IBM SPSS Custom Tables 26

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 21 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 26, Release 0, Modifikation 0 von IBM® SPSS Statistics und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Inhaltsverzeichnis

Benutzerdefinierte Tabellen

Die folgenden Funktionen für benutzerdefinierte Tabellen sind in SPSS Statistics Standard Edition oder der Option "Custom Tables" enthalten.

Schnittstelle für benutzerdefinierte Tabellen

Schnittstelle des Tabellenerstellungsprogramms

Mit dem Befehl "Benutzerdefinierte Tabellen" öffnen Sie eine Schnittstelle des Tabellenerstellungsprogramms, die mit Ziehen und Ablegen bedient werden kann und in der Sie eine Vorschau der Tabelle erhalten, während Sie Variablen und Optionen auswählen. Außerdem finden Sie hier einen Grad an Flexibilität, den ein normales Dialogfeld nicht bietet, beispielsweise die Möglichkeit, die Größe des Fensters und die Größe der einzelnen Fensterbereiche zu ändern.

Erstellen von Tabellen

In der Schnittstelle für benutzerdefinierte Tabellen wählen Sie die Variablen und Auswertungsmaße aus, die in der Tabelle angezeigt werden sollen..

Analysieren > **Tabellen** > **Benutzerdefinierte Tabellen**

Variablenliste. Die Variablen in der Datendatei werden im linken Teilfenster des Dialogs angezeigt. Bei benutzerdefinierten Tabellen werden zwei Messniveaus unterschieden, nach denen Variablen behandelt werden:

Kategorial. Daten mit einer begrenzten Anzahl von eindeutigen Werten oder Kategorien (beispielsweise Geschlecht oder Religion). Kategoriale Variablen können Zeichenfolgevariablen (alphanumerisch) oder numerische Variablen sein, die zur Darstellung von Kategorien numerischen Code verwenden (beispielsweise 0 = *männlich* und 1 = *weiblich*). Auch als qualitative Daten bezeichnet. Kategoriale Variablen können **nominal** oder **ordinal** sein.

- *Nominal*. Eine Variable kann als nominal behandelt werden, wenn ihre Werte Kategorien darstellen, die sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- *Ordinal*. Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.

Kategoriale Variablen definieren Kategorien (Zeilen, Spalten, Schichten) in der Tabelle. In der Standardeinstellung wird als Auswertungsstatistik die Anzahl (Anzahl der Fälle pro Kategorie) berechnet. In einer Standardtabelle für eine kategoriale Variable "Geschlecht" wird beispielsweise einfach die Anzahl der Männer und die Anzahl der Frauen aufgeführt.

Skala. Daten, die auf einer Intervall- oder Verhältnisskala gemessen werden und bei denen die Datenwerte sowohl die Reihenfolge der Werte als auch die Distanz zwischen den Werten festlegen. So ist beispielsweise ein Gehalt von $72.195 höher als ein Gehalt von $52.398 und die Distanz zwischen den Werten beträgt $19.797. Auch als quantitative oder stetige Daten bezeichnet.

Metrische Variablen werden normalerweise in den Kategorien von kategorialen Variablen ausgewertet. In der Standardeinstellung wird als Auswertungsstatistik der Mittelwert berechnet. So wird beispielsweise in einer Standardtabelle, in der das Einkommen nach Geschlecht kategorisiert ist, das durchschnittliche Einkommen von Männern und das durchschnittliche Einkommen von Frauen aufgeführt.

Es ist darüber hinaus möglich, metrische Variablen auszuwerten, ohne mithilfe von kategorialen Variablen Gruppen zu definieren. Dies ist in erster Linie zum **Stapeln** der Auswertungen von mehreren metrischen Variablen sinnvoll.

Mehrfachantwortsets

In benutzerdefinierten Tabellen kann außerdem eine besondere Art von Variable verwendet werden, die als **Mehrfachantwortset** bezeichnet wird. Bei Mehrfachantwortsets handelt es sich nicht um Variablen im üblichen Sinn. Mehrfachantwortsets können nicht im Dateneditor angezeigt werden, sie werden von anderen Prozeduren nicht erkannt. Mehrfachantwortsets verwenden mehrere Variablen, um Antworten auf Fragen aufzuzeichnen, auf welche der Befragte mehr als eine Antwort geben kann. Sie werden wie kategoriale Variablen behandelt und bieten weitestgehend dieselben Möglichkeiten wie kategoriale Variablen.

Der Variablentyp ist durch ein Symbol neben der jeweiligen Variablen in der Variablenliste gekennzeichnet.

Kategorien. Wenn Sie in der Variablenliste eine kategoriale Variable auswählen, werden die für diese Variable definierten Kategorien im Teilfenster "Variableninformationen" angezeigt. Die Kategorien werden außerdem im Erstellungsbereich angezeigt, wenn Sie die Variable in einer Tabelle verwenden. Wenn für die Variable keine Kategorien definiert wurden, werden im Teilfenster "Variableninformationen" und im Erstellungsbereich zwei Kategorieplatzhalter angezeigt: *Kategorie 1* und *Kategorie 2*.

Die definierten Kategorien, die im Tabellenerstellungsprogramm angezeigt werden, beruhen auf **Wertbeschriftungen**. Hierbei handelt es sich um aussagekräftige Beschriftungen, die verschiedenen Datenwerten zugeordnet sind (z. B. die beiden numerischen Werte 0 und 1 mit den jeweiligen Wertbeschriftungen *männlich* und *weiblich*). Sie können Wertbeschriftungen im Teilfenster "Variableninformationen" des Dateneditors definieren.

Erstellungsbereich. Tabellen werden durch Ziehen und Ablegen von Variablen auf den Zeilen und Spalten im Erstellungsbereich erstellt. Im Erstellungsbereich wird eine Vorschau der Tabelle angezeigt, die erstellt wird. Die Zellen im Erstellungsbereich enthalten keine tatsächlichen Datenwerte. Die Darstellung im Erstellungsbereich bietet jedoch eine relativ genaue Layoutansicht der resultierenden Tabelle. Bei kategorialen Variablen enthält die tatsächliche Tabelle möglicherweise mehr Kategorien als die Tabelle in der Vorschau, wenn sich in der Datendatei eindeutige Werte befinden, für die keine Wertbeschriftungen definiert wurden.

Grundregeln und Einschränkungen für das Erstellen von Tabellen

- Bei kategorialen Variablen beruhen Auswertungsstatistiken auf der innersten Variablen in der Quellendimension der Statistik.
- Die Standardquellendimension der Statistik (Zeile oder Spalte) für kategoriale Variablen beruht auf der Reihenfolge, in der Sie die Variablen in den Erstellungsbereich ziehen. Wenn Sie z. B. eine Variable zuerst in den Bereich für die Zeilen ziehen, wird die Zeilendimension als Standardquellendimension für die Statistik verwendet.
- Metrische Variablen können nur innerhalb der Kategorien der innersten Variablen in der Zeilen- oder der Spaltendimension ausgewertet werden. (Die metrische Variable kann auf einer beliebigen Ebene der Tabelle platziert werden, ihre Auswertung erfolgt jedoch auf der innersten Ebene.)
- Metrische Variablen können nicht innerhalb von anderen metrischen Variablen ausgewertet werden. Sie können Auswertungen von mehreren metrischen Variablen stapeln und metrische Variablen innerhalb der Kategorien von kategorialen Variablen auswerten. Es ist nicht möglich, metrische Variablen ineinander zu verschachteln oder eine metrische Variable in der Zeilendimension und eine andere in der Spaltendimension anzuordnen.
- Wenn eine Variable im aktiven Dataset mehr als 12.000 definierte Wertbeschriftungen enthält, können Sie das Tabellenerstellungsprogramm nicht zum Erstellen von Tabellen verwenden. Wenn Sie keine Variablen, die diese Beschränkung überschreiten, in Ihre Tabellen aufnehmen müssen, können Sie Variab-

lensets definieren und anwenden, die diese Variablen ausschließen. Wenn Sie Variablen mit mehr als 12.000 definierten Wertbeschriftungen aufnehmen müssen, können Sie die betreffenden Tabellen mit der Befehlssyntax CTABLES generieren.

So erstellen Sie eine Tabelle

1. Wählen Sie in den Menüs Folgendes aus:
   Analysieren > **Tabellen** > **Benutzerdefinierte Tabellen**
2. Ziehen Sie eine oder mehrere Variablen auf die Bereiche für Zeilen bzw. Spalten im Erstellungsbereich.
3. Klicken Sie auf **Erstellen**, um die Tabelle zu erstellen.

So löschen Sie eine Variable aus dem Erstellungsbereich:

1. Wählen Sie eine Variable aus, indem Sie im Erstellungsbereich darauf klicken.
2. Klicken Sie mit der rechten Maustaste und wählen Sie **Variable löschen** im Dropdown-Menü aus.

Verschachteln von Variablen

Ähnlich wie bei Kreuztabellen kann mit Verschachtelung die Beziehung zwischen zwei kategorialen Variablen aufgezeigt werden. Bei der Verschachtelung wird eine der Variablen jedoch in derselben Dimension innerhalb der anderen Variablen verschachtelt. So können Sie beispielsweise *Geschlecht* in *Alterskategorie* in der Zeilendimension verschachteln, um die Anzahl der Männer und Frauen in jeder Alterskategorie darzustellen.

Es ist außerdem möglich, metrische Variablen in kategorialen Variablen zu verschachteln. So können Sie beispielsweise *Einkommen* in *Geschlecht* verschachteln, sodass das mittlere Einkommen (oder der Median bzw. ein anderes Auswertungsmaß des Einkommens) für Frauen und Männer getrennt aufgeführt wird.

So verschachteln Sie Variablen:

1. Ziehen Sie eine kategoriale Variable in den Erstellungsbereich und legen Sie sie im Bereich für die Zeilen oder die Spalten ab.
2. Ziehen Sie eine kategoriale oder eine metrische Variable in den Erstellungsbereich und legen Sie sie auf einer kategorialen Zeilen- oder Spaltenvariablen ab.
3. Wählen Sie **Oberhalb aller Variablen verschachteln**, **Links verschachteln** oder **Rechts verschachteln** im Menü aus.

Tabelle 1. Verschachtelte kategoriale Variablen

Variable 1	Variable 2	Auswertungsstatistik
Kategorie 1	Kategorie 1	12
	Kategorie 2	34
	Kategorie 3	56
Kategorie 2	Kategorie 1	12
	Kategorie 2	34
	Kategorie 3	56

Anmerkung: Bei benutzerdefinierten Tabellen wird die Verarbeitung von geschichteten aufgeteilten Dateien nicht berücksichtigt. Um dasselbe Ergebnis wie bei geschichteten aufgeteilten Dateien zu erzielen, müssen Sie die Dateiaufteilungsvariablen in den äußersten Verschachtelungsebenen der Datei platzieren.

Statistiken bearbeiten

Im Teilfenster "Statistiken bearbeiten" können Sie folgende Schritte ausführen:

- Hinzufügen und Entfernen von Auswertungsstatistiken zu bzw. aus einer Tabelle

Es hängt vom Messniveau der jeweiligen Statistikquellenvariablen ab, welche Statistiken (und andere Optionen) im Teilfenster "Statistiken bearbeiten" verfügbar sind. Die Quelle der Statistiken (die Variable, auf der die Statistiken beruhen) wird durch die folgenden Faktoren bestimmt:

- **Messniveau**. Wenn eine Tabelle (oder ein Tabellenabschnitt in einer gestapelten Tabelle) eine metrische Variable enthält, beruhen Statistiken auf der metrischen Variablen.
- **Reihenfolge der Variablenauswahl**. Die Standardquellendimension der Statistik (Zeile oder Spalte) für kategoriale Variablen beruht auf der Reihenfolge, in der Sie die Variablen in den Erstellungsbereich ziehen. Wenn Sie z. B. eine Variable zuerst in den Bereich für die Zeilen ziehen, wird die Zeilendimension als Standardquellendimension für die Statistik verwendet.
- **Verschachtelung**. Bei kategorialen Variablen beruhen Statistiken auf der innersten Variablen in der Quellendimension der Statistik.

Auswertungsstatistiken für kategoriale Variablen: Als grundlegende Statistiken für kategoriale Variablen sind Häufigkeiten und Prozentsätze verfügbar. Zusätzlich können angepasste Auswertungsstatistiken für Gesamtsummen und Zwischenergebnisse festgelegt werden. Diese angepassten Auswertungsstatistiken umfassen Lagemaße (wie Mittelwert und Median) und Streuungsmaße (wie Standardabweichung), die sich für einige ordinale kategoriale Variablen eignen.

Häufigkeiten. Anzahl der Fälle in jeder Zelle einer Tabelle bzw. Anzahl der Antworten bei Mehrfachantwortsets. Wenn die Gewichtung aktiviert ist, entspricht dieser Wert der gewichteten Anzahl.

- Wenn die Gewichtung aktiviert ist, entspricht der Wert der gewichteten Anzahl.
- Die gewichtete Anzahl ist bei der Gewichtung globaler Datasets (**Daten** > **Fälle gewichten...**).

Ungewichtete Anzahl. Ungewichtete Anzahl von Fällen in jeder Zelle der Tabelle. Diese unterscheidet sich von der Anzahl nur, wenn eine Gewichtung verwendet wird.

Gewichtete Anzahl. Die gewichtete Anzahl, die in Gewichtungsberechnungen für eine effektive Basis verwendet wird. Wenn Sie keine Gewichtungsvariable für eine effektive Basis verwenden, ist die gewichtete Anzahl mit der Anzahl identisch.

Zeilenprozente. Prozentsätze in jeder Zeile. Die Summe der Prozentsätze in jeder Zeile einer Untertabelle beträgt 100 % (bei einfachen Prozentsätzen). Zeilenprozente sind in der Regel nur dann sinnvoll, wenn eine kategoriale *Spaltenvariable* vorhanden ist.

Spaltenprozente. Prozentsätze in jeder Spalte. Die Summe der Prozentsätze in jeder Spalte einer Untertabelle beträgt 100 % (bei einfachen Prozentsätzen). Spaltenprozente sind in der Regel nur dann sinnvoll, wenn eine kategoriale *Zeilenvariable* vorhanden ist.

Untertabellenprozente. Die Prozentsätze für jede Zelle beziehen sich auf die Untertabelle. Alle Prozentsätze für Zellen in der Untertabelle beruhen auf derselben Gesamtanzahl von Fällen; die Summe der Prozentsätze über die gesamte Untertabelle beträgt 100 %. In verschachtelten Tabelle werden die Untertabellen durch die Variable definiert, die der innersten Verschachtelungsebene vorangeht. So werden beispielsweise die Untertabellen in einer Tabelle mit *Familienstand* innerhalb von *Geschlecht* innerhalb von *Alterskategorie* durch die Variable *Geschlecht* definiert.

Tabellenprozente. Die Prozentsätze für jede Zelle beziehen sich auf die gesamte Tabelle. Alle Prozentsätze für Zellen beruhen auf derselben Gesamtanzahl von Fällen; die Summe der Prozentsätze über die gesamte Tabelle beträgt 100 % (bei einfachen Prozentsätzen).

Konfidenzintervalle

- Untere und obere Konfidenzgrenzen sind für Häufigkeiten, Prozentsätze, Mittelwerte, Mediane, Perzentile und Summen verfügbar.
- Die Textzeichenfolge "&[Konfidenzniveau]" in der Beschriftung enthält das Konfidenzniveau in der Spaltenbeschriftung in der Tabelle.

- Der Standardfehler ist für Häufigkeiten, Prozentsätze, Mittelwerte und Summen verfügbar.
- Konfidenzintervalle und Standardfehler sind für Mehrfachantwortsets nicht verfügbar.

Niveau

Das Konfidenzniveau für Konfidenzintervalle, ausgedrückt in Prozent. Der Wert muss größer als 0 und kleiner als 100 sein.

Mehrfachantwortsets

Die Prozentsätze von Mehrfachantwortsets können sich auf Fälle, Antworten oder Häufigkeiten beziehen. Weitere Informationen finden Sie in „Auswertungsstatistiken für Mehrfachantwortsets".

Prozentbasis: Abhängig von der Behandlung fehlender Werten in der Berechnungsgrundlage können Prozentsätze auf drei verschiedenen Arten berechnet werden:

Einfacher Prozentsatz. Die Prozentsätze beziehen sich auf die Anzahl der Fälle, die in der Tabelle verwendet werden. Die Summe der Prozentsätze beträgt stets 100 %. Wenn eine Kategorie aus der Tabelle ausgeschlossen ist, werden die Fälle dieser Kategorie bei der Berechnung nicht berücksichtigt. Fälle mit systemdefiniert fehlenden Werten werden immer aus der Berechnungsgrundlage ausgeschlossen. Fälle mit benutzerdefiniert fehlenden Werten werden bei der Berechnung nicht berücksichtigt, wenn benutzerdefiniert fehlende Kategorien aus der Tabelle ausgeschlossen sind (Standardeinstellung). Sie werden berücksichtigt, wenn benutzerdefiniert fehlende Kategorien in der Tabelle enthalten sind. Alle Prozentsätze ohne *Gültige N* oder *Gesamtanzahl* in der Bezeichnung sind einfache Prozentsätze.

Prozentsätze auf Grundlage der Gesamtanzahl. Der Prozentbasis werden die Fälle mit systemdefiniert fehlenden und benutzerdefiniert fehlenden Werten hinzugefügt. Die Summe der Prozentsätze kann weniger als 100 % betragen.

Prozentsätze auf Grundlage der gültigen N. Fälle mit benutzerdefiniert fehlenden Werten werden aus der einfachen Prozentbasis entfernt, auch wenn benutzerdefiniert fehlende Kategorien in der Tabelle enthalten sind.

Anmerkung: Fälle in manuell ausgeschlossenen Kategorien, die keine benutzerdefiniert fehlenden Kategorien sind, werden bei der Berechnung nie berücksichtigt.

Auswertungsstatistiken für Mehrfachantwortsets: Für Mehrfachantwortsets sind die folgenden zusätzlichen Auswertungsstatistiken verfügbar.

Antworten als Spalten%/Zeilen%/Schichten%. Die Prozentsätze basieren auf Antworten.

Antworten als Spalten%/Zeilen%/Schichten% (Basis: Anzahl). Der Zähler enthält die Antworten, der Nenner die Gesamtanzahl.

Anzahl als Spalten%/Zeilen%/Schichten% (Basis: Antworten). Der Zähler enthält die Anzahl, der Nenner die Gesamtanzahl der Antworten.

Antworten als Spalten%/Zeilen% in Schicht. Prozentsatz für mehrere Untertabellen. Die Prozentsätze basieren auf Antworten.

Antworten als Spalten%/Zeilen% in Schicht (Basis: Anzahl). Prozentsätze für mehrere Untertabellen. Der Zähler enthält die Antworten, der Nenner die Gesamtanzahl.

Antworten als Spalten%/Zeilen% in Schicht (Basis: Antworten). Prozentsätze für mehrere Untertabellen. Der Zähler enthält die Anzahl, der Nenner die Gesamtanzahl der Antworten.

Antworten. Anzahl der Antworten.

Antworten als Untertabellen%/Tabellen%. Die Prozentsätze basieren auf Antworten.

Antworten als Untertabellen%/Tabellen% (Basis: Anzahl). Der Zähler enthält die Antworten, der Nenner die Gesamtanzahl.

Anzahl als Untertabellen%/Tabellen% (Basis: Antworten). Der Zähler enthält die Anzahl, der Nenner die Gesamtanzahl der Antworten.

Auswertungsstatistiken für metrische Variablen und angepasste Gesamtsummen für kategoriale Variablen: Zusätzlich zu den für kategoriale Variablen verfügbaren Häufigkeiten und Prozentsätzen sind für metrische Variablen die folgenden Auswertungsstatistiken verfügbar, die auch als angepasste Auswertungen für Gesamtsummen und Zwischenergebnisse bei kategorialen Variablen verwendet werden können. Diese Auswertungsstatistiken sind für Mehrfachantwortsets oder (alphanumerische) Zeichenfolgevariablen nicht verfügbar.

Mittelwert. Arithmetisches Mittel; die Summe geteilt durch die Anzahl der Fälle.

Median. Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil.

Modalwert. Häufigster Wert. Wenn zwei oder mehr Werte gleich häufig auftreten, wird der kleinste Wert angezeigt.

Minimum. Kleinster (niedrigster) Wert.

Maximum. Größter (höchster) Wert.

Fehlend. Anzahl der fehlenden Werte (sowohl benutzerdefiniert fehlende als auch systemdefiniert fehlende Werte).

Perzentil. Sie können das 5., 25., 75., 95. und/oder 99. Perzentil abrufen.

Bereich. Differenz zwischen größtem und kleinstem Wert.

Standardabweichung Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68 % der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95 % der Fälle innerhalb von zwei Standardabweichungen. Wenn beispielsweise das durchschnittliche Alter 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65 Jahren (Quadratwurzel der Varianz).

Summe. Die Summe der Werte.

Prozent der Summe. Prozentsätze bezogen auf Summen. Verfügbar für Zeilen und Spalten (von Untertabellen), ganze Zeilen und Spalten (über mehrere Untertabellen), Schichten, Untertabellen und ganze Tabellen.

Gesamtanzahl. Anzahl der nicht fehlenden, benutzerdefiniert fehlenden und systemdefiniert fehlenden Werte. Fälle in manuell ausgeschlossenen Kategorien, die nicht zu den benutzerdefiniert fehlenden Kategorien gehören, werden hierbei nicht berücksichtigt.

Korrigierte Gesamtzahl. Die korrigierte Gesamtzahl wird in Gewichtungsberechnungen für eine effektive Basis verwendet. Wenn Sie keine Gewichtungsvariable für eine effektive Basis (Registerkarte **Optionen**) verwenden, ist die korrigierte Gesamtzahl mit der Gesamtzahl identisch. Diese Statistik ist für Mehrfachantwortsets nicht verfügbar.

Gültige Anzahl. Anzahl der nicht fehlenden Werte. Fälle in manuell ausgeschlossenen Kategorien, die nicht zu den benutzerdefiniert fehlenden Kategorien gehören, werden hierbei nicht berücksichtigt.

Korrigierte gültige Anzahl. Die korrigierte gültige Anzahl wird in Gewichtungsberechnungen für eine effektive Basis verwendet. Wenn Sie keine Gewichtungsvariable für eine effektive Basis (Registerkarte **Optionen**) verwenden, ist die korrigierte gültige Anzahl mit der gültigen Anzahl identisch. Diese Statistik ist für Mehrfachantwortsets nicht verfügbar.

Varianz. Ein Maß der Streuung um den Mittelwert, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. Die Einheit der Varianz entspricht der quadrierten Einheit der eigentlichen Variablen (quadrierte Standardabweichung).

Konfidenzintervalle

- Untere und obere Konfidenzgrenzen sind für Häufigkeiten, Prozentsätze, Mittelwerte, Mediane, Perzentile und Summen verfügbar.
- Die Textzeichenfolge "&[Konfidenzniveau]" in der Beschriftung enthält das Konfidenzniveau in der Spaltenbeschriftung in der Tabelle.
- Der Standardfehler ist für Häufigkeiten, Prozentsätze, Mittelwerte und Summen verfügbar.
- Konfidenzintervalle und Standardfehler sind für Mehrfachantwortsets nicht verfügbar.

Niveau

Das Konfidenzniveau für Konfidenzintervalle, ausgedrückt in Prozent. Der Wert muss größer als 0 und kleiner als 100 sein.

Gestapelte Tabellen

Jeder von einer gestapelten Variablen definierte Tabellenabschnitt wird als eigene Tabelle behandelt und die Auswertungsstatistiken werden dementsprechend berechnet.

Kategorien und Gesamtsummen

Mit angepassten Tabellen können Sie folgende Funktionen ausführen:

- Umordnen von Kategorien.
- Einfügen von Gesamtsummen.
- Für Variablen ohne Wertbeschriftungen können nur Kategorien sortiert und Gesamtsummen eingefügt werden.

So greifen Sie auf die Kategorie- und Gesamtsummenoptionen zu

1. Ziehen Sie eine kategoriale Variable bzw. ein Mehrfachantwortset in den Erstellungsbereich und legen Sie diese bzw. dieses dort ab.
2. Klicken Sie mit der rechten Maustaste im Erstellungsbereich auf die Variable und wählen Sie im Popup-Menü eine der Kategorie- oder Gesamtsummenoptionen aus.

So sortieren Sie Kategorien

1. Klicken Sie mit der rechten Maustaste im Erstellungsbereich auf die Variable, wählen Sie im Popup-Menü die Option **Kategorien sortieren** und dann die Sortiermethode aus:
   - Nach Wert
   - Nach Beschriftung
   - Nach Anzahl
   - Nach niedrigen Werten

Gesamt

1. Klicken Sie mit der rechten Maustaste im Erstellungsbereich auf die Variable, wählen Sie im Popup-Menü die Option **Gesamtsumme anzeigen** und dann die Anzeigeposition der Gesamtsumme aus:
   - Über der Kategorie

- Unter der Kategorie

Wenn die ausgewählte Variable innerhalb einer anderen Variablen verschachtelt ist, werden Gesamtsummen für jede Untertabelle eingefügt.

Benutzerdefinierte Tabellen: "Teststatistiken"

Die Funktion **Teststatistiken** stellt Signifikanztests für benutzerdefinierte Tabellen bereit.

Diese Tests sind weder für Tabellen, bei denen Kategoriebeschriftungen aus den Standardtabellendimensionen verschoben wurden, noch für berechnete Kategorien verfügbar.

Spaltenmittel- und Spaltenanteiletests
: Spaltenmitteltests stehen für metrische Variablen zur Verfügung. Spaltenanteiletests stehen für kategoriale Variablen zur Verfügung.

  Spaltenmittel vergleichen
  : Paarweiser Test auf Gleichheit der Spaltenmittel. Die Tabelle muss eine kategoriale Variable in den Spalten enthalten und eine metrische Variable als innerste Ebene der Zeilen. Die Tabelle muss den Mittelwert als Auswertungsstatistik beinhalten.

    Die Varianz kann für reguläre kategoriale Variablen auf der Grundlage aller Kategorien oder auf der Grundlage der verglichenen Kategorien geschätzt werden. Für Mehrfachantwortvariablen wird die Varianz für den Spaltenmitteltest immer nur auf der Grundlage der verglichenen Kategorien geschätzt.

  Spaltenanteile vergleichen
  : Paarweiser Test auf Gleichheit der Spaltenanteile. Die Tabelle muss mindestens eine kategoriale Variable sowohl in den Spalten als auch in den Zeilen enthalten. Die Tabelle muss Häufigkeiten oder Spaltenprozente beinhalten.

Signifikanzniveau
: Das Signifikanzniveau für Spaltenmittel- und Spaltenanteiletests.

  - Der Wert muss größer als 0 und kleiner als 1 sein.
  - Wenn Sie zwei Signifikanzniveaus angeben, werden Großbuchstaben zur Identifizierung der Signifikanzwerte kleiner-gleich dem niedrigeren Niveau verwendet. Kleinbuchstaben werden zur Identifizierung der Signifikanzwerte kleiner-gleich dem höheren Niveau verwendet.
  - Wenn Sie **Tiefgestellte Zeichen im APA-Stil** auswählen, wird der zweite Wert ignoriert.

p-Werte für Mehrfachvergleiche anpassen
: Bei der **Bonferroni**-Korrektur werden Anpassungen für die Familywise Error Rate (FWER) vorgenommen. Die **Benjamini-Hochberg**-Methode ist eine Anpassung für die False Discovery Rate (FDR). Diese Methode ist weniger konservativ als die Bonferroni-Korrektur.

Signifikante Unterschiede identifizieren
: Für Spaltenmittel- und Spaltenanteiletests können Sie signifikante Ergebnisse in einer separaten Tabelle oder in der Haupttabelle anzeigen.

  In einer eigenen Tabelle
  : Ergebnisse von Signifikanztests werden in einer separaten Tabelle angezeigt. Falls zwei Werte deutlich voneinander abweichen, wird in der dem höheren Wert zugehörigen Zelle ein Schlüssel angezeigt, der auf die Spalte mit dem kleineren Wert verweist.

    Signifikanzwerte anzeigen
    : Die Signifikanzwerte werden nach jedem Schlüsselwert in der Zelle in Klammern angezeigt. Diese Option ist nur verfügbar, wenn signifikante Ergebnisse in einer separaten Tabelle angezeigt werden.

In der Haupttabelle

Ergebnisse von Signifikanztests werden in der Haupttabelle angezeigt. Jede Spaltenkategorie in der Tabelle wird durch einen alphabetischen Schlüssel gekennzeichnet. Für jedes signifikante Paar wird der Schlüssel der Kategorie mit dem kleineren Spaltenmittel oder Spaltenanteil in der Kategorie mit dem größeren Spaltenmittel oder Spaltenanteil angezeigt.

- Wenn Sie mit der Maus über einen Schlüssel in der Spaltenbeschriftungszelle in einer Pivot-Tabelle fahren, werden alle Zellen in der Tabelle mit diesem Signifikanzschlüssel hervorgehoben. Für eine Tabelle mit mehreren Variablen in der Spaltendimension werden nur die Zellen in dieser Untertabelle hervorgehoben.
- Wenn Sie alle Zellen in einer Tabelle (bzw. Untertabelle) auswählen wollen, die denselben Signifikanzschlüssel aufweisen, klicken Sie mit der rechten Maustaste auf die Spaltenbeschriftungszelle und wählen **Auswählen** > **Alle Zellen mit diesem Signifikanzschlüssel auswählen** aus.

Tiefgestellte Zeichen im APA-Stil

Signifikante Unterschiede werden durch eine Formatierung im APA-Stil angegeben, bei der tiefgestellte Buchstaben verwendet werden. Falls zwei Werte deutlich voneinander abweichen, weisen diese Werte unterschiedliche tiefgestellte Buchstaben auf. Diese tiefgestellten Buchstaben sind keine Fußnoten. Ist diese Option aktiviert, wird der in der aktuellen Tabellenvorlage definierte Fußnotenstil überschrieben und tatsächliche Fußnoten werden als hochgestellte Zahlen angezeigt. Wenn Sie alle Zellen in derselben Zeile auswählen wollen, die denselben Signifikanzschlüssel aufweisen, klicken Sie mit der rechten Maustaste auf eine Zelle, die einen Signifikanzschlüssel aufweist, und wählen **Zellen mit ähnlicher Signifikanz auswählen** aus.

Tests auf Unabhängigkeit (Chi-Quadrat)

Chi-Quadrat-Test auf Unabhängigkeit für Tabellen, in denen sowohl in den Zeilen als auch in den Spalten mindestens eine Kategorievariable vorhanden ist.

Zwischensummen anstelle von Kategorien für Zwischensummen verwenden

In Signifikanztests werden statt den Kategorien für Zwischenergebnisse die Zwischenergebnisse selbst verwendet. Andernfalls werden nur die für Zwischenergebnisse verwendeten ausgeblendeten Kategorien in den Tests durch die Zwischenergebnisse selbst ersetzt.

Mehrfachantwortvariablen in Tests einschließen

Kategorien von Mehrfachantwortsets werden in Signifikanztests eingeschlossen. Andernfalls werden Mehrfachantwortsets nicht in Signifikanztests eingeschlossen.

Beispieldateien

Die zusammen mit dem Produkt installierten Beispieldateien finden Sie im Unterverzeichnis *Samples* des Installationsverzeichnisses. Für jede der folgenden Sprachen gibt es einen eigenen Ordner innerhalb des Unterverzeichnisses "Samples": Deutsch, Englisch, Französisch, Italienisch, Japanisch, Koreanisch, Polnisch, Russisch, Spanisch, Traditionelles Chinesisch und Vereinfachtes Chinesisch.

Nicht alle Beispieldateien stehen in allen Sprachen zur Verfügung. Wenn eine Beispieldatei nicht in einer Sprache zur Verfügung steht, enthält der jeweilige Sprachordner eine englische Version der Beispieldatei.

Beschreibungen

Im Folgenden finden Sie Kurzbeschreibungen der in den verschiedenen Beispielen in der Dokumentation verwendeten Beispieldateien.

- **accidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die alters- und geschlechtsabhängige Risikofaktoren für Autounfälle in einer bestimmten Region untersucht. Jeder Fall entspricht einer Kreuzklassifikation von Alterskategorie und Geschlecht.
- **adl.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die Vorteile einer vorgeschlagenen Therapieform für Schlaganfallpatienten zu ermitteln. Ärzte teilten weibliche Schlaganfallpatienten nach dem Zufallsprinzip jeweils einer von zwei Gruppen zu. Die erste Gruppe erhielt die physische Standardtherapie, die zweite erhielt eine zusätzliche Emotionaltherapie. Drei Monate nach den Behandlungen wurden die Fähigkeiten der einzelnen Patienten, übliche Alltagsaktivitäten auszuführen, als ordinale Variablen bewertet.
- **advert.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Einzelhändlers geht, die Beziehungen zwischen den in Werbung investierten Beträgen und den daraus resultierenden Umsätzen zu untersuchen. Zu diesem Zweck hat er die Umsätze vergangener Jahre mit den zugehörigen Werbeausgaben zusammengestellt.
- **aflatoxin.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests von Maisernten auf Aflatoxin geht, ein Gift, dessen Konzentration stark zwischen und innerhalb von Ernteerträgen schwankt. Ein Getreideverarbeitungsbetrieb hat aus 8 Ernteerträgen je 16 Proben erhalten und das Aflatoxinniveau in Teilen pro Milliarde (PPB - Parts per Billion) gemessen.
- **anorectic.sav.** Bei der Ausarbeitung einer standardisierten Symptomatologie anorektischen/bulimischen Verhaltens führten Forscher[1] eine Studie mit 55 Jugendlichen mit bekannten Essstörungen durch. Jeder Patient wurde vier Mal über einen Zeitraum von vier Jahren untersucht, es fanden also insgesamt 220 Beobachtungen statt. Bei jeder Beobachtung erhielten die Patienten Scores für jedes von 16 Symptomen. Die Symptomscores fehlen für Patient 71 zum Zeitpunkt 2, Patient 76 zum Zeitpunkt 2 und Patient 47 zum Zeitpunkt 3, wodurch 217 gültige Beobachtungen verbleiben.
- **anticonvulsants.sav.** Wissenschaftler aus der Medizinforschung können ein verallgemeinertes lineares gemischtes Modell verwenden, um zu ermitteln, ob ein neues Antikonvulsivum die Häufigkeit epileptischer Anfälle bei einem Patienten verringern kann. Messwiederholungen bei ein und demselben Patienten sind in der Regel positiv korreliert. Daher sollte ein gemischtes Modell mit einigen Zufallseffekten angemessen sein. Für das Zielfeld (Anzahl der Anfälle) werden positive ganzzahlige Werte verwendet. Daher könnte ein verallgemeinertes lineares gemischtes Modell mit einer Poisson-Verteilung und einer Log-Verknüpfung geeignet sein.
- **bankloan.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Bank geht, den Anteil der nicht zurückgezahlten Kredite zu reduzieren. Die Datei enthält Informationen zum Finanzstatus und demografischen Hintergrund von 850 früheren und potenziellen Kunden. Bei den ersten 700 Fällen handelt es sich um Kunden, denen bereits ein Kredit gewährt wurde. Bei den letzten 150 Fällen handelt es sich um potenzielle Kunden, deren Kreditrisiko die Bank als gering oder hoch einstufen möchte.
- **bankloan_binning.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen zum Finanzstatus und demografischen Hintergrund von 5.000 früheren Kunden enthält.
- **bankloan_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Bank geht, die Merkmale von Kunden zu erkennen, die sehr wahrscheinlich mit einem Kredit in Verzug geraten, und diese Merkmale dann zum Erkennen niedriger und hoher Kreditrisiken zu verwenden.
- **bankloan_cs_noweights.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Bank geht, die Merkmale von Kunden zu erkennen, die sehr wahrscheinlich mit einem Kredit in Verzug geraten, und diese Merkmale dann zum Erkennen niedriger und hoher Kreditrisiken zu verwenden. Die Stichprobengewichtungen sind in der Datei nicht enthalten.

1. Van der Ham, T., J. J. Meulman, D. C. Van Strien, and H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry,* 170, 363-368.

- **behavior.sav.** In einem klassischen Beispiel[2] wurden 52 Schüler/Studenten gebeten, die Kombinationen aus 15 Situationen und 15 Verhaltensweisen auf einer 10-Punkte-Skala von 0 = "ausgesprochen angemessen" bis 9 = "ausgesprochen unangemessen" zu bewerten. Die Werte werden über die einzelnen Personen gemittelt und als Unähnlichkeiten verwendet.
- **behavior_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine zweidimensionale Lösung für *behavior.sav*.
- **brakes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik geht, die Scheibenbremsen für Hochleistungsfahrzeuge herstellt. Die Datendatei enthält Messungen des Durchmessers von 16 Scheiben aus 8 Produktionsmaschinen. Der Zieldurchmesser für die Scheiben ist 322 Millimeter.
- **breakfast.sav.** In einer klassischen Studie[3] wurden 21 MBA-Studenten der Wharton School mit ihren Lebensgefährten darum gebeten, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 = "am meisten bevorzugt" bis 15 = "am wenigsten bevorzugt" zu ordnen. Die Präferenzen wurden in sechs unterschiedlichen Szenarios erfasst, von "Overall preference" (Allgemein bevorzugt) bis "Snack, with beverage only" (Imbiss, nur mit Getränk).
- **breakfast-overall.sav.** Diese Datei enthält die Daten zu den bevorzugten Frühstücksartikeln, allerdings nur für das erste Szenario, "Overall preference" (Allgemein bevorzugt).
- **broadband_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die die Anzahl der Abonnenten eines Breitbandservice, nach Region geordnet, enthält. Die Datendatei enthält die monatlichen Abonnentenzahlen für 85 Regionen über einen Zeitraum von vier Jahren.
- **broadband_2.sav** Diese Datendatei stimmt mit *broadband_1.sav* überein, enthält jedoch Daten für weitere drei Monate.
- **cable_survey.sav.** Entscheidungsträger eines Kabelproviders für TV-, Telefon- und Internet-Services wollen mehr über potenzielle Kunden erfahren. Sie führen eine Umfrage bei 2000 Personen in ihren Serviceregionen durch und fragen für jeden der drei Services, ob die Personen (1) den Service nicht haben, (2) den Service über andere Provider beziehen oder (3) den Service dieses Unternehmens nutzen. Die Umfrage erfasst darüber hinaus einige demografische Informationen, wie Geschlecht, Alter (4 Stufen), Bildung (3 Stufen), Einkommen (3 Stufen), Wohnsitztyp (4 Stufen), Wohndauer an aktueller Adresse (3 Stufen), Anzahl der Personen im Haushalt usw.
- **car_insurance_claims.sav.** Ein an anderer Stelle[4] vorgestelltes und analysiertes Dataset bezieht sich auf Schadensansprüche für Autos. Die durchschnittliche Höhe der Schadensansprüche lässt sich mit Gammaverteilung modellieren. Dazu wird eine inverse Verknüpfungsfunktion verwendet, um den Mittelwert der abhängigen Variablen mit einer linearen Kombination aus Alter des Versicherungsnehmers, Fahrzeugtyp und Fahrzeugalter in Bezug zu setzen. Die Anzahl der eingereichten Schadensansprüche kann als Skalierungsgewichtung verwendet werden.
- **car_sales.sav.** Diese Datendatei enthält hypothetische Verkaufsschätzungen, Listenpreise und physische Spezifikationen für verschiedene Fahrzeugfabrikate und -modelle. Die Listenpreise und physischen Spezifikationen wurden von *edmunds.com* und Herstellerwebsites entnommen.
- **car_sales_uprepared.sav.** Hierbei handelt es sich um eine modifizierte Version der Datei *car_sales.sav*, die keinerlei transformierte Versionen der Felder enthält.
- **carpet.sav.** In einem bekannten Beispiel[5] möchte ein Unternehmen einen neuen Teppichreiniger vermarkten und dazu den Einfluss von fünf Faktoren auf die Präferenz durch den Verbraucher untersuchen: Verpackungsgestaltung, Markenname, Preis, Gütesiegel *Good Housekeeping* und Geld-zurück-Garantie. Die Verpackungsgestaltung setzt sich aus drei Faktorenebenen zusammen, die sich durch die Position der Auftragebürste unterscheiden. Außerdem gibt es drei Markennamen (*K2R*, *Glory* und *Bissell*), drei Preisstufen sowie je zwei Ebenen (Nein oder Ja) für die letzten beiden Faktoren. 10 Kunden stufen 22 Profile ein, die durch diese Faktoren definiert sind. Die Variable *Preference* enthält den Rang

2. Price, R. H., and D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology,* 30, 579-586.

3. Green, P. E., and V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

4. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

5. Green, P. E., and Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

der durchschnittlichen Einstufung für die verschiedenen Profile. Ein niedriger Rang bedeutet eine starke Präferenz. Diese Variable gibt ein Gesamtmaß der Präferenz für die Profile an.

- **carpet_prefs.sav.** Diese Datendatei beruht auf denselben Beispielen, wie für *carpet.sav* beschrieben, enthält jedoch die tatsächlichen Einstufungen durch jeden der 10 Kunden. Die Kunden wurden gebeten, die 22 Produktprofile in der Reihenfolge ihrer Präferenzen einzustufen. Die Variablen *PREF1* bis *PREF22* enthalten die IDs der zugeordneten Profile, wie in *carpet_plan.sav* definiert.
- **catalog.sav.** Diese Datendatei enthält hypothetische monatliche Verkaufszahlen für drei Produkte, die von einem Versandhaus verkauft werden. Daten für fünf mögliche Prädiktorvariablen wurden ebenfalls aufgenommen.
- **catalog_seasfac.sav.** Diese Datendatei ist mit *catalog.sav* identisch, außer, dass ein Set von saisonalen Faktoren, die mithilfe der Prozedur "Saisonale Zerlegung" berechnet wurden, sowie die zugehörigen Datumsvariablen hinzugefügt wurden.
- **cellular.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Mobiltelefonunternehmens geht, die Kundenabwanderung zu verringern. Propensity-Scores für die Abwanderungsneigung (von 0 bis 100) werden auf die Kunden angewendet. Kunden mit einem Score von 50 oder höher streben vermutlich einen Anbieterwechsel an.
- **ceramics.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Herstellers geht, der ermitteln möchte, ob ein neue, hochwertige Keramiklegierung eine größere Hitzebeständigkeit aufweist als eine Standardlegierung. Jeder Fall entspricht einem Test einer der Legierungen; die Temperatur, bei der das Keramikwälzlager versagte, wurde erfasst.
- **cereal.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Umfrage geht, bei der 880 Personen nach ihren Frühstückgewohnheiten befragt wurden. Außerdem wurden Alter, Geschlecht, Familienstand und Vorliegen bzw. Nichtvorliegen eines aktiven Lebensstils (auf der Grundlage von mindestens zwei Trainingseinheiten pro Woche) erfasst. Jeder Fall entspricht einem Befragten.
- **clothing_defects.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Bekleidungsfabrik geht. Aus jeder in der Fabrik produzierten Charge entnehmen die Kontrolleure eine Stichprobe an Bekleidungsartikeln und zählen die Anzahl der Bekleidungsartikel die inakzeptabel sind.
- **coffee.sav.** Diese Datendatei enthält Daten zum wahrgenommenen Image von sechs Eiskaffeemarken[6]. Bei den 23 Attributen des Eiskaffee-Image sollten die Teilnehmer jeweils alle Marken auswählen, die durch dieses Attribut beschrieben werden. Die sechs Marken werden als "AA", "BB", "CC", "DD", "EE" und "FF" bezeichnet, um Vertraulichkeit zu gewährleisten.
- **contacts.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Kontaktlisten einer Gruppe von Vertretern geht, die Computer an Unternehmen verkaufen. Die einzelnen Kontaktpersonen werden anhand der Abteilung, in der sie in ihrem Unternehmen arbeiten und anhand ihrer Stellung in der Unternehmenshierarchie in Kategorien eingeteilt. Außerdem werden der Betrag des letzten Verkaufs, die Zeit seit dem letzten Verkauf und die Größe des Unternehmens, in dem die Kontaktperson arbeitet, aufgezeichnet.
- **credit_card.sav.** Eine hypothetische Studie der Kreditkartenverwendung, die den monatlichen Einsatz der Hauptkreditkarte der Teilnehmer über zwei Jahre untersucht und die Ausgaben dabei nach dem Transaktionstyp (Lebensmittel, Einzelhandel, Unterhaltung, Reisen und Sonstiges) unterteilt. Jeder Datensatz im Dataset entspricht jeweils einem bestimmten Monat und Transaktionstyp. Die für die einzelnen Teilnehmer erfassten Daten erfordern also 2 Jahre × 12 Monate pro Jahr × 5 Typen von Transaktionen = 120 Datensätze.
- **creditpromo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Kaufhauses geht, die Wirksamkeit einer kürzlich durchgeführten Kreditkartenwerbeaktion einzuschätzen. Dazu wurden 500 Karteninhaber nach dem Zufallsprinzip ausgewählt. Die Hälfte erhielt eine Werbebeilage, die einen reduzierten Zinssatz für Einkäufe in den nächsten drei Monaten ankündigte. Die andere Hälfte erhielt eine Standardwerbebeilage.

6. Kennedy, R., C. Riquier, and B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing,* 5, 56-70.

- **cross_sell.sav.** Ein Onlineversand hat einen Buchclub und einen CD-Club. Jeden Monat gibt es Sonderaktionen für die Clubmitglieder. Das Unternehmen möchte ein Modell der pro Monat insgesamt aufgrund der Sonderaktionen getätigten Käufe auf der Basis der gesamten Buchverkäufe und CD-Verkäufe sowie dem Typ des Angebots erstellen, das den Clubmitgliedern gemacht wurde. Die 2SLS-Regression (Two-Stage Least-Squares Regression) ist für diese Situation geeignet, da Geld, das für die Sonderaktionen ausgegeben wird, nicht für Bücher oder CDs ausgegeben wird. Daher gibt es zwischen der Antwort und diesen beiden Prädiktoren eine Rückkopplungsschleife.
- **customer_dbase.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, das die Informationen in seinem Data Warehouse nutzen möchte, um spezielle Angebote für Kunden zu erstellen, die mit der größten Wahrscheinlichkeit darauf ansprechen. Nach dem Zufallsprinzip wurde ein Subset des Kundenstamms ausgewählt. Diese Kunden erhielten die speziellen Angebote und die Reaktionen wurden aufgezeichnet.
- **customer_information.sav.** Eine hypothetische Datendatei mit Kundenmailing-Daten wie Name und Adresse.
- **customer_subset.sav.** Ein Subset von 80 Fällen aus der Datei *customer_dbase.sav*.
- **debate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die paarige Antworten auf eine Umfrage unter den Zuhörern einer politischen Debatte enthält (Antworten vor und nach der Debatte). Jeder Fall entspricht einem Befragten.
- **debate_aggregate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der die Antworten aus *debate.sav* aggregiert wurden. Jeder Fall entspricht einer Kreuzklassifikation der favorisierten Politiker vor und nach der Debatte.
- **demo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Kundendatenbank geht, die zum Zwecke der Zusendung monatlicher Angebote erworben wurde. Neben verschiedenen demografischen Informationen ist erfasst, ob der Kunde auf das Angebot geantwortet hat.
- **demo_cs_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den ersten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einer anderen Stadt. Außerdem sind IDs für Region, Provinz, Landkreis und Stadt erfasst.
- **demo_cs_2.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den zweiten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einem anderen Stadtteil aus den im ersten Schritt ausgewählten Städten. Außerdem sind IDs für Region, Provinz, Landkreis, Stadt, Stadtteil und Wohneinheit erfasst. Die Informationen zur Stichprobenziehung aus den ersten beiden Stufen des Stichprobenplans sind ebenfalls enthalten.
- **demo_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfrageinformationen enthält, die mit einem komplexen Stichprobendesign erfasst wurden. Jeder Fall entspricht einer anderen Wohneinheit. Es sind verschiedene Informationen zum demografischen Hintergrund und zur Stichprobenziehung erfasst.
- **diabetes_costs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen enthält, die von einer Versicherungsgesellschaft für an Diabetes erkrankte Versicherungsnehmer verwaltet werden. Jeder Fall entspricht einem anderen Versicherungsnehmer.
- **dietstudy.sav.** Diese hypothetische Datendatei enthält die Ergebnisse einer Studie der "Stillman-Diät"[7]. Jeder Fall entspricht einem Teilnehmer und enthält dessen Gewicht vor und nach der Diät in amerikanischen Pfund sowie mehrere Messungen des Triglyceridspiegels (in mg/100 ml).
- **dmdata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen über Demografie und Einkäufe für ein Direktmarketing-Unternehmen enthält. *dmdata2.sav* enthält Informationen für ein Subset von Kontakten, die eine Testsendung erhalten, und *dmdata3.sav* enthält Informationen zu den verbleibenden Kontakten, die keine Testsendung erhalten.
- **dvdplayer.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Entwicklung eines neuen DVD-Players geht. Mithilfe eines Prototyps hat das Marketingteam Zielgruppendaten

7. Rickman, R., N. Mitchell, J. Dingman, and J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association,* 228:, 54-58.

erfasst. Jeder Fall entspricht einem befragten Benutzer und enthält demografische Daten zu dem Benutzer sowie dessen Antworten auf Fragen zum Prototyp.

- **Employee data.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die mitarbeiterspezifische Informationen enthält (Ausbildungsstufe, Anstellungsstatus, aktuelles Gehalt, Berufserfahrung usw.).
- **german_credit.sav.** Diese Daten sind aus dem Dataset "German credit" im Repository of Machine Learning Databases[8] an der Universität von Kalifornien in Irvine entnommen.
- **grocery_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *grocery_coupons.sav*, wobei die wöchentlichen Einkäufe zusammengefasst sind, sodass jeder Fall einem anderen Kunden entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und der verzeichnete ausgegebene Betrag ist nun die Summe der Beträge, die in den vier Wochen der Studie ausgegeben wurden.
- **grocery_coupons.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfragedaten enthält, die von einer Lebensmittelkette erfasst wurden, die sich für die Kaufgewohnheiten ihrer Kunden interessiert. Jeder Kunde wird über vier Wochen beobachtet und jeder Fall entspricht einer Kundenwoche und enthält Informationen zu den Geschäften, in denen der Kunde einkauft sowie zu anderen Merkmalen, beispielsweise welcher Betrag in der betreffenden Woche für Lebensmittel ausgegeben wurde.
- **guttman.sav.** Bell[9] legte eine Tabelle zur Darstellung möglicher sozialer Gruppen vor. Guttman[10] verwendete einen Teil dieser Tabelle, bei der fünf Variablen, die Aspekte beschreiben, wie soziale Interaktion, das Gefühl der Gruppenzugehörigkeit, die physische Nähe der Mitglieder und die Formalität der Beziehung, mit sieben theoretischen sozialen Gruppen gekreuzt wurden: "crowds" (Menschenmassen, beispielsweise die Zuschauer eines Fußballspiels), "audience" (Zuhörerschaften, beispielsweise die Personen im Theater oder bei einer Vorlesung), "public" (Öffentlichkeit, beispielsweise Zeitungsleser oder Fernsehzuschauer), "mobs" (Mobs, wie Menschenmassen, jedoch mit wesentlich stärkerer Interaktion), "primary groups" (Primärgruppen, vertraulich), "secondary groups" (Sekundärgruppen, freiwillig) und "modern community" (die moderne Gesellschaft, ein lockerer Zusammenschluss, der aus einer engen physischen Nähe und dem Bedarf an spezialisierten Dienstleistungen entsteht).
- **health_funding.sav.** Hierbei handelt es sich um eine hypothetische Datei, die Daten zur Finanzierung des Gesundheitswesens (Betrag pro 100 Personen), Krankheitsraten (Rate pro 10.000 Personen der Bevölkerung) und Besuche bei medizinischen Einrichtungen/Ärzten (Rate pro 10.000 Personen der Bevölkerung) enthält. Jeder Fall entspricht einer anderen Stadt.
- **hivassay.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Bemühungen eines pharmazeutischen Labors, einen Schnelltest zur Erkennung von HIV-Infektionen zu entwickeln. Die Ergebnisse des Tests sind acht kräftiger werdende Rotschattierungen, wobei kräftigeren Schattierungen auf eine höhere Infektionswahrscheinlichkeit hindeuten. Bei 2.000 Blutproben, von denen die Hälfte mit HIV infiziert war, wurde ein Labortest durchgeführt.
- **hourlywagedata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zum Stundenlohn von Pflegepersonal in Praxen und Krankenhäusern mit unterschiedlich langer Berufserfahrung.
- **insurance_claims.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen möchte. Jeder Fall entspricht einem Anspruch.
- **insure.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die die Risikofaktoren untersucht, die darauf hinweisen, ob ein Kunde die Leistungen einer mit einer Laufzeit von 10 Jahren abgeschlossenen Lebensversicherung in Anspruch neh-

8. Blake, C. L., and C. J. Merz. 1998. "UCI Repository of machine learning databases." Verfügbar unter http://www.ics.uci.edu/~mlearn/MLRepository.html.

9. Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

10. Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika,* 33, 469-506.

men wird. Jeder Fall in der Datendatei entspricht einem Paar von Verträgen, je einer mit Leistungsforderung und der andere ohne, wobei die beiden Versicherungsnehmer in Alter und Geschlecht übereinstimmen.

- **judges.sav.** Hierbei handelt es sich um eine hypothetische Datendatei mit den Wertungen von ausgebildeten Kampfrichtern (sowie eines Sportliebhabers) zu 300 Kunstturnleistungen. Jede Zeile stellt eine Leistung dar; die Kampfrichter bewerteten jeweils dieselben Leistungen.
- **kinship_dat.sav.** Rosenberg und Kim[11] haben 15 Bezeichnungen für den Verwandtschaftsgrad untersucht (Tante, Bruder, Cousin, Tochter, Vater, Enkelin, Großvater, Großmutter, Enkel, Mutter, Neffe, Nichte, Schwester, Sohn, Onkel). Die beiden Analytiker baten vier Gruppen von College-Studenten (zwei weibliche und zwei männliche Gruppen), diese Bezeichnungen auf der Grundlage der Ähnlichkeiten zu sortieren. Zwei Gruppen (eine weibliche und eine männliche Gruppe) wurden gebeten, die Bezeichnungen zweimal zu sortieren; die zweite Sortierung sollte dabei nach einem anderen Kriterium erfolgen als die erste. So wurden insgesamt sechs "Quellen" erzielt. Jede Quelle entspricht einer Ähnlichkeitsmatrix mit 15 x 15 Elementen. Die Anzahl der Zellen ist dabei gleich der Anzahl der Personen in einer Quelle minus der Anzahl der gemeinsamen Platzierungen der Objekte in dieser Quelle.
- **kinship_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine dreidimensionale Lösung für *kinship_dat.sav*.
- **kinship_var.sav.** Diese Datendatei enthält die unabhängigen Variablen *gender* (Geschlecht), *gener* (Generation) und *degree* (Verwandtschaftsgrad), die zur Interpretation der Dimensionen einer Lösung für *kinship_dat.sav* verwendet werden können. Insbesondere können sie verwendet werden, um den Lösungsraum auf eine lineare Kombination dieser Variablen zu beschränken.
- **marketvalues.sav.** Diese Datendatei betrifft Hausverkäufe in einem Neubaugebiet in Algonquin, Illinois, in den Jahren 1999–2000. Diese Verkäufe sind in Grundbucheinträgen dokumentiert.
- **nhis2000_subset.sav.** Die "National Health Interview Survey (NHIS)" ist eine große, bevölkerungsbezogene Umfrage unter der US-amerikanischen Zivilbevölkerung. Es werden persönliche Interviews in einer landesweit repräsentativen Stichprobe von Haushalten durchgeführt. Für die Mitglieder jedes Haushalts werden demografische Informationen und Beobachtungen zum Gesundheitsverhalten und Gesundheitsstatus eingeholt. Diese Datendatei enthält ein Subset der Informationen aus der Umfrage des Jahres 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Datendatei und Dokumentation öffentlich zugänglich. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Zugriff erfolgte 2003.
- **ozone.sav.** Die Daten enthalten 330 Beobachtungen zu sechs meteorologischen Variablen zur Vorhersage der Ozonkonzentration aus den übrigen Variablen. Bei früheren Untersuchungen[12, 13] fanden Wissenschaftler einige Nichtlinearitäten unter diesen Variablen, die die Standardverfahren bei der Regression behindern.
- **pain_medication.sav.** Diese hypothetische Datendatei enthält die Ergebnisse eines klinischen Tests für ein entzündungshemmendes Medikament zur Schmerzbehandlung bei chronischer Arthritis. Von besonderem Interesse ist die Zeitdauer, bis die Wirkung des Medikaments einsetzt und wie es im Vergleich mit bestehenden Medikamenten abschneidet.
- **patient_los.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen zu Patienten, die wegen des Verdachts auf Herzinfarkt in das Krankenhaus eingeliefert wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **patlos_sample.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen für eine Stichprobe von Patienten, denen während der Behandlung eines Herzinfarkts Thrombolytika verabreicht wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.

---

11. Rosenberg, S., and M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research,* 10, 489-502.

12. Breiman, L., and J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association,* 80, 580-598.

13. Hastie, T., and R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

- **poll_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die öffentliche Unterstützung für einen Gesetzentwurf zu ermitteln, bevor er im Parlament eingebracht wird. Die Fälle entsprechen registrierten Wählern. Für jeden Fall sind County, Gemeinde und Wohnviertel des Wählers erfasst.
- **poll_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *poll_cs.sav* aufgeführten Wähler. Die Stichprobe wurde gemäß dem in der Plandatei *poll.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Beachten Sie jedoch Folgendes: Da im Stichprobenplan die PPS-Methode (Probability Proportional To Size - Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*poll_jointprob.sav*). Die zusätzlichen Variablen zum demografischen Hintergrund der Wähler und ihrer Meinung zum vorgeschlagenen Gesetzentwurf wurden nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **property_assess.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen Bezirk (County) zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien, die im vergangenen Jahr in dem betreffenden County verkauft wurden. Jeder Fall in der Datendatei enthält die Gemeinde, in der sich die Immobilie befindet, den Bewerter, der die Immobilie besichtigt hat, die seit dieser Bewertung verstrichene Zeit, den zu diesem Zeitpunkt ermittelten Wert sowie den Verkaufswert der Immobilie.
- **property_assess_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen US-Bundesstaat zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien in dem betreffenden Bundesstaat. Jeder Fall in der Datendatei enthält das County, die Gemeinde und das Wohnviertel, in dem sich die Immobilie befindet, die seit der letzten Bewertung verstrichene Zeit sowie den zu diesem Zeitpunkt ermittelten Wert.
- **property_assess_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *property_assess_cs.sav* aufgeführten Immobilien. Die Stichprobe wurde gemäß dem in der Plandatei *property_assess.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Die zusätzliche Variable *Current value* (Aktueller Wert) wurde nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **recidivism.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Zeit bis zu seiner zweiten Festnahme, sofern diese innerhalb von zwei Jahren nach der ersten Festnahme erfolgte.
- **recidivism_cs_sample.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter, der im Juni 2003 erstmals aus der Haft entlassen wurde, und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Daten zu seiner zweiten Festnahme, sofern diese bis Ende Juni 2006 erfolgte. Die Straftäter wurden aus per Stichprobenziehung ermittelten Polizeidirektionen ausgewählt (gemäß dem in *recidivism_cs.csplan* angegebenen Stichprobenplan). Da hierbei eine PPS-Methode (Probability Proportional To Size - Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Eine hypothetische Datendatei mit Kauftransaktionsdaten wie Kaufdatum, gekauften Artikeln und Geldbetrag für jede Transaktion.
- **salesperformance.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Evaluierung von zwei neuen Verkaufsschulungen geht. 60 Mitarbeiter, die in drei Gruppen unterteilt sind, erhalten jeweils eine Standardschulung. Zusätzlich erhält Gruppe 2 eine technische Schulung und Gruppe 3 eine Praxisschulung. Die einzelnen Mitarbeiter wurden am Ende der Schulung einem Test unterzogen und der erzielte Score wurde erfasst. Jeder Fall in der Datendatei stellt einen Lehrgangsteilnehmer dar und enthält die Gruppe, der der Lehrgangsteilnehmer zugeteilt wurde sowie der von ihm in der Prüfung erreichte Score.

- **satisf.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Zufriedenheitsumfrage, die von einem Einzelhandelsunternehmen in 4 Filialen durchgeführt wurde. Insgesamt wurden 582 Kunden befragt. Jeder Fall gibt die Antworten eines einzelnen Kunden wieder.
- **screws.sav.** Diese Datendatei enthält Informationen zu den Eigenschaften von Schrauben, Bolzen, Muttern und Reißnägeln[14].
- **shampoo_ph.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik für Haarpflegeprodukte geht. In regelmäßigen Zeitintervallen werden Messwerte von sechs separaten Ausgangschargen erhoben und ihr pH-Wert erfasst. Der Zielbereich ist 4,5-5,5.
- **ships.sav.** Ein an anderer Stelle[15] vorgestelltes und analysiertes Dataset bezieht sich auf die durch Wellen verursachten Schäden an Frachtschiffen. Die Vorfallshäufigkeiten können unter Angabe von Schiffstyp, Konstruktionszeitraum und Betriebszeitraum gemäß einer Poisson-Rate modelliert werden. Das Aggregat der Betriebsmonate für jede Zelle der durch die Kreuzklassifizierung der Faktoren gebildeten Tabelle gibt die Werte für die Risikoanfälligkeit an.
- **site.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, neue Standorte für die betriebliche Expansion auszuwählen. Das Unternehmen beauftragte zwei Berater unabhängig voneinander mit der Bewertung der Standorte. Neben einem umfassenden Bericht gaben die Berater auch eine zusammenfassende Wertung für jeden Standort als "good" (gut) "fair" (mittelmäßig) oder "poor" (schlecht) ab.
- **smokers.sav.** Diese Datendatei wurde aus der Umfrage "National Household Survey of Drug Abuse" aus dem Jahr 1998 abstrahiert und stellt eine Wahrscheinlichkeitsstichprobe US-amerikanischer Haushalte dar. (http://dx.doi.org/10.3886/ICPSR02934) Daher sollte der erste Schritt bei der Analyse dieser Datendatei darin bestehen, die Daten entsprechend den Bevölkerungstrends zu gewichten.
- **stocks.sav.** Diese hypothetische Datendatei umfasst Börsenkurse und -volumina für ein Jahr.
- **stroke_clean.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozeduren in Statistics Base Edition bereinigt wurde.
- **stroke_invalid.sav.** Diese hypothetische Datendatei enthält den ursprünglichen Zustand einer medizinischen Datenbank, der mehrere Dateneingabefehler aufweist.
- **stroke_survival.** In dieser hypothetischen Datendatei geht es um die Überlebenszeiten von Patienten, die nach einem Rehabilitationsprogramm wegen eines ischämischen Schlaganfalls mit einer Reihe von Problemen zu kämpfen haben. Nach dem Schlaganfall werden das Auftreten von Herzinfarkt, ischämischem Schlaganfall und hämorrhagischem Schlaganfall sowie der Zeitpunkt des Ereignisses aufgezeichnet. Die Stichprobe ist auf der linken Seite abgeschnitten, da sie nur Patienten enthält, die bis zum Ende des Rehabilitationsprogramms, das nach dem Schlaganfall durchgeführt wurde, überlebten.
- **stroke_valid.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozedur "Daten validieren" überprüft wurde. Sie enthält immer noch potenziell anomale Fälle.
- **survey_sample.sav.** Diese Datendatei enthält Umfragedaten einschließlich demografischer Daten und verschiedener Meinungskennzahlen. Sie beruht auf einem Subset der Variablen aus der NORC General Social Survey aus dem Jahr 1998. Allerdings wurden zu Demonstrationszwecken einige Daten abgeändert und weitere fiktive Variablen hinzugefügt.
- **tcm_kpi.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die die Werte der wöchentlichen wesentlichen Leistungsindikatoren für ein Unternehmen enthält. Sie enthält darüber hinaus wöchentliche Daten für zahlreiche steuerbare Metriken über denselben Zeitraum.
- **tcm_kpi_upd.sav.** Diese Datendatei stimmt mit *tcm_kpi.sav* überein, enthält jedoch Daten für weitere vier Wochen.

14. Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

15. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

- **telco.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Telekommunikationsunternehmens geht, die Kundenabwanderung zu verringern. Jeder Fall entspricht einem Kunden und enthält verschiedene Informationen zum demografischen Hintergrund und zur Servicenutzung.
- **telco_extra.sav.** Diese Datendatei ähnelt der Datei *telco.sav*, allerdings wurden die Variablen "tenure" und die Log-transformierten Variablen zu den Kundenausgaben entfernt und durch standardisierte Log-transformierte Variablen ersetzt.
- **telco_missing.sav.** Diese Datendatei ist ein Subset der Datendatei *telco.sav*, allerdings wurde ein Teil der demografischen Datenwerte durch fehlende Werte ersetzt.
- **testmarket.sav.** Diese hypothetische Datendatei bezieht sich auf die Pläne einer Fast-Food-Kette, einen neuen Artikel in ihr Menü aufzunehmen. Es gibt drei mögliche Kampagnen zur Verkaufsförderung für das neue Produkt. Daher wird der neue Artikel in Filialen in mehreren zufällig ausgewählten Märkten eingeführt. An jedem Standort wird eine andere Form der Verkaufsförderung verwendet und die wöchentlichen Verkaufszahlen für das neue Produkt werden für die ersten vier Wochen aufgezeichnet. Jeder Fall entspricht einer Standortwoche.
- **testmarket_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *testmarket.sav*, wobei die wöchentlichen Verkaufszahlen zusammengefasst sind, sodass jeder Fall einem Standort entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren und die verzeichneten Verkaufszahlen sind nun die Summe der Verkaufszahlen während der vier Wochen der Studie.
- **tree_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_credit.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält.
- **tree_missing_data.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält und eine große Anzahl fehlender Werte aufweist.
- **tree_score_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_textdata.sav.** Eine einfache Datendatei mit nur zwei Variablen, die vor allem den Standardzustand von Variablen vor der Zuweisung von Messniveau und Wertbeschriftungen zeigen soll.
- **tv-survey.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Studie, die von einem Fernsehstudio durchgeführt wurde, das überlegt, ob die Laufzeit eines erfolgreichen Programms verlängert werden soll. 906 Personen wurden gefragt, ob sie das Programm unter verschiedenen Bedingungen ansehen würden. Jede Zeile entspricht einem Befragten; jede Spalte entspricht einer Bedingung.
- **ulcer_recurrence.sav.** Diese Datei enthält Teilinformationen aus einer Studie zum Vergleich der Wirksamkeit zweier Therapien zur Vermeidung des Wiederauftretens von Geschwüren. Es stellt ein gutes Beispiel für intervallzensierte Daten dar und wurde an anderer Stelle[16] vorgestellt und analysiert.
- **ulcer_recurrence_recoded.sav.** In dieser Datei sind die Daten aus *ulcer_recurrence.sav* so umstrukturiert, dass das Modell der Ereigniswahrscheinlichkeit für jedes Intervall der Studie berechnet werden kann und nicht nur die Ereigniswahrscheinlichkeit am Ende der Studie. Sie wurde an anderer Stelle[17] vorgestellt und analysiert.
- **verd1985.sav.** Diese Datendatei enthält eine Umfrage[18]. Die Antworten von 15 Subjekten auf 8 Variablen wurden aufgezeichnet. Die relevanten Variablen sind in drei Sets unterteilt. Set 1 umfasst *age* und *marital*, Set 2 besteht aus *pet* und *news* und in Set 3 finden sich *music* und *live*. Die Variable *pet* wird mehrfach nominal skaliert und die Variable *age* ordinal. Alle anderen Variablen werden einzeln nominal skaliert.

16. Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.
17. Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.
18. Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (in Dutch)*. Leiden: Department of Data Theory, University of Leiden.

- **virus.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Internet-Service-Providers geht, der die Auswirkungen eines Virus auf seine Netze ermitteln möchte. Dabei wurde vom Moment der Virusentdeckung bis zu dem Zeitpunkt, zu dem die Virusinfektion unter Kontrolle war, der (ungefähre) prozentuale Anteil infizierter E-Mail in den Netzen erfasst.
- **wheeze_steubenville.sav.** Hierbei handelt es sich um ein Subset der Daten aus einer Langzeitstudie zu den gesundheitlichen Auswirkungen der Luftverschmutzung auf Kinder[19]. Die Daten enthalten wiederholte binäre Messungen des Keuchens von Kindern aus Steubenville, Ohio, im Alter von 7, 8, 9 und 10 Jahren sowie eine unveränderliche Angabe, ob die Mutter im ersten Jahr der Studie rauchte oder nicht.
- **workprog.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einem Arbeitsprogramm der Regierung, das versucht, benachteiligten Personen bessere Arbeitsplätze zu verschaffen. Eine Stichprobe potenzieller Programmteilnehmer wurde beobachtet. Von diesen Personen wurden nach dem Zufallsprinzip einige für die Teilnahme an dem Programm ausgewählt. Jeder Fall entspricht einem Programmteilnehmer.
- **worldsales.sav.** Diese hypothetische Datendatei enthält Verkaufserlöse nach Kontinent und Produkt.

19. Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, and B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases,* 129, 366-374.

Bemerkungen

Die vorliegenden Informationen wurden für Produkte und Services entwickelt, die auf dem deutschen Markt angeboten werden. IBM stellt dieses Material möglicherweise auch in anderen Sprachen zur Verfügung. Für den Zugriff auf das Material in einer anderen Sprache kann eine Kopie des Produkts oder der Produktversion in der jeweiligen Sprache erforderlich sein.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Trotz sorgfältiger Bearbeitung können technische Ungenauigkeiten oder Druckfehler in dieser Veröffentlichung nicht ausgeschlossen werden. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Die angeführten Leistungsdaten und Kundenbeispiele dienen nur zur Illustration. Die tatsächlichen Ergebnisse beim Leistungsverhalten sind abhängig von der jeweiligen Konfiguration und den Betriebsbedingungen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

COPYRIGHTLIZENZ:

Diese Veröffentlichung enthält Beispielanwendungsprogramme, die in Quellensprache geschrieben sind und Programmiertechniken in verschiedenen Betriebsumgebungen veranschaulichen. Sie dürfen diese Beispielprogramme kostenlos kopieren, ändern und verteilen, wenn dies zu dem Zweck geschieht, Anwendungsprogramme zu entwickeln, zu verwenden, zu vermarkten oder zu verteilen, die mit der Anwendungsprogrammierschnittstelle für die Betriebsumgebung konform sind, für die diese Beispielprogramme geschrieben werden. Diese Beispiele wurden nicht unter allen denkbaren Bedingungen getestet. Daher kann IBM die Zuverlässigkeit, Wartungsfreundlichkeit oder Funktion dieser Programme weder zusagen noch gewährleisten. Die Beispielprogramme werden ohne Wartung (auf "as-is"-Basis) und ohne jegliche Gewährleistung zur Verfügung gestellt. IBM übernimmt keine Haftung für Schäden, die durch die Verwendung der Beispielprogramme entstehen.

Kopien oder Teile der Beispielprogramme bzw. daraus abgeleiteter Code müssen folgenden Copyrightvermerk beinhalten:

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corp in den USA und/oder anderen Ländern. Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite "Copyright and trademark information" unter www.ibm.com/legal/copytrade.shtml.

Adobe, das Adobe-Logo, PostScript und das PostScript-Logo sind Marken oder eingetragene Marken der Adobe Systems Incorporated in den USA und/oder anderen Ländern.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Marken oder eingetragene Marken der Intel Corporation oder ihrer Tochtergesellschaften in den USA oder anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA und/oder anderen Ländern.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.

Index

Gedruckt in Deutschland