehlssyntax zum Durchführen der Diskriminanzanalyse verwenden, können Sie mit dem Unterbefehl FUNCTIONS die Anzahl der zu berücksichtigenden Funktionen vorgeben und auf diese Weise die dritte Funktion ausschließen. Hierzu können Sie zunächst alle gewünschten Einstellungen in dem Dialogfeld vornehmen und diese anschließend mit der Schaltfläche *Einfügen* in die Syntaxsprache übersetzen lassen. Anschließend können Sie in den Syntaxbefehl folgende Zeile einfügen:

```
/functions 2
```

Der gesamte Befehl hat einschließlich dieses Unterbefehls - je nach zugrundeliegenden Einstellungen - ungefähr die folgende Struktur:

```
DISCRIMINANT
  /GROUPS=v325(1 6)
  /VARIABLES=v3 v37 v112 v142 v261
  /ANALYSIS ALL
  /FUNCTIONS 2
  /PRIORS  EQUAL
  /STATISTICS=TABLE
  /CLASSIFY=NONMISSING POOLED .
```

Der dargestellte Befehl verwendet weitgehend die im bisherigen Beispiel benutzten Dialogfeldeinstellungen (siehe S. 611), allerdings wurde zusätzlich der Unterbefehl /STATISTICS=TABLE eingefügt. Dieser Befehl entspricht der Option *Zusammenfassende Tabelle* aus dem Dialogfeld der Schaltfläche *Klassifizieren*. Die mit diesem Befehl angeforderte Tabelle wird in Abbildung 25.11 wiedergegeben.

Die obere Tabelle entspricht den bisher betrachteten Ergebnissen, bei denen alle drei Funktionen berücksichtigt werden. Die untere Tabelle bezieht sich auf eine Analyse, in der die dritte Funktion ausgeschlossen wurde. Dies ist die Analyse, die mit dem dargestellten Syntaxbefehl aufgerufen wird.

Es zeigt sich, daß die dritte Funktion tatsächlich nur einen sehr geringen Einfluß auf das Ergebnis der Diskriminanzanalyse hat. Bei der Berücksichtigung aller drei Funktionen werden 80 + 49 + 9 + 31 = 169 der insgesamt 328 Personen der richtigen Partei zugeordnet. Dies ist eine „Trefferquote“ von 52%. Wird die dritte Funktion nicht herangezogen, verringert sich die Anzahl der richtig zugeordneten Personen auf 80 + 52 + 6 + 23 = 161, so daß sich eine nur geringfügig niedrigere Trefferquote von 49% ergibt.

Klassifizierungsergebnisse[a]

		WAHLABSICHT, BUNDESTAGSWAHL; BEFR.	Vorhergesagte Gruppenzugehörigkeit				
			CDU-CSU	SPD	F.D.P.	BUENDNIS90-GRUENE	Gesamt
Original	Anzahl	CDU-CSU	80	23	22	17	143
		SPD	16	49	7	27	99
		F.D.P.	2	5	9	4	21
		BUENDNIS90-GRUENE	7	19	8	31	65
	%	CDU-CSU	56,2	16,3	15,4	12,1	100,0
		SPD	16,2	49,5	6,6	27,6	100,0
		F.D.P.	11,4	25,8	42,8	20,0	100,0
		BUENDNIS90-GRUENE	10,9	29,1	11,8	48,1	100,0

a. 51,7% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

Klassifizierungsergebnisse[a]

		WAHLABSICHT, BUNDESTAGSWAHL;	Vorhergesagte Gruppenzugehörigkeit				
			CDU-CSU	SPD	F.D.P.	BUENDNIS90-GRUENE	Gesamt
Original	Anzahl	CDU-CSU	80	23	30	9	143
		SPD	12	52	16	18	99
		F.D.P.	5	7	6	4	21
		BUENDNIS90-GRUENE	5	22	15	23	65
	%	CDU-CSU	56,2	16,3	20,8	6,7	100,0
		SPD	12,6	53,1	16,2	18,0	100,0
		F.D.P.	22,8	31,5	28,6	17,1	100,0
		BUENDNIS90-GRUENE	8,2	33,7	22,7	35,4	100,0

a. 49,4% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

Abbildung 25.11: Klassifizierungsergebnisse für eine Analyse mit drei Diskriminanzfunktionen (obere Tabelle) und für eine Analyse mit zwei Funktionen (untere Tabelle)

25.3.2 Streudiagramm der Gruppenzugehörigkeiten

Bei einer Diskriminanzanalyse mit mindestens zwei Diskriminanzfunktionen kann man ein Streudiagramm erstellen lassen, auf dessen Achsen die Funktionswerte der beiden ersten Diskriminanzfunktionen abgetragen werden. In der Diagrammfläche wird dann für jeden Fall der Datendatei ein Punkt eingezeichnet, dessen Lage von den Funktionswerten der beiden Diskriminanzfunktionen abhängt.

Durch unterschiedliche Formen oder Farben der in dem Diagramm dargestellten Punkte läßt sich die tatsächliche Gruppenzugehörigkeit (in diesem Fall also die tatsächliche Parteianhängerschaft) erkennen.[280] Abbildung 25.12 zeigt das Streudiagramm, das mit den auf S. 611 beschriebenen Dialogfeldeinstellungen erzeugt wurde. Das Diagramm wurde mit der Option *Kombinierte Gruppen* aus dem Dialogfeld *Klassifizieren* angefordert.

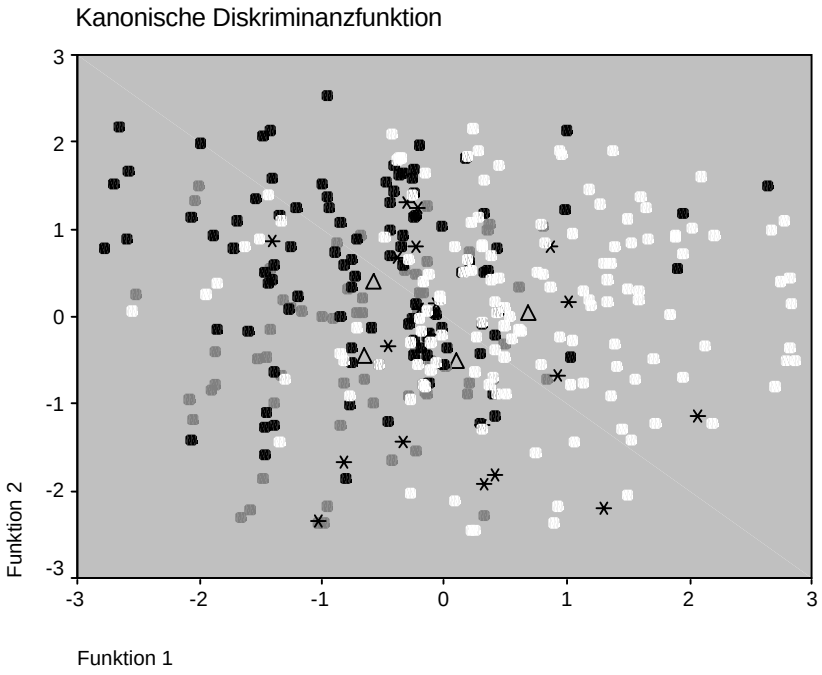


Abbildung 25.12: Streudiagramm, in dem die tatsächlichen Gruppenzugehörigkeiten auf der durch die beiden Diskriminanzfunktionen aufgespannten Fläche dargestellt werden

In dem dargestellten Diagramm wurde die Legende ausgeblendet, daher sei die Bedeutung der unterschiedlichen Symbole kurz genannt:

- **Schwarzer Punkt:** Wähler der SPD
- **Weißer Punkt:** Wähler von CDU/CSU
- **Grauer Punkt:** Wähler von Bündnis90/Grüne

[280] In früheren SPSS-Versionen konnte eine entsprechende Grafik auch für Diskriminanzanalysen mit nur zwei Diskriminanzfunktionen erstellt werden. Es wurde dann ein aus Textzeichen zusammengesetztes Histogramm erzeugt. In den neueren Versionen von SPSS steht diese Grafik jedoch nicht mehr zur Verfügung.

* **Stern:** Wähler der FDP

Δ **Dreieck:** Mittelpunkt einer der vier Gruppen

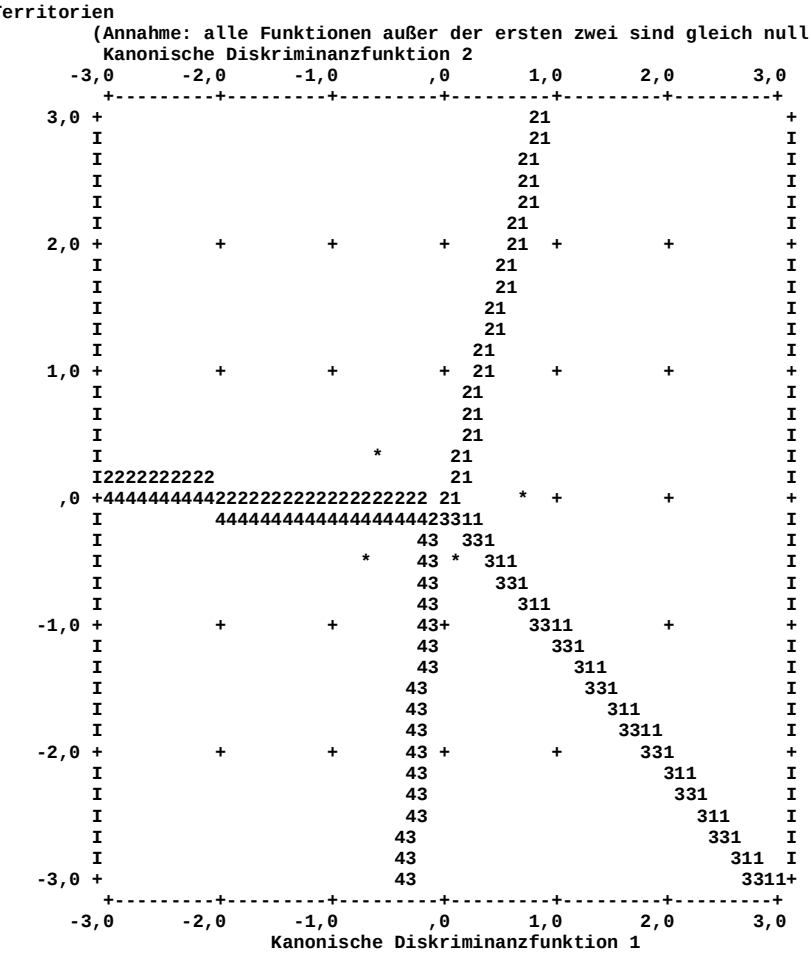
Auf der horizontalen Achse werden die Werte der ersten Funktion, auf der vertikalen Achse die der zweiten Funktion abgetragen. Die einzelnen Fälle der Datendatei werden in der Grafik jeweils durch einen Punkt repräsentiert, dessen Position sich aus den für den Fall berechneten Funktionswerten ergibt. Je nach Gruppenzugehörigkeit (Parteianhängerschaft) haben diese Punkte unterschiedliche Farben und/oder Formen. Es ist zu erkennen, daß sich die Punkte aller vier Gruppen grundsätzlich über die gesamte Diagrammfläche verteilen, allerdings ist in den Punkten folgendes Muster zu erkennen: Weiße Punkte (CDU/CSU-Wähler) finden sich eher in der rechten Diagrammhälfte und damit im Bereich hoher Werte der ersten Diskriminanzfunktion. Schwarze Punkte (SPD-Wähler) liegen überwiegend im linken oberen Viertel der Diagrammfläche, also im Bereich hoher Werte der zweiten Funktion und niedriger Werte der ersten Diskriminanzfunktion. Die Anhänger der Grünen weisen dagegen hauptsächlich niedrige Werte der ersten und zudem niedrige Werte der zweiten Funktion auf. Am diffusesten ist die Verteilung der FDP-Anhänger, denn die Sterne verteilen sich ohne deutlich erkennbares Muster mehr oder weniger über die gesamte Diagrammfläche.

Aus dieser Anordnung der Punkte im Streudiagramm ergibt sich folgende Handlungsanweisung für die Diskriminanzanalyse: Zeigt sich für eine Person ein hoher Wert bei der ersten Diskriminanzfunktion, ist diese der Gruppe der CDU/CSU-Wähler zuzuordnen. Bei einem niedrigen Funktionswert bei der ersten und einem hohen Wert bei der zweiten Funktion handelt es sich dagegen wahrscheinlich um einen SPD-Wähler. Sind dagegen beide Funktionswerte niedrig, wird man die Person in die Gruppe der Wähler von Bündnis90/Grüne einordnen. Schwierigkeiten bereitet dagegen das Erkennen von FDP-Wählern. Dies spiegelt sich auch in den Klassifizierungsergebnissen (siehe Abbildung 25.11, S. 616) wider, da (bei Verwendung aller drei Funktionen, obere Tabelle) nur 9 der insgesamt 21 FDP-Wähler richtig erkannt und zugleich 37 Personen fälschlich als Wähler der FDP klassifiziert wurden.

Die vier Dreiecke kennzeichnen die Funktionsmittelwerte der vier Gruppen. Die Beschriftungen der Dreiecke wurden ebenfalls ausgeblendet, jedoch entspricht die Lage der Dreiecke den Erwartungen, die sich aus der Verteilung der Punkte ergeben: Das linke obere Dreieck kennzeichnet die SPD-Gruppe und das linke untere Dreieck die Bündnis90/Grünen. Das am weitesten rechts liegende Dreieck steht für die CDU/CSU und das verbleibende, ziemlich in der Mitte der Grafik plazierte Dreieck markiert die Gruppe der FDP-Wähler. Insgesamt liegen die Dreiecke alle recht nahe beieinander. Dies ist ein weiterer Hinweis darauf, daß die Unterscheidung zwischen den Gruppen bei weitem nicht eindeutig und nur mit deutlichen Unsicherheiten möglich ist.

25.3.3 Gebietskarte der Gruppenzuordnungen

Eine weitere Möglichkeit zur Darstellung der Gruppenzuordnungen in Abhängigkeit von den beiden ersten Diskriminanzfunktionen ist das Erstellen einer Gebietskarte. Auch diese steht nur zur Verfügung, wenn mindestens zwei Diskriminanzfunktionen verwendet werden, wie es bei der Unterscheidung von mehr als zwei Gruppen in der abhängigen Variablen per Voreinstellung der Fall ist. Die Gebietskarte wurde in den Dialogfeldeinstellungen (siehe S. 611) mit der Option *Territorien* aus dem Dialogfeld *Klassifizieren* angefordert. Abbildung 25.13 stellt die Gebietskarte dar, die wie in früheren SPSS-Versionen noch immer aus einfachen Textzeichen zusammengesetzt ist.



(*wird fortgesetzt*)

```
Symbole für Territorien

Symbol  Grp.  Label
------  ----  ------------------
   1       1  CDU-CSU
   2       2  SPD
   3       3  F.D.P.
   4       6  BUENDNIS90-GRUENE
   *          Markiert Gruppenzentroide
```

Abbildung 25.13: Gebietskarte mit der Gruppenzuordnung in Abhängigkeit von den Funktionswerten

Auch hier werden wieder auf der horizontalen Achse die Werte der ersten und auf der vertikalen Achse die der zweiten Diskriminanzfunktion abgetragen. Die Diagrammfläche ist in vier Bereiche aufgeteilt, wobei jeder der vier Bereiche einer der vier Gruppen und damit einer der vier Parteien entspricht. Der obere, linke Bereich ist der Bereich von Gruppe 2, da dieser durch die Ziffer *2* von der übrigen Diagrammfläche abgetrennt ist. Aus der Legende unterhalb des Diagramms geht hervor, daß die Gruppe 2 den SPD-Wählern entspricht. Ergibt sich für eine Person aus der ersten und der zweiten Diskriminanzfunktion eine Kombination von Funktionswerten, die in dem Gebietsdiagramm in den linken, oberen Bereich fällt, wird diese Person daher als SPD-Wähler klassifiziert. Dies ist ungefähr dann der Fall, wenn sich bei der ersten Diskriminanzfunktion ein Wert unter 1 und bei der zweiten Funktion ein Wert über 0 ergibt.

Das Gebietsdiagramm spiegelt das Muster wider, das bereits in dem Streudiagramm (Abbildung 25.12, S. 617) zu erkennen war: Hohe Funktionswerte der ersten Diskriminanzfunktion kennzeichnen tendenziell einen CDU/CSU-Wähler, niedrige Werte bei beiden Funktionen einen Wähler der Grünen und ein niedriger Wert bei der ersten in Verbindung mit einem hohen Wert bei der zweiten Funktion markieren einen Wähler der SPD. FDP-Wähler weisen tendenziell hohe Werte bei der ersten und niedrige Werte bei der zweiten Funktion auf, allerdings ist dieser Bereich insgesamt sehr schmal.

Die vier Sternchen in dem Diagramm kennzeichnen wiederum die Funktionsmittelwerte der vier Gruppen und entsprechen damit den Dreiecken aus dem Streudiagramm. Die Pluszeichen haben dagegen keine inhaltliche Bedeutung, sondern sollen lediglich ein Gitternetz andeuten und damit der Orientierung innerhalb des Diagramms dienen.

25.4 Selektionsmethoden

25.4.1 Allgemeines Verfahren

In den beiden bisherigen Beispielen wurden die unabhängigen Variablen, anhand derer die Zuordnung zu den verschiedenen Gruppen der abhängigen Variablen vorzunehmen war, von vornherein fest vorgegeben. Alternativ kann man auch so vorgehen, daß man lediglich eine Reihe potentiell als erklärende Variablen in Be-

tracht kommende Variablen angibt, und die Auswahl der am besten für die Vorhersage der Gruppenzugehörigkeit geeigneten Variablen von der Diskriminanzanalyse vornehmen läßt. Dabei werden die Variablen nach einem schrittweisen Verfahren ausgewählt. Ein solches Verfahren steht bei SPSS auch bei der Regressionsanalyse zur Verfügung, wo zudem noch weitere Selektionsmethoden angeboten werden.

Die *Schrittweise* Selektionsmethode bezieht vor der Ausführung der ersten Selektionsstufe keine einzige Variable in die Analyse ein. Erst in mehreren aufeinanderfolgenden Schritten wird nach und nach über die Einbeziehung einzelner Variablen entschieden. Bei der Selektion der Variablen werden die folgenden Stufen durchlaufen:

- Im ersten Schritt wird aus allen potentiellen unabhängigen Variablen eine Variable als erklärende Variable des Modells ausgewählt. Dabei wird die Variable gewählt, für die sich der beste Wert des Selektionskriteriums ergibt. Als Selektionskriterium dient per Voreinstellung das Maß Wilks' Lambda, Sie können jedoch auch andere Kriterien verwenden. Für Wilks' Lambda wird ein möglichst kleiner Wert angestrebt. (Das Maß kennzeichnet das Verhältnis der Streuung innerhalb der Gruppen zur gesamten Streuung, siehe auch *Wilks' Lambda*, S. 602) Bei der Verwendung von Wilks' Lambda wird damit auf der ersten Stufe des Selektionsverfahrens von allen potentiellen erklärenden Variablen die Variable ausgewählt, für die sich das kleinste Wilks' Lambda ergibt.
- Im zweiten Schritt wird aus den verbleibenden, noch nicht aufgenommen Variablen erneut eine Variable ausgewählt, wobei wiederum das Selektionskriterium aus dem ersten Schritt zur Anwendung kommt. Anschließend wird die zuerst aufgenommene Variable daraufhin überprüft, ob sie nach wie vor dem Kriterium zum Verbleib in der Gleichung genügt, das heißt, es wird untersucht, ob ihr Erklärungswert auch unter Berücksichtigung der zweiten Variablen groß genug ist, um eine Verwendung dieser Variablen in der Diskriminanzfunktion zu rechtfertigen. Wie groß der Einfluß einer Variablen sein muß, damit sie in dem Modell verbleibt, können Sie manuell festlegen.[281] Genügt die erste Variable nicht mehr den gestellten Anforderungen, wird sie wieder aus der Funktion entfernt.
- Anschließend folgen weitere Schritte, in denen wie im zweiten Schritt verfahren wird: Zunächst wird aus den verbleibenden potentiellen erklärenden Variablen diejenige herausgesucht, die das Auswahlkriterium am stärksten positiv beeinflußt. Nachdem diese Variable in die Funktion aufgenommen wurde, wird für die zuvor aufgenommenen Variablen geprüft, ob sie nach wie vor die Kriterien zum Verbleib in der Funktion erfüllen. Ist dies bei einzelnen Variablen nicht der Fall, werden sie wieder aus dem Modell entfernt.

[281] Dies geschieht in dem Dialogfeld der Schaltfläche *Methode*, siehe, S. 630.

- Das Selektionsverfahren wird beendet, wenn eines der folgenden drei Ereignisse eintritt:
  - Alle Variablen sind in das Modell aufgenommen.
  - Keine der noch nicht aufgenommenen Variablen erfüllt das Aufnahmekriterium.[282]
  - Es wurde eine Höchstzahl an Iterationsschritten erreicht.

25.4.2 Ergebnisse des Beispiels

Im folgenden soll eine schrittweise Auswahl der unabhängigen Variablen für die Unterscheidung der vier Parteien (CDU/CSU, SPD, Bündnis90/Grüne, FDP) erfolgen. Hierzu werden die bisherigen Dialogfeldeinstellungen (S. 611) übernommen, es wird lediglich zusätzlich im Hauptdialogfeld die Option *Schrittweise Methode verwenden* ausgewählt. Die Voreinstellungen in dem Dialogfeld der Schaltfläche *Methode* werden unverändert übernommen. Bei Verwendung einer Selektionsmethode wird der Output der Diskriminanzanalyse um eine Beschreibung der Selektionsschritte und -ergebnisse ergänzt. Abbildung 25.14 gibt diesen zusätzlichen Output wieder.

Aufgenommene/Entfernte Variablen[a,b,c,d]

Schritt	Aufgenommen	Wilks-Lambda											
						Exaktes F				Näherungsweises F			
		Statistik	df1	df2	df3	Statistik	df1	df2	Signifikanz	Statistik	df1	df2	Signifikanz
1	LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	,722	1	3	323,426	41,536	3	323,426	,000				
2	ALLGEMEINER SCHULABSCHLUSS	,677	2	3	323,426	23,120	6	644,853	,000				
3	ALTER: BEFRAGTE<R>	,652	3	3	323,426					16,714	9	782,418	,000

Bei jedem Schritt wird die Variable aufgenommen, die das gesamte Wilks-Lambda minimiert.

a. Maximale Anzahl der Schritte ist 10.

b. Minimaler partieller F-Wert für die Aufnahme ist 3.84.

c. Maximaler partieller F-Wert für den Ausschluß ist 2.71.

d. F-Niveau, Toleranz oder VIN sind für eine weitere Berechnung unzureichend.

(*wird fortgesetzt*)

[282] Die Schwelle des Aufnahmekriteriums können Sie in dem Dialogfeld *Methode* vorgeben, siehe S. 630.

Variablen in der Analyse

Schritt		Toleranz	F-Wert für den Ausschluß	Wilks-Lambda
1	LINKS-RECHTS-SELBST EINSTUFUNG, BEFR.	1,000	41,536	
2	LINKS-RECHTS-SELBST EINSTUFUNG, BEFR.	,991	41,143	,937
	ALLGEMEINER SCHULABSCHLUSS	,991	7,079	,722
3	LINKS-RECHTS-SELBST EINSTUFUNG, BEFR.	,952	40,213	,896
	ALLGEMEINER SCHULABSCHLUSS	,931	4,611	,680
	ALTER: BEFRAGTE<R>	,895	4,182	,677

Variablen, die NICHT in der Analyse sind

Schritt		Toleranz	Minimale Toleranz	F-Wert für die Aufnahme	Wilks-Lambda
0	ERHEBUNGSGEBIET: WEST - OST	1,000	1,000	1,425	,987
	ALTER: BEFRAGTE<R>	1,000	1,000	7,524	,935
	LINKS-RECHTS-SELBST EINSTUFUNG, BEFR.	1,000	1,000	41,536	,722
	ALLGEMEINER SCHULABSCHLUSS	1,000	1,000	7,306	,937
	BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	1,000	1,000	,210	,998
1	ERHEBUNGSGEBIET: WEST - OST	,992	,992	1,150	,714
	ALTER: BEFRAGTE<R>	,953	,953	6,640	,680
	ALLGEMEINER SCHULABSCHLUSS	,991	,991	7,079	,677
	BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	,995	,995	,593	,718
2	ERHEBUNGSGEBIET: WEST - OST	,985	,982	,715	,673
	ALTER: BEFRAGTE<R>	,895	,895	4,182	,652
	BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	,888	,885	1,226	,670
3	ERHEBUNGSGEBIET: WEST - OST	,977	,888	,438	,649
	BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	,862	,811	1,072	,645

Wilks-Lambda

						Exaktes F				Näherungsweises F			
Schritt	Anzahl der Variablen	Lambda	df1	df2	df3	Statistik	df1	df2	Signifikanz	Statistik	df1	df2	Signifikanz
1	1	,722	1	3	323	41,536	3	323,426	,000				
2	2	,677	2	3	323	23,120	6	644,853	,000				
3	3	,652	3	3	323					16,714	9	782,418	,000

Abbildung 25.14: Beschreibung des Selektionsprozesses zur Auswahl der unabhängigen Variablen

In der obersten Tabelle mit der Überschrift *Aufgenommene/Entfernte Variablen* werden in den Fußnoten die Kriterien mitgeteilt, an denen sich das Selektionsverfahren orientierte: Die maximale Anzahl der Iterationsschritte war von vornherein auf zehn beschränkt. Diese Höchstzahl von Iterationen wird von SPSS festgelegt und ergibt sich als das Doppelte der Anzahl potentieller unabhängige Variablen.[283]

Als Kriterium für die Aufnahme oder den Ausschluß einer Variablen dient Wilks' Lambda. Da die Veränderung, die sich an Wilks' Lambda durch die Berücksichtigung einer Variablen ergibt, auch rein zufällig aus der vorliegenden Stichprobe resultieren kann, wird die Signifikanz der Veränderung zur Beurteilung herangezogen. Die Signifikanz wird anhand des F-Wertes gemessen. Damit eine Variable in das Modell aufgenommen wird, muß der partielle F-Wert für diese Variable eine vorgegebene Untergrenze überschreiten. Diese Untergrenze entspricht per Voreinstellung dem Wert 3,84, kann jedoch in dem Dialogfeld *Methode* verändert werden. In diesem Beispiel wurde die Voreinstellung beibehalten, wie aus den Angaben unter der Tabelle *Aufgenommene/Entfernte Variablen* hervorgeht. Wird eine neue Variable in das Modell aufgenommen, kann sich der partielle F-Wert der bereits im Modell enthaltenen Variablen dadurch verändern. Verringert sich der F-Wert einer Variablen, nimmt die Signifikanz für deren Einfluß auf die Gruppenzuordnung eines Falles ab. Wenn der F-Wert eine vorgegebene Obergrenze unterschreitet, wird die Variable daher wieder aus dem Modell ausgeschlossen. Diese Obergrenze beträgt per Voreinstellung 2,71, kann jedoch ebenfalls in dem Dialogfeld *Methode* geändert werden.

In den beiden Tabellen *Variablen in der Analyse* und *Variablen, die NICHT in der Analyse sind* werden die einzelnen Schritte des Selektionsprozesses beschrieben. Die Tabelle *Variablen, die NICHT in der Analyse sind* führt im obersten Bereich die fünf potentiellen erklärenden Variablen auf. Für jede dieser Variablen werden auch Wilks' Lambda und der F-Wert angegeben. Das kleinste Wilks' Lambda und damit den größten F-Wert weist die Variable *Links-Rechts-Selbsteinstufung* auf. Daher wird diese Variable im ersten Selektionsschritt ausgewählt und in das Modell aufgenommen. Das Ergebnis des ersten Selektionsschrittes kann in der Tabelle *Variablen in der Analyse* abgelesen werden. Nach dem ersten Schritt ist die Variable *Links-Rechts-Selbsteinstufung* als einzige Variable im Modell enthalten.

Die Tabelle *Variablen, die NICHT in der Analyse sind* gibt wiederum an, daß die vier Variablen *Erhebungsgebiet*, *Alter*, *Allgemeiner Schulabschluß* und *Nettoeinkommen* nach dem ersten Schritt noch nicht in das Modell aufgenommen wurden. Für diese vier Variablen werden neben Wilks' Lambda und dem F-Wert auch die *Toleranz* sowie die *Minimale Toleranz* angegeben. Die Toleranz ist ein Maß dafür, wie stark die unabhängigen Variablen untereinander korreliert sind. Besteht zwischen zwei Variablen ein starker linearer Zusammenhang, kann der Erklärungswert dieser Variablen nicht zuverlässig bestimmt werden, da sich nicht ermitteln läßt, welcher der beiden korrelierten Variablen der Einfluß zuzuschreiben ist. Noch nicht in das Modell aufgenommene unabhängige Variablen, die einen star-

[283] Durch Verwendung der Befehlssyntax können Sie die Höchstzahl der Iterationen manuell verändern, indem Sie den Unterbefehl `MAXSTEPS` einfügen.

ken linearen Zusammenhang zu den bereits im Modell enthaltenen Variablen aufweisen, sollen daher auch dann nicht in das Modell aufgenommen werden, wenn sich für sie ein hoher F-Wert ergibt. Dabei wird die Entscheidung über Aufnahme oder Nichtaufnahme anhand des Toleranzwertes getroffen. Die Toleranz ist definiert als $1 - R_i^2$, wobei R_i^2 der quadrierte multiple Korrelationskoeffizient zwischen der jeweils betrachteten Variablen und den bereits im Modell enthaltenen Variablen ist. Liegt hier eine starke Korrelation vor, ergibt sich ein hohes R_i^2 und damit ein niedriger Toleranzwert. Per Voreinstellung wird eine Variable dann nicht in das Modell aufgenommen, wenn der Toleranzwert unter 0,001 liegt. Zudem wird einer Variablen die Aufnahme in das Modell verweigert, wenn ihre Aufnahme das Toleranzniveau einer bereits in dem Modell enthaltenen Variablen auf einen Wert unter 0,001 verringern würde.[284] Nach dem ersten Iterationsschritt sind die Toleranzwerte aller noch nicht im Modell enthalten Variablen so hoch, daß von daher kein Grund für die Nichtaufnahme in das Modell besteht.

Anstatt die Korrelation der ausgeschlossenen Variablen zu der Gesamtheit der im Modell enthaltenen Variablen zu betrachten, kann man auch die Korrelation zwischen einer ausgeschlossenen und den einzelnen im Modell enthaltenen Variablen untersuchen. Dies geschieht in der Spalte *Minimale Toleranz*, in der der niedrigste Toleranzwert (und damit indirekt die höchste Korrelation) zwischen der jeweiligen noch nicht im Modell enthaltenen Variablen und den einzelnen bereits in das Modell aufgenommenen Variablen angegeben wird. Da sich nach dem ersten Iterationsschritt erst eine Variable im Modell befindet, sind die *Toleranz* und die *Minimale Toleranz* auf dieser Stufe identisch.

Der F-Wert sowie der Toleranzwert werden auch in der Tabelle *Variablen in der Analyse* für die bereits in das Modell aufgenommenen Variablen angegeben. Hier bezieht sich der Toleranzwert auf die Korrelation zwischen einer im Modell enthaltenen Variablen und der Gesamtheit der anderen bereits ins Modell aufgenommenen Variablen. Da das Modell nach dem ersten Schritt erst eine Variable umfaßt, kann der Toleranzwert noch nicht berechnet werden und ist daher auf 1 gesetzt. Auf späteren Stufen sind die Toleranz- und F-Werte der im Modell enthaltenen Variablen entscheidend für die Frage, ob eine Variable wieder aus dem Modell ausgeschlossen werden soll.

Im zweiten Selektionsschritt wird die Variable *Allgemeiner Schulabschluß* in das Modell aufgenommen, da diese nach dem ersten Schritt unter den verbliebenen Variablen den höchsten F-Wert (bzw. das niedrigste Wilks' Lambda) aufwies und zudem der F-Wert mit 7,079 über dem Grenzwert von 3,84 (s.o.) liegt und der Toleranzwert mit 0,953 hinreichend hoch ist.

Auch nach der Aufnahme der zweiten unabhängigen Variablen in das Modell sind der F-Wert und die Toleranz der beiden nun im Modell enthaltenen Variablen so groß, daß keine der Variablen wieder aus dem Modell auszuschließen ist.

---

[284] Der Grenzwert von 0,001 ist die Voreinstellung, die in den Dialogfeldern der Diskriminanzanalyse auch nicht abgeändert werden kann. Wenn Sie allerdings die Befehlssyntax verwenden, können Sie mit dem Unterbefehl `TOLERANCE`, der im Anschluß an den Befehl `METHOD` aufgeführt werden muß, einen beliebigen Toleranzwert zwischen 0 und 1 vorgeben.

Nach den gleichen Kriterien wird im dritten Selektionsschritt die Variable *Alter* in das Modell aufgenommen. In der Tabelle *Variablen, die NICHT in der Analyse sind* ist zu erkennen, daß nach dem dritten Schritt die Variablen *Erhebungsgebiet* und *Nettoeinkommen* noch nicht aufgenommen wurden. Da für beide Variablen der F-Wert mit 0,438 bzw. 1,072 deutlich unter dem Grenzwert von 3,84 liegt, werden diese Variablen dem Modell auch im vierten Schritt nicht hinzugefügt. Der gesamte Selektionsprozeß wird nach dem dritten Schritt beendet, da keine der verbleibenden Variablen einen hinreichend großen partiellen F-Wert aufweist, um in das Modell aufgenommen zu werden.

Wenn Sie ein Selektionsverfahren anwenden und damit aus einer Reihe potentieller erklärender Variablen einige für die Diskriminanzanalyse gut geeignete Variablen auswählen, beachten Sie, daß damit keinesfalls immer die Variablen ausgewählt werden, die den engsten kausalen Zusammenhang zur abhängigen Variablen aufweisen. Die ausgewählten Variablen zeichnen sich lediglich dadurch aus, daß sie für die Fälle der Stichprobe gut geeignet sind, auf unterschiedliche Werte der abhängigen Variablen zu schließen. Wenn jedoch tatsächlich kein kausaler Zusammenhang zur abhängigen Variablen vorliegt, besteht die Gefahr, daß sich der Erklärungswert der ausgewählten Variablen auf die Stichprobe beschränkt, die Analyse also nicht für Prognosezwecke oder für die Anwendung auf Fälle außerhalb der Stichprobe geeignet ist. Soweit möglich, ist es daher stets besser, zunächst eine Theorie über mögliche Zusammenhänge zwischen erklärenden Variablen und der abhängigen Variablen zu entwickeln. Wenn die in der Theorie plausibel formulierten Zusammenhänge auch tatsächlich beobachtet werden können, muß der formulierte Zusammenhang zwar immer noch nicht zwingend zutreffen, die Wahrscheinlichkeit dafür ist jedoch sehr viel größer, als wenn die erklärenden Variablen durch Versuch und Irrtum aus einer Reihe in Frage kommender Variablen ausgewählt werden.

25.5 Einstellungen der Diskriminanzanalyse

25.5.1 Allgemeine Vorgehensweise

Um eine Diskriminanzanalyse auszuführen, wählen Sie den Befehl

STATISTIK
KLASSIFIZIEREN ▸
DISKRIMINANZANALYSE...

Dieser Befehl öffnet das Dialogfeld aus Abbildung 25.15. Unmittelbar nach dem Öffnen das Dialogfeldes wird dessen unterer Bereich mit dem Eingabefeld *Auswahlvariable* nicht angezeigt. Dieser Bereich muß erst mit der Schaltfläche *Auswählen* eingeblendet werden.

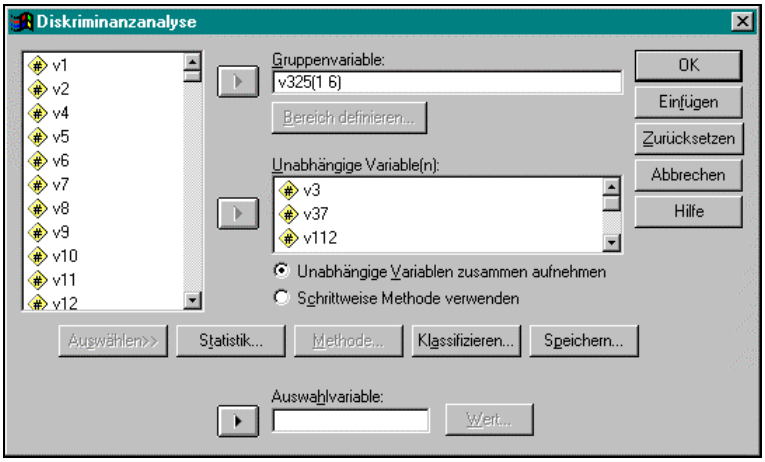


Abbildung 25.15: Dialogfeld des Befehls STATISTIK, KLASSIFIZIEREN, DISKRIMINANZANALYSE

In der Variablenliste werden sämtliche numerischen Variablen der Datendatei aufgeführt. Textvariablen können in der Diskriminanzanalyse nicht verwendet werden. Um eine Diskriminanzanalyse durchzuführen, nehmen Sie die folgenden Einstellungen vor:

- **Abhängige Variable:** Geben Sie in dem Feld *Gruppenvariable* die abhängige Variable an, deren Werte durch die Diskriminanzanalyse erklärt bzw. vorhergesagt werden sollen. Zusätzlich muß für diese Variable in dem Dialogfeld der Schaltfläche *Bereich definieren* der zu berücksichtigende Wertebereich festgelegt werden, siehe Abschnitt *Wertebereich für die abhängige Variable festlegen*, S. 628.
- **Erklärende Variable(n):** Verschieben Sie die erklärende(n) Variable(n) des Modells in das Feld *Unabhängige Variable(n)*. Nach der Angabe der abhängigen und mindestens einer unabhängigen Variablen sind die Mindestspezifikationen vorgenommen. Wenn Sie bei den anderen Optionen die Voreinstellungen verwenden möchten, können Sie die Analyse nun mit der Schaltfläche *OK* starten.
- **Methode:** Per Voreinstellung werden alle ausgewählten unabhängigen Variablen auch tatsächlich in dem Diskriminanzmodell verwendet. Von dieser Voreinstellung können Sie jedoch zugunsten einer *Schrittweisen Methode* abweichen. Eine schrittweise Methode wählt zunächst die am besten geeignetsten unabhängigen Variablen aus, und berücksichtigt in dem abschließenden Modell nur diese Variablen (siehe Abschnitt *25.5.3, Selektionsmethode*, S. 630).
- **Auswahl:** Soll die Analyse nicht für sämtliche Fälle der Datendatei, sondern nur für eine Auswahl von Fällen durchgeführt werden, können Sie dies mit einer *Auswahlvariablen* veranlassen. Voraussetzung hierfür ist jedoch, daß alle zu berücksichtigenden Fälle dadurch gekennzeichnet sind, daß sie in der Aus-

wahlvariablen den gleichen Wert aufweisen, der zudem in keinem der nicht zu berücksichtigenden Fälle vorkommen darf (siehe Abschnitt *Fälle auswählen*, S. 629).

- **Statistik:** Sie können den Standard-Output der Analyse um zahlreiche Statistiken zur Beschreibung der Analyseergebnisse ergänzen (siehe Abschnitt *25.5.4, Statistiken*, S. 632).
- **Klassifizieren:** Auch in dem Dialogfeld dieser Schaltfläche können Sie unter anderem zusätzlichen Output wie zum Beispiel eine Gebietskarte oder ein Streudiagramm anfordern. Zudem können Sie hier auch Einstellungen für das Analyseverfahren vornehmen und damit die A-priori-Wahrscheinlichkeiten sowie die Behandlung von fehlenden Werten steuern (siehe Abschnitt *25.5.5, Klassifizieren*, S. 634).
- **Speichern:** Das Dialogfeld dieser Schaltfläche bietet die Möglichkeit, die von der Diskriminanzanalyse vorgenommenen Gruppenzuordnungen sowie die Wahrscheinlichkeit für die jeweilige Gruppenzugehörigkeit in der Datendatei als neue Variablen zu speichern, siehe Abschnitt *25.5.6, Speichern von Ergebnissen in der Datendatei*, S. 637.

25.5.2 Variablen angeben und Fälle auswählen

Für die Diskriminanzanalyse müssen Sie eine abhängige und mindestens eine unabhängige Variable festlegen. Die unabhängigen (erklärenden) Variablen können Sie hierzu einfach in das Feld *Unabhängige Variable(n)* verschieben. Entsprechend wird die abhängige Variable in das Feld *Gruppenvariable* eingefügt, jedoch muß für diese Variable zudem der zu berücksichtigende Wertebereich festgelegt werden.

Wertebereich für die abhängige Variable festlegen

Um den Wertebereich für die unabhängige Variable festzulegen, öffnen Sie mit der Schaltfläche *Bereich definieren* das Dialogfeld aus Abbildung 25.16. Diese Schaltfläche ist nur aktiv, wenn die abhängige Variable bereits in das Feld *Gruppenvariable* eingefügt wurde und zudem dort aktuell markiert ist.

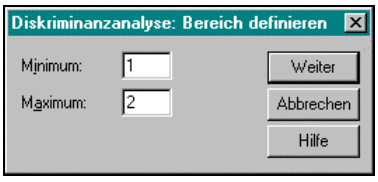


Abbildung 25.16: Dialogfeld der Schaltfläche „Bereich definieren“

Die Gruppen, die in der abhängigen Variablen unterschieden werden sollen, müssen durch unterschiedliche ganzzahlige Werte in der abhängigen Variablen gekennzeichnet sein. Die Werte müssen zudem einem zusammenhängenden Werte-

bereich entstammen, es ist jedoch nicht erforderlich, daß alle Werte dieses Wertebereichs tatsächlich vertreten sind. Wenn Sie beispielsweise drei Gruppen unterscheiden, können diese durch die Werte *1*, *3* und *4* gekennzeichnet sein, sofern der Wert *2* in der abhängigen Variablen nicht vertreten ist. Fälle, die in der abhängigen Variablen einen Wert außerhalb des Wertebereichs aufweisen, werden in der Analyse als ungruppierte Fälle angesehen. Bei der Bestimmung der Diskriminanzfunktion werden diese Fälle nicht berücksichtigt, bei der Anwendung der Funktion werden jedoch auch diese Fälle einer der gültigen Gruppen zugewiesen.

Um den Wertebereich zu definieren, geben Sie den niedrigsten zu berücksichtigenden Wert in des Feld *Minimum* und den höchsten zu berücksichtigenden Wert in das Feld *Maximum* ein. Jeder in der abhängigen Variablen vorkommende ganzzahlige Wert bildet eine Gruppe. Enthält die Variable auch Werte mit Dezimalstellen, werden diese abgeschnitten. Der Wert *3,9* wird damit als *3* und der Wert *0,9* als *0* angesehen.

Fälle auswählen

Sie können die Analyse auf eine Auswahl von Fällen aus der Datendatei beschränken. Die Fälle müssen dadurch gekennzeichnet sein, daß sie in einer beliebigen numerischen Auswahlvariablen alle den gleichen Wert aufweisen, der zudem in keinem der nicht zu berücksichtigenden Fälle vorkommen darf. Um eine derartige Auswahl der Fälle vorzunehmen, fügen Sie die entsprechende Variable in das Feld *Auswahlvariable* ein. Dieses Feld wird unmittelbar nach dem Öffnen des Dialogfeldes nicht angeboten, sondern erst durch die Schaltfläche *Auswählen* eingeblendet. Nachdem Sie die Auswahlvariable in das Feld eingefügt haben, müssen Sie noch den Wert angeben, den die zu berücksichtigenden Fälle in dieser Variablen aufweisen. Öffnen Sie hierzu mit der Schaltfläche *Wert* das Dialogfeld aus Abbildung 25.17, und geben Sie den Wert in das Feld *Wert der Auswahlvariablen* ein. Dieser Wert darf auch Dezimalstellen aufweisen.

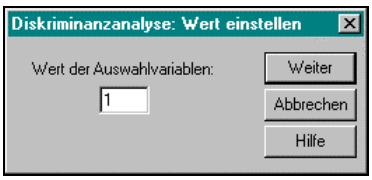


Abbildung 25.17: Dialogfeld der Schaltfläche „Wert“ für die Auswahlvariable

25.5.3 Selektionsmethode

Sie können bei der Diskriminanzanalyse eine Selektionsmethode verwenden, um aus den Variablen, die Sie in das Feld *Unabhängige Variable(n)* eingefügt haben, die besten erklärenden Variablen auswählen zu lassen. Hierzu dienen zunächst die beiden folgenden Optionen:

- ***Unabhängige Variablen zusammen aufnehmen:*** Diese Option ist voreingestellt. Alle ausgewählten unabhängigen Variablen werden zur Erklärung der Gruppenzugehörigkeit verwendet.
- ***Schrittweise Methode verwenden:*** Wählen Sie diese Option, um aus den unabhängigen Variablen zunächst die zur Vorhersage der Gruppenzugehörigkeit am besten geeigneten Variablen auszuwählen. Diese Auswahl kann sowohl dazu führen, daß alle angegebenen Variablen benutzt werden, als auch zu dem Ergebnis, daß keine der Variablen zur Vorhersage der Werte aus der abhängigen Variablen geeignet ist. Wenn Sie eine *Schrittweise Methode* verwenden, können Sie diese in dem Dialogfeld *Methode* noch näher spezifizieren, Sie können jedoch auch die Voreinstellungen von dort übernehmen.

Merkmale des Selektionsprozesses festlegen

Sie können zwischen fünf verschiedenen schrittweisen Methoden wählen. Die Methoden unterscheiden sich in dem Auswahlkriterium für die Aufnahme einer Variablen in das Diskriminanzmodell. Zusätzlich können Sie für jede der Methoden die Grenzwerte vorgeben, anhand derer die Eignung einer Variablen festgelegt wird. Um diese Einstellungen vorzunehmen, öffnen Sie mit der Schaltfläche *Methode* das Dialogfeld aus Abbildung 25.18. In diesem Dialogfeld können Sie auch den Umfang des Output bestimmen, mit dem der Selektionsprozeß beschrieben wird.

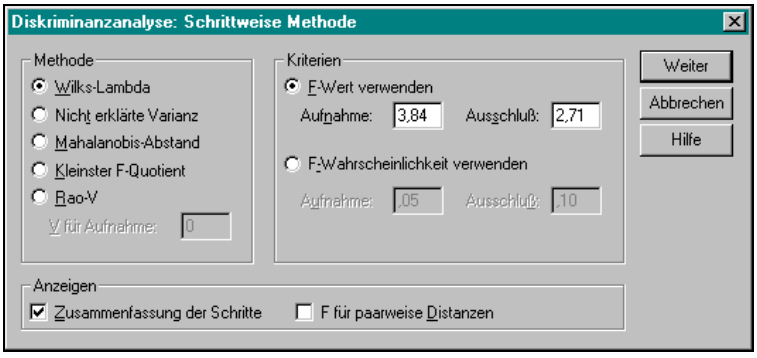


Abbildung 25.18: Dialogfeld der Schaltfläche „Methode“ zur Beschreibung des Selektionsprozesses

Methode

Mit der Methode legen Sie die Maßzahl fest, an der sich die Auswahl der Variablen orientiert:

- ***Wilks' Lambda:*** Diese Option ist voreingestellt, so daß der Wert von Wilks' Lambda berechnet wird, um daran den Erklärungswert einer Variablen zu messen. Es wird jeweils die Variable ausgewählt, mit der sich für das gesamte bisherige Modell der kleinste Wilks' Lambda-Wert ergibt.
- ***Nicht erklärte Varianz:*** Hiermit wird jeweils die Variable ausgewählt, bei der sich die geringste nicht erklärte Varianz für das Modell ergibt.
- ***Mahalanobis-Abstand:*** Gesucht wird jeweils die Variable, durch die sich der größte Mahalanobis-Abstand für die beiden am dichtesten beieinanderliegenden Gruppen ergibt.
- ***Kleinster F-Quotient:*** Auf jeder Stufe wird die Variable ausgewählt, bei der der geringste F-Quotient, der sich zwischen jeweils zwei der Gruppen ergibt, am größten ist.
- ***Raos V:*** Hiermit wird die Variable in das Modell aufgenommen, durch die Raos V am stärksten zunimmt. Bei dieser Option können Sie zusätzlich einen Mindestwert angeben, um den sich Raos V mindestens vergrößern muß, damit eine Variable in das Modell aufgenommen wird. Geben Sie diesen Wert in das Feld *V für Aufnahme* an. Per Voreinstellung ist eine 0 angegeben, so daß kein Mindestwert gefordert wird.

Kriterien

In der Gruppe *Methode* haben Sie die Maßzahl festgelegt, die als Kriterium für die Auswahl der Variablen dienen soll. Auf jeder Stufe wird die Variable in das Modell aufgenommen, die die jeweilige Maßzahl am stärksten positiv beeinflußt. Da bei dem Selektionsverfahren jedoch nur solche Variablen ausgewählt werden sollen, die einen Mindesterklärungsgehalt aufweisen, ist zusätzlich eine Untergrenze für den Einfluß der Variablen erforderlich. Diese Grenze wird anhand der sich jeweils ergebenden F-Statistik festgelegt. Dabei können Sie wahlweise den *F-Wert* oder die *Signifikanz von F* verwenden:

- ***F-Wert verwenden:*** Diese Option ist voreingestellt. Damit wird der erforderliche Mindesteinfluß einer Variablen direkt über den F-Wert festgelegt. Geben Sie hierzu die beiden folgenden Werte an:
  - ***Aufnahme:*** Legen Sie den F-Wert fest, den eine Variable mindestens liefern muß, um in das Modell aufgenommen zu werden. Per Voreinstellung wird ein F-Wert von mindestens 3,84 gefordert. Sie können jedoch einen beliebigen anderen Wert größer 0 verwenden.
  - ***Ausschluß:*** Geben Sie hier den F-Wert an, den eine Variable, die bereits im Modell enthalten ist, erreichen oder unterschreiten muß, um aus dem Modell wieder ausgeschlossen zu werden. Dieser F-Wert muß unter dem

Aufnahmekriterium liegen und größer als 0 sein. Voreingestellt ist ein Wert von 2,71.

- ***F-Wahrscheinlichkeit verwenden:*** Anstatt direkt die F-Werte zu betrachten, können Sie auch die Signifikanz der Werte berechnen und als Kriterium ansetzen. Geben Sie hierzu entsprechend die beiden folgenden Werte vor:
  - ***Aufnahme:*** Legen Sie die Irrtumswahrscheinlichkeit fest, die sich für den F-Wert höchstens ergeben darf, damit eine Variable in das Modell aufgenommen wird.
  - ***Ausschluß:*** Ergibt sich für eine in dem Modell bereits enthaltene Variable eine Irrtumswahrscheinlichkeit gleich oder größer dem hier angegebenen Wert, wird sie aus dem Modell wieder herausgenommen.

Anzeigen

In dieser Gruppe können Sie den Umfang des Output festlegen, mit dem die einzelnen Schritte des Selektionsprozesses beschrieben werden:

- ***Zusammenfassung der Schritte:*** Hiermit werden die einzelnen Schritte des Selektionsverfahrens beschrieben. Dazu wird für jeden Schritt angegeben, welche Variablen in das Modell neu aufgenommen und ggf. welche Variablen wieder ausgeschlossen wurden, aus welchen Variablen sich das Modell vor und nach einem Schritt zusammensetzt und welche Variablen noch nicht berücksichtigt wurden. Zudem werden für jeden Schritt Wilks' Lambda sowie der entsprechende F-Wert und die Signifikanz für das Modell angegeben. Für die Auswahl der Variablen werden zudem die Toleranzwerte sowie die bei der jeweiligen Selektionsmethode verwendeten Maßzahlen mitgeteilt.
- ***F für paarweise Distanzen:*** Für die einzelnen Gruppen werden jeweils paarweise F-Koeffizienten berechnet und in einer Matrix ausgegeben. Wenn Sie die Option *Ergebnisse bei jedem Schritt* angekreuzt haben, wird diese Matrix ebenfalls für jeden Schritt angegeben, andernfalls wird die Matrix nur für das Endergebnis, also für die Zuordnung nach dem letzten Schritt, erstellt.

25.5.4 Statistiken

Abbildung 25.19 zeigt das Dialogfeld der Schaltfläche *Statistik*, in dem Sie den Standard-Output der Diskriminanzanalyse um zusätzliche Tabellen mit Statistiken wie deskriptiven Maßzahlen, Funktionskoeffizienten und Kovarianzmatrizen ergänzen können.

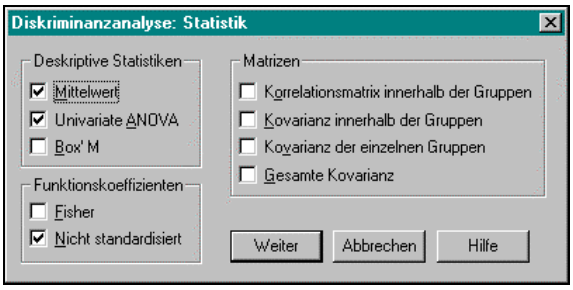


Abbildung 25.19: Dialogfeld der Schaltfläche „Statistik“ zum Anfordern von weiterem Prozedur-Output

Deskriptive Statistiken

- ***Mittelwert:*** Für jede erklärende (unabhängige) Variable werden der Mittelwert und die Standardabweichung jeweils getrennt für die verschiedenen Gruppen sowie für die Gesamtheit der Fälle berechnet. Die Angaben für die Gesamtheit der Fälle beziehen sich nur auf die in die Analyse einbezogenen Fälle und damit nicht zwingend auf die gesamte Datendatei. Für die mit dieser Option angeforderten Angaben wird nicht eine eigenständige Tabelle erstellt, sondern sie werden in die Tabelle *Gruppenstatistik* eingefügt. Eine solche Tabelle einschließlich der durch die Option *Mittelwert* angeforderten Maßzahlen ist in Abbildung 25.8, S. 608 dargestellt.
- ***Univariate ANOVA:*** Für jede der erklärenden Variablen wird mit einer Varianzanalyse die Hypothese getestet, daß die Variablenwerte für alle Gruppen der abhängigen Variablen gleich sind. Für jede Variable werden Wilks' Lambda, der F-Wert, die Freiheitsgrade und die Signifikanz der Hypothese angegeben. Kleine Signifikanzen deuten darauf hin, daß die Mittelwerte in den verschiedenen Gruppen nicht gleich sind. Abbildung 25.9, S. 610 zeigt eine Tabelle, die mit der Option *Univariate ANOVA* erstellt wurde.
- ***Box‘ M:*** Hiermit wird ein Test auf Gleichheit der Gruppen-Kovarianzmatrizen durchgeführt. Die Nullhypothese des Tests besagt, daß die Stichproben der verschiedenen Kovarianzmatrizen derselben Grundgesamtheit entstammen. Für den Test werden Box‘ M, das approximierte F, die Freiheitsgrade und die Signifikanz mitgeteilt.

Funktionskoeffizienten

- ***Fisher:*** Kreuzen Sie diese Option an, um die Koeffizienten von Fishers linearer Diskriminanzfunktion anzugeben. Diese Koeffizienten können Sie zur Klassifizierung verwenden.
- ***Nicht standardisiert:*** Diese Option gibt die nicht standardisierten Koeffizienten der Diskriminanzfunktionen aus. Eine entsprechende Tabelle ist in Abbildung 25.1, S. 595 wiedergegeben.

Matrizen

- ***Korrelationsmatrix innerhalb der Gruppen:*** Gibt in einer Matrix für die einzelnen Variablenkombinationen die gepoolten Korrelationen innerhalb der Gruppen an. Für jedes Variablenpaar wird dabei nur eine Korrelation ausgewiesen, die aus den Korrelationen innerhalb der einzelnen Gruppen errechnet wurde. Dies kann zu deutlich anderen Ergebnissen führen als die Korrelation, die für alle Fälle ohne eine Unterscheidung verschiedener Fallgruppen berechnet wird.
- ***Kovarianz innerhalb der Gruppen:*** Erstellt eine Matrix mit den Kovarianzen für die einzelnen Variablenpaare. Die angegebenen Kovarianzen werden aus den einzelnen Kovarianzen innerhalb der Gruppen gepoolt. Zusätzlich wird die Anzahl der Freiheitsgrade ausgewiesen.
- ***Kovarianz der einzelnen Gruppen:*** Für jede Gruppe wird eine eigene Matrix erstellt, die jeweils für die verschiedenen Paare der erklärenden Variablen die Kovarianzen angibt. Aus diesen Angaben werden die gepoolten Kovarianzen der vorhergehenden Option als gewichtetes Mittel errechnet. Als Gewichte dienen dabei die Anzahl der Fälle in den verschiedenen Gruppen.
- ***Gesamte Kovarianz:*** Diese Option berechnet die Kovarianzen der einzelnen Variablenpaare für die gesamte Stichprobe, also für alle in die Analyse einbezogenen Fälle. Zusätzlich wird die Anzahl der Freiheitsgrade angegeben.

25.5.5 Klassifizieren

In dem Dialogfeld aus Abbildung 25.20 können Sie die für die Analyse zu verwendenden A-priori-Wahrscheinlichkeiten festlegen und die Behandlung von fehlenden Werten steuern. Zudem können Sie weiteren Output wie zum Beispiel Diagramme anfordern. Sie öffnen das Dialogfeld mit der Schaltfläche *Klassifizieren*.

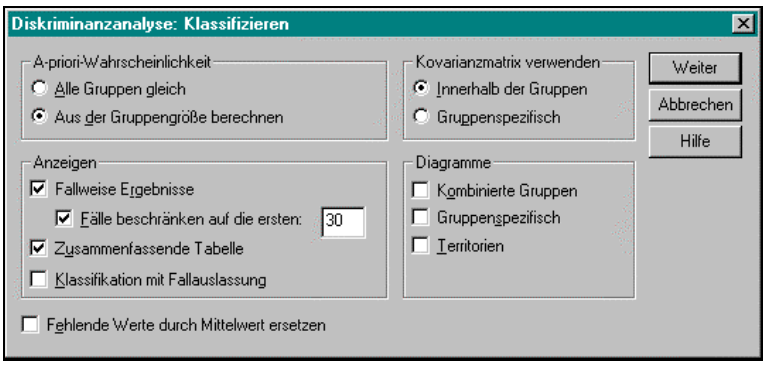


Abbildung 25.20: Dialogfeld der Schaltfläche „Klassifizieren“

A-priori-Wahrscheinlichkeit

- ***Alle Gruppen gleich:*** Diese Option ist voreingestellt, so daß für alle Gruppen die gleiche A-priori-Wahrscheinlichkeit angenommen wird. Bei der Unterscheidung zwischen vier Gruppen wird damit also für jede Gruppe eine Eintrittswahrscheinlichkeit von ¼ unterstellt.
- ***Aus der Gruppengröße berechnen:*** Wählen Sie diese Option, wenn Sie davon ausgehen, daß die Häufigkeiten der einzelnen Gruppen in der Stichprobe die A-priori-Wahrscheinlichkeiten widerspiegeln. Unterscheiden Sie zwei Gruppen, von der eine in der Stichprobe in 200 Fällen und die andere in 300 Fällen vertreten ist, wird für die Gruppe mit 200 Fällen eine Wahrscheinlichkeit von

$$\frac{200}{200+300} = 0{,}4 = 40\%$$

angenommen. Entsprechend beträgt die A-priori-Wahrscheinlichkeit der zweiten Gruppe 60%.

Kovarianzmatrix verwenden

In dieser Gruppe wählen Sie die Kovarianzmatrix, die zur Klassifizierung der Fälle verwendet werden soll:

- ***Innerhalb der Gruppen:*** Diese Option ist voreingestellt. Mit ihr werden die gepoolten Varianzen innerhalb der Gruppen zur Klassifizierung herangezogen.
- ***Gruppenspezifisch:*** Wählen Sie diese Option, um die Klassifizierung anhand der Kovarianzmatrizen der einzelnen Gruppen vorzunehmen.

Anzeigen

In dieser Gruppe können Sie zusätzlichen Output anfordern:

- ***Fallweise Ergebnisse:*** Diese Option erstellt eine Liste der einzelnen Fälle, für die jeweils die tatsächliche Gruppe sowie die durch die Analyse vorgenommene Gruppenzuordnung angegeben wird. Zusätzlich werden die Wahrscheinlichkeiten (A-posteriori-Wahrscheinlichkeit und bedingte Wahrscheinlichkeit) für die Gruppenzugehörigkeiten und der Wert der Diskriminanzfunktion angegeben. Eine entsprechende Tabelle ist in Abbildung 25.2, S. 598 dargestellt.
- ***Zusammenfassende Tabelle:*** Hiermit erstellen Sie eine Tabelle, in der die Häufigkeiten angegeben werden, mit denen die verschiedenen Kombinationen aus tatsächlicher Gruppenzugehörigkeit und Gruppenzuordnung vorkommen. Eine entsprechende Tabelle zeigt Abbildung 25.7, S. 607.
- ***Klassifikation mit Fallauslassung:*** Hiermit wird eine Klassifikation nach der *U-Methode* vorgenommen. Dabei wird jeder Fall anhand einer Funktion klassifiziert, die ausschließlich auf der Basis der übrigen Fälle, also unter Auslassung des jeweils betrachteten Falles, bestimmt wurde.

Diagramme

Mit Hilfe von Diagrammen können Sie die Funktionswerte den tatsächlichen oder den zugeordneten Gruppenzugehörigkeiten gegenüberstellen:

- ***Kombinierte Gruppen:*** Mit dieser Option wird nur dann ein Diagramm erstellt, wenn bei der Diskriminanzanalyse mindestens zwei Diskriminanzfunktionen verwendet werden.[285] Ist dies der Fall, wird ein Streudiagramm erzeugt, in dem auf den beiden Achsen die Funktionswerte der beiden ersten Diskriminanzfunktionen abgetragen werden. In dem Diagramm werden die einzelnen Fälle eingezeichnet und deren tatsächliche Gruppenzugehörigkeit ausgewiesen. Ein solches Diagramm ist in Abbildung 25.12, S. 617 dargestellt.
- ***Gruppenspezifisch:*** Mit dieser Option erstellen Sie eine ähnliche Grafik wie mit der Option *Kombinierte Gruppen*. Der Unterschied besteht lediglich darin, daß nicht alle Fälle in einer Grafik dargestellt werden, sondern für jede Gruppe ein eigenes Streudiagramm gezeichnet wird. Auch für die nicht gruppierten Fälle wird dabei eine Grafik erstellt.
- ***Territorien:*** Es wird eine aus Textzeichen zusammengesetzte Grafik in den Ausgabenavigator eingefügt, die für die beiden ersten Diskriminanzfunktionen die aus den Wertekombinationen beider Funktionen resultierenden Gruppenzuordnungen anzeigt. Diese Grafik kann nur erstellt werden, wenn mindestens zwei Diskriminanzfunktionen berechnet wurden. Eine Grafik dieser Art ist in Abbildung 25.13, S. 620 dargestellt.

Fehlende Werte durch Mittelwert ersetzen

Fälle, die in einer der in die Analyse einbezogenen Variablen (abhängige oder unabhängige Variable) einen fehlenden Wert aufweisen, werden bei der Bestimmung der Diskriminanzfunktionen ausgeschlossen. Die Berechnung der Funktionswerte und die anschließende Klassifizierung werden jedoch auch für solche Fälle vorgenommen, die in der abhängigen Variablen einen fehlenden Wert bzw. einen Wert außerhalb des angegebenen Wertebereichs aufweisen. Fälle mit einem fehlenden Wert in einer der erklärenden Variablen werden dagegen per Voreinstellung auch hierbei nicht berücksichtigt. Sie können diese Fälle mit fehlenden Werten in unabhängigen Variablen in die Klassifizierung mit einbeziehen, indem Sie die fehlenden Werte zur Berechnung der Funktionswerte durch den Mittelwert der jeweiligen erklärenden Variablen ersetzen. Kreuzen Sie hierzu die Option *Fehlende Werte durch Mittelwert ersetzen* an. Die fehlenden Werte werden dann nur bei der Berechnung der Gruppenzuordnung ersetzt, in der Datendatei bleiben sie dagegen unverändert.

[285] Bei früheren SPSS-Versionen wurde für den Fall, daß die Analyse nur eine Funktion verwendet, ebenfalls eine Grafik erzeugt, die sich aus einfachen Textzeichen zusammensetzte und in Form eines Histogramms die tatsächlichen Gruppenzugehörigkeiten den Funktionswerten gegenüberstellte.

25.5.6 Speichern von Ergebnissen in der Datendatei

Sie können in die Datendatei neue Variablen einfügen lassen, in denen für jeden Fall der Datei der Diskriminanzwert, die von der Analyse vorgenommene Gruppenzuordnung und die Wahrscheinlichkeit der Zugehörigkeit zu den verschiedenen Gruppen angegeben werden. Öffnen Sie hierzu mit der Schaltfläche *Speichern* das Dialogfeld aus Abbildung 25.21.

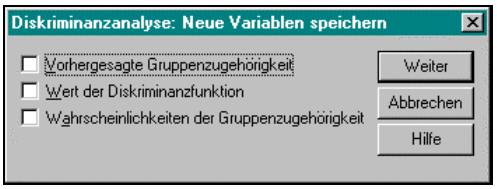


Abbildung 25.21: Dialogfeld der Schaltfläche „Speichern“

- ***Vorhergesagte Gruppenzugehörigkeit:*** Hiermit wird eine Variable erstellt, die die vorhergesagte Gruppenzugehörigkeit angibt. Der Variablen wird ein Label der Art *Vorhergesagte Gruppe aus Analyse 1* zugewiesen.
- ***Wert der Diskriminanzfunktion:*** Für jede Diskriminanzfunktion wird eine Variable erstellt, in der die Funktionswerte der einzelnen Fälle aufgeführt werden. Die Variablen erhalten Labels der Art *Werte der Diskriminanzfunktion 1 aus Analyse 1*.
- ***Wahrscheinlichkeiten der Gruppenzugehörigkeit:*** Für jede Gruppe der abhängigen Variablen wird eine Variable erstellt, die für die einzelnen Fälle die A-posteriori-Wahrscheinlichkeit der Zugehörigkeit zur jeweiligen Gruppe angibt. Die Variablen werden mit Labels der Art *Wahrscheinlichkeiten für die Mitgliedschaft in Gruppe 1 aus Analyse 1* versehen.

Die durch diese drei Optionen erstellten Variablen erhalten einen Namen, der sich aus dem Ausdruck *dis*, einer fortlaufenden Nummer, einem Unterstreichungszeichen und einer zweiten fortlaufenden Nummer zusammensetzt.