IBM SPSS Custom Tables 26

Comunicado

Antes de usar estas informações e o produto suportado por elas, leia as informações nos “Avisos” na página 19.

Informações sobre o produto

Esta edição aplica-se à versão 26, liberação 0, modificação 0 do IBM® SPSS Statistics e a todas as liberações e modificações subsequentes até que seja indicado de outra forma em novas edições.

Índice

Tabelas customizadas

Os recursos de tabelas customizadas a seguir estão incluídos em SPSS Statistics Standard Edition ou a opção Tabelas customizadas.

Interface de tabelas customizadas

Interface do construtor de tabela

As Tabelas customizadas usam uma interface simples do construtor de tabela arrastar e soltar que permite visualizar sua tabela conforme você seleciona variáveis e opções. Também fornecem um nível de flexibilidade não localizado em uma caixa de diálogo típica, incluindo a capacidade de mudar o tamanho da janela e o tamanho das áreas de janela na janela.

Construindo tabelas

Selecione as variáveis e as medidas de sumarização que aparecerão em suas tabelas na interface de Tabelas customizadas.

Analisar > **Tabelas** > **Tabelas customizadas**

Lista de variáveis. As variáveis no arquivo de dados são exibidas na área de janela esquerda do diálogo. As Tabelas customizadas distinguem entre dois diferentes níveis de medição para variáveis e trata-as de forma diferente, dependendo do nível de medição:

Categórico. Dados com um número limitado de valores ou categorias distintas (por exemplo, sexo ou religião). Variáveis categóricas podem ser variáveis de sequência de caracteres (alfanuméricas) ou variáveis numéricas que usam códigos numéricos para representar categorias (por exemplo, 0 = *male* e 1 = *female*). Também referidos como dados qualitativos. Variáveis categóricas podem ser **nominais** ou **ordinais**

- *Nominal*. Uma variável pode ser tratada como nominal quando seus valores representarem categorias sem ranqueamento intrínseco (por exemplo, o departamento da empresa na qual um funcionário trabalha). Exemplos de variáveis nominais incluem região, código de endereçamento postal e filiação religiosa.
- *Ordinal*. Uma variável pode ser tratada como ordinal quando seus valores representarem categorias com algum ranqueamento intrínseco (por exemplo, níveis de satisfação de serviço de muito insatisfeito para muito satisfeito). Exemplos de variáveis ordinais incluem escores de atitude que representam o grau de satisfação ou de confiança e os escores de classificação de preferência.

As variáveis categóricas definem categorias (linha, colunas e camadas) na tabela e a estatística de sumarização padrão é a contagem (número de casos em cada categoria). Por exemplo, uma tabela padrão de variável de gênero categórico simplesmente exibiria o número de homens e o número de mulheres.

Escala. Dados medidos em um intervalo ou razão de escala, em que os valores de dados indicam a ordem dos valores e a distância entre os valores. Por exemplo, um salário de $72.195 é maior que um salário de $52.398, e a distância entre os dois valores é de $19.797. Também é conhecida como dados quantitativos ou contínuos.

As variáveis de escala geralmente são sumarizadas em categorias de variáveis categóricas e a estatística de sumarização padrão é a média. Por exemplo, uma tabela padrão de receita em categorias de gênero exibiria a receita média para homens e a receita média para mulheres.

Também é possível sumarizar variáveis de escala por elas mesmas, sem usar uma variável categórica para definir grupos. Isso é útil principalmente para **empilhar** sumarizações de diversas variáveis de escala.

Conjuntos de múltiplas respostas

As Tabelas customizadas também suportam um tipo especial de variável chamado **conjunto de múltiplas respostas**. Os conjuntos de múltiplas respostas não são realmente variáveis no sentido normal. Não é possível vê-los no Editor de dados e outros procedimentos não os reconhecem. Os conjuntos de múltiplas respostas usam variáveis múltiplas para registrar respostas a perguntas em que o respondente pode dar mais de uma resposta. Os conjuntos de múltiplas respostas são tratados como variáveis categóricas e a maioria das coisas que podem ser feitas com variáveis categóricas também podem ser feitas com conjuntos de múltiplas respostas.

Um ícone próximo a cada variável na lista de variáveis identifica o tipo de variável.

Categorias. Quando você seleciona uma variável categórica na lista de variáveis, as categorias definidas para a variável são exibidas na área de janela Informações informações variável. Essas categorias também serão exibidas na área de janela de tela quando a variável for usada em uma tabela. Se a variável não tiver categorias definidas, a área de janela de Informações de variável e a área de janela da tela exibirão duas categorias de item temporário: *Categoria 1* e *Categoria 2*.

As categorias definidas exibidas no construtor de tabela são baseadas em **rótulos de valor**, rótulos descritivos designados a diferentes valores de dados (por exemplo, valores numéricos de 0 e 1, com rótulos de valor de *homem* e *mulher*). É possível definir rótulos de valor na área de janela Informações de variável no Editor de dados.

Área de janela de tela. Construa uma tabela arrastando e soltando variáveis nas linhas e colunas da área de janela de tela. A área de janela de tela exibe uma visualização da tabela que será criada. A área de janela de tela não mostra valores de dados reais nas células, mas deve fornecer uma visualização bastante precisa do layout da tabela final. Para variáveis categóricas, a tabela real pode conter mais categorias que a visualização se o arquivo de dados contiver valores exclusivos para os quais nenhum rótulo de valor foi definido.

Regras básicas e limitações para construir uma tabela

- Para variáveis categóricas, as estatísticas básicas são baseadas na variável mais interna na dimensão de origem estatística.
- A dimensão de origem estatística padrão (linha ou coluna) para variáveis categóricas é baseada na ordem em que você arrasta e solta variáveis na área de janela de tela. Por exemplo, se você arrastar uma variável para a bandeja de linhas primeiro, a dimensão da linha será a dimensão de origem estatística padrão.
- As variáveis de escala podem ser sumarizadas apenas nas categorias da variável mais interna na dimensão da linha ou da coluna. (É possível posicionar a variável de escala em qualquer nível da tabela, mas ela é sumarizada no nível mais interno.)
- As variáveis de escala não podem ser sumarizadas em outras variáveis de escala. É possível empilhar sumarizações de variáveis de escala múltiplas ou sumarizar variáveis de escala em categorias de variáveis categóricas. Não é possível aninhar uma variável de escala em outra ou colocar uma variável de escala na dimensão da linha e outra variável de escala na dimensão da coluna.
- Se alguma variável no conjunto de dados ativo contiver mais de 12.000 rótulos de valor definidos, não será possível usar o construtor de tabela para criar tabelas. Se não precisar incluir variáveis que excedem essa limitação em suas tabelas, é possível definir e aplicar conjuntos de variáveis que excluem essas variáveis. Se precisar incluir variáveis com mais de 12.000 rótulos de valor definidos, é possível usar a sintaxe de comando `CTABLES` para gerar as tabelas.

Para construir uma tabela

1. Nos menus, escolha:

   Analisar > **Tabelas** > **Customizar tabelas**
2. Arraste e solte uma ou mais variáveis para as áreas de linha e/ou coluna da área de janela de tela.
3. Clique em **Criar** para criar a tabela.

Para excluir uma variável da área de janela da tela

1. Selecione (clique) uma variável na área de janela da tela.
2. Clique com o botão direito do mouse e selecione **Excluir variável** no menu suspenso.

Aninhando variáveis

O aninhamento, assim como a tabulação cruzada, pode mostrar o relacionamento entre duas variáveis categóricas, exceto que uma variável está aninhada na outra na mesma dimensão. Por exemplo, é possível aninhar *Gênero* em *Categoria de idade* na dimensão da linha, mostrando o número de homens e mulheres em cada categoria de idade.

Também é possível aninhar uma variável de escala em uma variável categórica. Por exemplo, é possível aninhar *Receita* em *Gênero*, mostrando valores médios (ou mediana ou outra medida de sumarização) de receita separados para homens e mulheres.

Para aninhar variáveis

1. Arraste e solte uma variável categórica na área de linha ou de coluna da área de janela de tela.
2. Arraste e solte uma variável categórica ou de escala na parte superior de uma variável categórica de linha ou coluna.
3. Selecione **Aninhar todas as variáveis acima**, **Aninhar à esquerda** ou **Aninhar à direita** no menu.

Tabela 1. Variáveis categóricas aninhadas

Variável 1	Variável 2	Estatística de resumo
Categoria 1	Categoria 1	12
	Categoria 2	34
	Categoria 3	56
Categoria 2	Categoria 1	12
	Categoria 2	34
	Categoria 3	56

Nota: As Tabelas customizadas não honram o processamento de arquivo dividido em camadas. Para obter o mesmo resultado como arquivos divididos em estratos, coloque as variáveis de arquivo dividido nas estratos de aninhamento mais externas da tabela.

Editar Estatísticas

A área de janela Editar estatísticas permite:

- Incluir e remover estatísticas básicas de uma tabela.

As estatísticas (e outras opções) disponíveis na área de janela Editar estatísticas dependem do nível de medição da variável de origem estatística. A origem de estatísticas (a variável na qual as estatísticas se baseiam) é determinada por:

- **Nível de medição**. Se uma tabela (ou uma seção da tabela em uma tabela empilhada) contiver uma variável de escala, as estatísticas serão baseadas na variável de escala.

- **Ordem de seleção de variáveis**. A dimensão de origem estatística padrão (linha ou coluna) para variáveis categóricas é baseada na ordem em que você arrasta e solta variáveis na área de janela de tela. Por exemplo, se você arrastar uma variável para a área de linhas primeiro, a dimensão da linha será a dimensão de origem estatística padrão.
- **Aninhamento**. Para variáveis categóricas, as estatísticas se baseiam na variável mais interna na dimensão de origem estatística.

Estatísticas básicas para variáveis categóricas: As estatísticas básicas disponíveis para variáveis categóricas são contagens e porcentagens. Também é possível especificar estatísticas básicas customizada para totais e subtotais. Essas estatísticas básicas customizada incluem medidas de tendência central (como média e mediana) e dispersão (como desvio padrão) que podem ser adequadas para algumas variáveis categóricas ordinais.

Contagem. Número de casos em cada célula da tabela ou número de respostas para conjuntos de múltiplas respostas. Se peso estiver em vigor, esse valor é a contagem ponderada.

- Se peso estiver em vigor, o valor é a contagem ponderada.
- A contagem ponderada é igual ao peso do conjunto de dados global (**Dados** > **Casos de ponderação...**).

Contagem não ponderada. Número não ponderado de casos em cada célula da tabela. Isso difere da contagem apenas se a ponderação estiver em vigor.

Contagem Ajustada. A contagem ajustada usada em cálculos de ponderação de base efetiva. Se você não usar uma variável de ponderação de base efetiva, a contagem ajustada será igual à contagem.

Porcentagens de linhas. Porcentagens em cada linha. As porcentagens em cada linha de uma subtabela (para porcentagens simples) somam 100%. As porcentagens de linhas geralmente serão úteis apenas se você tiver uma variável de *coluna* categórica.

Porcentagens da coluna. Porcentagens em cada coluna. As porcentagens em cada coluna de uma subtabela (para porcentagens simples) somam 100%. As porcentagens da coluna geralmente serão úteis apenas se você tiver uma variável de *linha* categórica.

Porcentagens da subtabela. As porcentagens em cada célula são baseadas na subtabela. Todas as porcentagens de células na subtabela são baseadas no mesmo número total de casos e somam 100% na subtabela. Em tabelas aninhadas, a variável que precede o nível de aninhamento mais interno define as subtabelas. Por exemplo, em uma tabela de *Estado civil* em *Gênero* na *Categoria de idade*, *Gênero* define subtabelas.

Porcentagens da tabela. As porcentagens para cada célula são baseadas na tabela inteira. Todas as porcentagens de células são baseadas no mesmo número total de casos e somam 100% (para porcentagens simples) na tabela inteira.

Intervalos de confiança

- Os limites de confiança inferior e superior estão disponíveis para contagens, porcentagens, média, mediana, percentis e soma.
- A sequência de caracteres de texto "&[Nível de confiança]" no rótulo inclui o nível de confiança no rótulo da coluna na tabela.
- O erro padrão está disponível para contagens, porcentagens, média e soma.
- Os intervalos de confiança e o erro padrão não estão disponíveis para conjuntos de múltiplas respostas.

Nível O nível de confiança para intervalos de confiança, expresso como uma porcentagem. O valor deve ser maior que 0 e menor que 100.

Conjuntos de múltiplas respostas

Os conjuntos de múltiplas respostas podem ter porcentagens baseadas em casos, respostas ou contagens. Consulte o tópico "Estatísticas básicas para conjuntos de múltiplas respostas" para obter mais informações

Base de porcentagem: As porcentagens podem ser calculadas de três maneiras diferentes, determinadas pelo tratamento de valores omissos na base computacional:

Porcentagem simples. As porcentagens são baseadas no número de casos usados na tabela e sempre somam 100%. Se uma categoria for excluída da tabela, os casos nessa categoria são excluídos da base. Casos com valores omissos do sistema são sempre excluídos da base. Os casos com valores omissos de usuário serão excluídos se as categorias com usuário desconhecido forem excluídas da tabela (o padrão) ou incluídos se as categorias com usuário desconhecido forem incluídas na tabela. Qualquer porcentagem que não tenha *N válido* ou *N total* em seu nome é uma porcentagem simples.

Porcentagem de N total. Os casos com valores omissos do sistema e omissos de usuário são incluídos na base de porcentagem Simples. As porcentagens podem somar menos de 100%.

Porcentagem de N válido. Os casos com valores omissos de usuário são removidos da base de porcentagem Simples, mesmo que as categorias com usuário desconhecido sejam incluídas na tabela.

Nota: Os casos em categorias excluídas manualmente diferentes de categorias sem usuário são sempre excluídos da base.

Estatísticas básicas para conjuntos de múltiplas respostas: As seguintes estatísticas básicas adicionais estão disponíveis para conjuntos de múltiplas respostas.

% de Respostas de Col/Linha/Camada. Porcentagem baseada em respostas.

% de Respostas de Col/Linha/Camada (Base: Contagem). As respostas são o numerador e a contagem total é o denominador.

% de Contagem de Col/Linha/Camada (Base: Respostas). Contagem é o numerador e total de respostas é o denominador.

% de Respostas de Col/Linha de estrato. Porcentagem em subtabelas. Porcentagem baseada em respostas.

% de Respostas de Col/Linha de estrato (Base: Contagem). Porcentagens em subtabelas. As respostas são o numerador e a contagem total é o denominador.

% de Respostas de Col/Linha de estrato (Base: Respostas). Porcentagens em subtabelas. Contagem é o numerador e total de respostas é o denominador.

Respostas. Contagem de respostas.

% de Respostas de Subtabela/Tabela. Porcentagem baseada em respostas.

% de Respostas de Subtabela/Tabela (Base: Contagem). As respostas são o numerador e a contagem total é o denominador.

% de Contagem de Subtabela/Tabela (Base: Respostas). Contagem é o numerador e total de respostas é o denominador.

Estatísticas básicas para variáveis de escala e totais customizados categóricos: Além das contagens e porcentagens disponíveis para variáveis categóricas, as seguintes estatísticas básicas estão disponíveis para variáveis de escala e como sumarizações de total e subtotal customizadas para variáveis categóricas. Essas estatísticas básicas não estão disponíveis para conjuntos de múltiplas respostas ou variáveis de sequência de caracteres (alfanuméricas).

Média. Média aritmética; a soma dividida pelo número de casos.

Mediana. Valor acima e abaixo do qual está a metade dos casos; o 50º percentil.

Modo. Valor mais frequente. Se houver um empate, o menor valor será mostrado.

Mínimo. Menor (mais baixo) valor.

Máxima. Maior (mais alto) valor.

Omisso. Contagem de valores omissos (omissos de usuário e do sistema).

Percentil. É possível incluir o 5º, 25º, 75º, 95º e/ou 99º percentis.

Intervalo. Diferença entre os valores máximo e mínimo.

Desvio padrão. Medida de dispersão em torno da média. Em uma distribuição normal, 68% dos casos estão contidos em um desvio padrão da média e 95% dos casos estão contidos em dois desvios padrão. Por exemplo, se a média de idade for 45, com um desvio padrão de 10, 95% dos casos estariam entre 25 e 65 em uma distribuição normal (a raiz quadrada da variância).

Soma. Soma dos valores.

Porcentagem da soma. Porcentagens baseadas em somas. Disponível para linhas e colunas (em subtabelas), linhas e colunas inteiras (entre subtabelas), estratos, subtabelas e tabelas inteiras.

N total. Contagem de valores não omissos, omissos de usuário e omissos do sistema. Não inclui os casos em categorias excluídas manualmente diferentes de categorias com usuário desconhecido.

N Total Ajustado. O N total ajustado usado em cálculos de ponderação de base efetiva. Se você não usar uma variável de ponderação de base efetiva (guia Opções), o N total ajustado é o mesmo que o N total. Essa estatística não está disponível para conjuntos de múltiplas respostas.

N válido. Contagem de valores não omissos. Não inclui os casos em categorias excluídas manualmente diferentes de categorias com usuário desconhecido.

N Válido Ajustado. O em válido ajustado usado em cálculos de ponderação de base efetiva. Se você não usar uma variável de ponderação de base efetiva (guia Opções), o N válido ajustado é o mesmo que o N válido. Essa estatística não está disponível para conjuntos de múltiplas respostas.

Variância. Uma medida de dispersão ao redor da média, igual à soma dos desvios quadrados da média dividido por um menor que o número de casos. A variância é medida em unidades que são o quadrado das unidades da própria variável (o quadrado do desvio padrão).

Intervalos de confiança

- Os limites de confiança inferior e superior estão disponíveis para contagens, porcentagens, média, mediana, percentis e soma.
- A sequência de caracteres de texto "&[Nível de confiança]" no rótulo inclui o nível de confiança no rótulo da coluna na tabela.

- O erro padrão está disponível para contagens, porcentagens, média e soma.
- Os intervalos de confiança e o erro padrão não estão disponíveis para conjuntos de múltiplas respostas.

Nível O nível de confiança para intervalos de confiança, expresso como uma porcentagem. O valor deve ser maior que 0 e menor que 100.

Tabelas empilhadas

Cada seção da tabela definida por uma variável de empilhamento é tratada como uma tabela separada, e as estatísticas básicas são calculadas de forma apropriada.

Categorias e totais

As Tabelas customizadas permitem:

- Reordenar categorias.
- Inserir totais.
- Para variáveis sem rótulos de valor definido, é possível apenas ordenar categorias e inserir totais.

Para acessar as opções de categorias e totais

1. Arraste e solte uma variável categórica ou conjunto de múltiplas respostas para a área de janela de tela.
2. Clique com o botão direito na variável na área de janela da tela e selecione uma das opções de categoria ou total no menu pop-up.

Para classificar categorias

1. Clique com o botão direito na variável na área de janela da tela, selecione **Classificar categorias** no menu pop-up e, em seguida, selecione o método de classificação:
   - Por valor
   - Por rótulo
   - Por contagem
   - Por inferior

Totais

1. Clique com o botão direito em uma variável na área de janela da tela, selecione **Mostrar total** no menu pop-up e, em seguida, selecione onde exibir o total:
   - Categoria acima
   - Categoria abaixo

Se a variável selecionada estiver aninhada dentro de outra variável, os totais serão inseridos para cada subtabela.

Tabelas customizadas: Estatísticas do teste

O recurso Estatísticas do teste fornece testes de significância para Tabelas customizadas.

Esses testes não estão disponíveis para tabelas nas quais os rótulos de categoria são movidos para fora da dimensão de tabela padrão ou para categorias calculadas.

Médias de coluna e testes de proporção da coluna
Os testes de médias de coluna estão disponíveis para variáveis de escala. Os testes proporções da coluna estão disponíveis para variáveis categóricas.

Comparar médias de colunas

Testes de pares da igualdade de médias de coluna. A tabela deve ter uma variável categórica nas colunas e uma variável de escala como o nível mais interno das linhas. A tabela deve incluir a média como uma estatística de sumarização.

Para variáveis categóricas ordinárias, a variância pode ser estimada a partir de todas as categorias ou apenas das categorias que são comparadas. Para variáveis de múltiplas respostas, a variância para o teste de médias é sempre baseada apenas nas categorias que são comparadas.

Comparar proporções de coluna

Testes de pares da igualdade de proporções da coluna. A tabela deve ter pelo menos uma variável categórica nas colunas e linhas. A tabela deve incluir contagens ou porcentagens da coluna.

Nível de Significância

O nível de significância para testes de médias de coluna e de proporções da coluna.

- O valor deve ser maior que 0 e menor que 1.
- Se você especificar dois níveis de significância, letras maiúsculas serão usadas para identificar valores de significância menores ou iguais ao menor nível. Letras minúsculas são usadas para identificar valores de significância menores ou iguais ao maior nível.
- Se você selecionar **Usar subscritos de estilo de APA**, o segundo valor será ignorado.

Ajustar valores p para comparações múltiplas

A correção de **Bonferroni** é ajustada para a taxa de erro da família (FWER). O método de **Benjamini-Hochberg** é um ajustamento de taxa de descoberta falsa (FDR). Esse método é menos conservador que a correção de Bonferroni.

Identificar diferenças significativas

Para testes de médias de coluna e de proporções da coluna, é possível exibir resultados significativos em uma tabela separada ou na tabela principal.

Em uma tabela separada

Os resultados de testes de significância são exibidos em uma tabela separada. Se dois valores forem significativamente diferentes, a célula correspondente ao valor maior exibirá uma chave que identifica a coluna com o valor menor.

Exibir valores de significância

Os valores de significância são exibidos entre parênteses após cada valor da chave na célula. Esta opção está disponível apenas quando os resultados significativos são exibidos em uma tabela separada.

Na tabela principal

Os resultados do teste de significância são exibidos na tabela principal. Cada categoria de coluna na tabela é identificada com uma chave alfabética. Para cada par significativo, a chave da categoria com a menor média ou proporção da coluna aparece na categoria com a maior média ou proporção da coluna.

- Quando você passar o mouse sobre uma chave na célula de rótulo da coluna em uma tabela dinâmica, todas as células na tabela com aquela chave de significância serão destacadas. Para uma tabela com variáveis múltiplas na dimensão da coluna, somente células naquela subtabela serão destacadas.
- Para selecionar todas as células em uma tabela (ou subtabela) que têm a mesma chave de significância, clique com o botão direito na célula de rótulo da coluna e escolha **Selecionar** > **Selecionar todas as células com essa chave de significância**.

Usar subscritos de estilo APA

Identifique diferenças significativas com formatação de estilo APA que usa letras subscritas. Se dois valores forem significativamente diferentes, esses valores exibirão diferentes letras subscritas. Esses subscritos não são notas de rodapé.

Quando essa opção estiver em vigor, o estilo de nota de rodapé definido no TableLook atual será substituído e as notas de rodapé serão exibidas como números sobrescritos. Para selecionar todas as células na mesma linha com a mesma chave de significância, clique com o botão direito em uma célula que tem uma chave de significância e escolha **Selecionar células com significância semelhante**

Testes de independência (qui-quadrado)
Teste qui-quadrado de independência para tabelas nas quais existe pelo menos uma variável de categoria nas linhas e colunas.

Usar subtotais em vez de categorias subtotalizadas
Cada subtotal substitui suas categorias para o teste de significância. Caso contrário, apenas os subtotais para os quais os subtotais das categorias estão ocultos substituirão suas categorias para teste.

Incluir variáveis de múltiplas respostas nos testes
As categorias de conjuntos de múltiplas respostas são incluídas em testes de significância. Caso contrário, os conjuntos de múltiplas respostas não são incluídos em testes de significância.

Arquivos de Amostra

Os arquivos de amostra instalados com o produto podem ser localizados no subdiretório *Amostras* do diretório de instalação. Há uma pasta separada dentro do subdiretório Amostras para cada um dos seguintes idiomas: inglês, francês, alemão, italiano, japonês, coreano, polonês, russo, chinês simplificado, espanhol e chinês tradicional.

Nem todos os arquivos de amostra estão disponíveis em todos os idiomas. Se um arquivo de amostra não estiver disponível em um idioma, essa pasta de idiomas conterá uma versão em inglês do arquivo de amostra.

PDV

A seguir, encontram-se breves descrições dos arquivos de amostra usadas em vários exemplos em toda a documentação.

- **accidents.sav.** Esse é um arquivo de dados hipotéticos que diz respeito a uma empresa de seguros que está estudando fatores de risco de idade e sexo para acidentes de automóveis em uma determinada região. Cada caso corresponde a uma classificação cruzada de categoria de idade e sexo.
- **adl.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços para determinar os benefícios de um tipo proposto de terapia para pacientes que sofreram acidente vascular cerebral. Os médicos designaram aleatoriamente pacientes do sexo feminino que sofreram acidente vascular cerebral a um dos dois grupos. O primeiro recebeu a terapia física padrão e o segundo recebeu uma terapia emocional adicional. Três meses seguindo os tratamentos, as habilidades de cada paciente em realizar atividades comuns da vida diária foram pontuadas como variáveis ordinais.
- **advert.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços do varejista em examinar o relacionamento entre o dinheiro gasto em publicidade e as vendas resultantes. Para essa finalidade, eles coletaram dados de vendas passadas e os custos associados de publicidade.
- **aflatoxin.sav.** Esse é um arquivo de dados hipotéticos que diz respeito ao teste de safras de milho para aflatoxina, um veneno cuja concentração varia muito entre e dentro das produções da safra. Um processador de grãos recebeu 16 amostras de cada uma das 8 produções da safra e mediu os níveis de aflatoxina em partes por bilhão (PPB).
- **anorectic.sav.** Enquanto trabalhando em direção a uma sintomatologia padronizada do comportamento anorético/bulímico, os pesquisadores [1] fizeram um estudo de 55 adolescentes com transtornos

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

alimentares conhecidos. Cada paciente foi visto quatro vezes durante quatro anos, para um total de 220 observações. Em cada observação, os pacientes foram pontuados para cada um dos 16 sintomas. As pontuações dos sintomas estão faltando para o paciente 71 no tempo 2, paciente 76 no tempo 2 e paciente 47 no tempo 3, deixando 217 observações válidas.

- **anticonvulsants.sav.** Os pesquisadores médicos podem usar um modelo linear generalizado misto para determinar se uma nova droga anticonvulsiva pode reduzir a taxa de crises epilépticas de um paciente. Medidas repetidas do mesmo paciente são tipicamente correlacionadas positivamente; portanto, um modelo misto com alguns efeitos aleatórios deve ser apropriado. O campo de destino, o número de convulsões, assume valores de número inteiro positivo, de modo que um modelo linear generalizado misto com distribuição de Poisson e ligação de log pode ser apropriado.
- **bankloan.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de um banco em reduzir a taxa de padrões de empréstimo. O arquivo contém informações financeiras e demográficas sobre 850 clientes passados e potenciais. Os primeiros 700 casos são clientes que anteriormente receberam empréstimos. Os últimos 150 casos são clientes potenciais que o banco precisa classificar os riscos de crédito como bons ou ruins.
- **bankloan_binning.sav.** Esse é um arquivo de dados hipotéticos que contém informações financeiras e demográficas sobre 5.000 clientes antigos.
- **bankloan_cs.sav.** Esse é um arquivo de dados hipotéticos referente aos esforços de um banco para identificar características indicativas de pessoas propensas a serem inadimplentes em empréstimos e, em seguida, usar essas características para identificar riscos de crédito bons e ruins.
- **bankloan_cs_noweights.sav.** Esse é um arquivo de dados hipotéticos referente aos esforços de um banco para identificar características indicativas de pessoas propensas a serem inadimplentes em empréstimos e, em seguida, usar essas características para identificar riscos de crédito bons e ruins. As ponderações de amostragem não são incluídas no arquivo.
- **behavior.sav.** Em um exemplo clássico [2], foi solicitado que 52 estudantes classificassem as combinações de 15 situações e 15 comportamentos em uma escala de 10 pontos variando de 0="extremamente adequado" a 9="extremamente inadequado." Na média de indivíduos, os valores são assumidos como dissimilaridades.
- **behavior_ini.sav.** Este arquivo de dados contém uma configuração inicial para uma solução bidimensional para *behavior.sav*.
- **brakes.sav.** Esse é um arquivo de dados hipotéticos que diz respeito ao controle de qualidade em uma fábrica que produz freios de disco para automóveis de alto desempenho. O arquivo de dados contém medidas de diâmetro de 16 discos de cada uma das 8 máquinas de produção. O diâmetro de destino para os freios é de 322 milímetros.
- **breakfast.sav.** Em um estudo clássico [3], foi solicitado que 21 estudantes de MBA da Wharton School e seus cônjuges classificassem 15 itens de café da manhã em ordem de preferência com 1="mais preferencial" a 15="menos preferencial". Suas preferências foram registradas em seis cenários diferentes, de "Preferência geral" a "Petiscos, apenas com bebidas".
- **breakfast-overall.sav.** Este arquivo de dados contém as preferências dos itens de café da manhã para o primeiro cenário, "Preferência geral", apenas.
- **broadband_1.sav.** Esse é um arquivo de dados hipotéticos que contém o número de assinantes, por região, de um serviço nacional de banda larga. O arquivo de dados contém os números de assinantes mensais para 85 regiões durante um período de quatro anos.
- **broadband_2.sav.** Este arquivos de dados é idêntico para *broadband_1.sav*, mas contém dados para três meses adicionais.
- **cable_survey.sav.** Executivos em um fornecedor de cabo de serviços de televisão, telefone e Internet querem saber mais sobre clientes potenciais. Eles conduzem uma pesquisa de opinião de 2.000 pessoas em suas regiões de serviço e perguntam se elas (1) não têm o serviço; (2) assinam o serviço com outros

2. Price, R. H., e 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., e V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

provedores; ou (3) têm o serviço com a empresa, para cada um dos três serviços. A pesquisa também coleta algumas informações demográficas, como sexo, categoria de idade (4 níveis), categoria de educação (3 níveis), categoria de renda (3 níveis), categoria de tipo de residência (4 níveis), anos na categoria de endereço atual (3 níveis), número de pessoas na casa, e assim por diante.

- **car_insurance_claims.sav.** Um conjunto de dados apresentado e analisado em outro lugar [4] diz respeito a reclamações por danos nos carros. A quantia média de reclamações pode ser modelada como tendo uma distribuição gama, usando uma função de ligação inversa para relacionar a média da variável dependente a uma combinação linear da idade do beneficiário do seguro, do tipo de veículo e da idade do veículo. O número de reclamações arquivadas pode ser usado como um peso de ajuste de escala.
- **car_sales.sav.** Este arquivo de dados contém estimativas de vendas hipotéticas, preços de lista e especificações físicas para várias marcas e modelos de veículos. Os preços de listas e as especificações físicas foram obtidos alternadamente de *edmunds.com* e de sites do fabricante.
- **car_sales_uprepared.sav.** Esta é uma versão modificada de *car_sales.sav* que não inclui nenhuma versão transformada dos campos.
- **carpet.sav.** Em um exemplo popular [5], uma empresa interessada em fazer propaganda de um novo limpador de carpete quer examinar a influência de cinco fatores de preferência do consumidor—design de pacote, nome de marca, preço, um selo *Good Housekeeping* e garantia de retorno financeiro. Há três níveis de fatores para design de pacote, cada um diferindo no local da escova do aplicador; três nomes de marcas (*K2R*, *Glory* e *Bissell*); três níveis de preço e dois níveis (não ou sim) para cada um dos dois últimos fatores. Dez consumidores classificam 22 perfis definidos por esses fatores. A variável *Preferência* contém a classificação média para cada perfil. Classificações baixas correspondem à preferência alta. Esta variável reflete uma medida geral de preferência para cada perfil.
- **carpet_prefs.sav.** Este arquivo de dados é baseado no mesmo exemplo descrito para *carpet.sav*, mas contém as classificações reais coletadas de cada um dos 10 consumidores. Foi solicitado que os consumidores classificassem os 22 perfis de produto do mais preferencial ao menos preferencial. As variáveis *PREF1* a *PREF22* contêm os identificadores dos perfis associados, conforme definido em *carpet_plan.sav*.
- **catalog.sav.** Este arquivo de dados contém dados de vendas mensais hipotéticos para três produtos vendidos por uma empresa de catálogo. Dados para cinco variáveis preditoras possíveis também estão incluídos.
- **catalog_seasfac.sav.** Este arquivo de dados é o mesmo que *catalog.sav*, exceto pela adição de um conjunto de fatores sazonais calculados a partir do procedimento de Decomposição Sazonal juntamente com as variáveis de data que acompanham.
- **cellular.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de uma empresa de telefonia celular para reduzir a perda de clientes. Os escores de propensão à perda de clientes são aplicados às contas, variando de 0 a 100. A pontuação de contas de 50 ou acima pode estar buscando a mudança dos provedores.
- **ceramics.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de um fabricante em determinar se uma nova liga premium tem uma resistência maior ao calor do que uma liga padrão. Cada caso representa um teste separado de uma das ligas; o calor no qual o rolamento falhou é registrado.
- **cereal.sav.** Esse é um arquivo de dados hipotéticos que diz respeito a uma pesquisa com 880 pessoas sobre suas preferências no café da manhã, observando também sua idade, sexo, estado civil e se elas têm ou não um estilo de vida ativo (com base em se elas se exercitam pelo menos duas vezes por semana). Cada caso representa um entrevistado separado.
- **clothing_defects.sav.** Esse é um arquivo de dados hipotéticos que diz respeito ao processo de controle de qualidade em uma fábrica de vestuário. De cada lote produzido na fábrica, os inspetores pegam uma amostra de roupas e contam o número de peças que não são aceitáveis.

---

4. McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2ª ed. Londres: Chapman & Hall.

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

- **coffee.sav.** Este arquivo de dados pertence às imagens percebidas de seis marcas de café gelado [6] . Para cada um dos 23 atributos de imagens de café gelado, as pessoas selecionaram todas as marcas que foram descritas pelo atributo. As seis marcas são indicadas como AA, BB, CC, DD, EE e FF para preservar a confidencialidade.
- **contacts.sav.** Esse é um arquivo de dados hipotéticos que diz respeito às listas de contato para um grupo de representantes de vendas de computadores corporativos. Cada contato é categorizado pelo departamento da empresa em que trabalham e suas classificações da empresa. Também registrada é a quantia da última venda realizada, o tempo desde a última venda e o tamanho da empresa de contato.
- **credit_card.sav.** Um estudo hipotético de uso de cartão de crédito acompanha os gastos mensais de cada sujeito em seu cartão primário por dois anos, com gastos divididos por tipo de transação (Supermercado, Varejo, Entretenimento, Viagem e outros). Cada registro no conjunto de dados corresponde a um determinado mês de gastos e tipo de transação, portanto, os dados coletados para cada assunto requerem 2 anos × 12 meses por ano × 5 tipos de transações = 120 registros.
- **creditpromo.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de uma loja de departamentos para avaliar a eficácia de uma promoção recente de cartões de crédito. Para essa finalidade, 500 titulares de cartões foram selecionados aleatoriamente. Metade recebeu um anúncio promovendo uma taxa de juros reduzida em compras feitas nos próximos três meses. Metade recebeu um anúncio sazonal padrão.
- **cross_sell.sav.** Uma empresa de encomendas por correio tem um clube de livros e um clube de CDs. A cada mês, ela disponibiliza ofertas especiais aos membros do clube. A empresa quer criar um modelo para o total de compras de ofertas especiais do mês com base no total de compras de livros, compras de CD e o tipo de oferta oferecida aos membros do clube. A Regressão de mínimos quadrados de dois estágios é apropriada para essa situação porque o dinheiro gasto em ofertas especiais é dinheiro não gasto em livros ou CD'; portanto, há um loop de feedback entre a resposta e esses dois preditores.
- **customer_dbase.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de uma empresa em usar as informações em seu data warehouse para fazer ofertas especiais para clientes que têm maior probabilidade de responder. Um subconjunto da base de clientes foi selecionado aleatoriamente e recebeu as ofertas especiais e suas respostas foram registradas.
- **customer_information.sav.** Um arquivo de dados hipotéticos que contém informações de correspondência do cliente, como nome e endereço.
- **customer_subset.sav.** Um subconjunto de 80 casos de *customer_dbase.sav*.
- **debate.sav.** Esse é um arquivo de dados hipotéticos que diz respeito às respostas pairwise a uma pesquisa de participantes de um debate político antes e depois do debate. Cada caso corresponde a um entrevistado separado.
- **debate_aggregate.sav.** Esse é um arquivo de dados hipotéticos que agrega as respostas em *debate.sav*. Cada caso corresponde a uma classificação cruzada de preferência antes e depois do debate.
- **demo.sav.** Esse é um arquivo de dados hipotéticos que diz respeito a um banco de dados de clientes adquiridos, com o objetivo de enviar por e-mail as ofertas mensais. É registrado se o cliente respondeu ou não à oferta, juntamente com várias informações demográficas.
- **demo_cs_1.sav.** Esse é um arquivo de dados hipotéticos que diz respeito à primeira etapa dos esforços de uma empresa em compilar um banco de dados de informações de pesquisas. Cada caso corresponde a uma cidade diferente e a região, província, distrito e identificação da cidade são registrados.
- **demo_cs_2.sav.** Esse é um arquivo de dados hipotéticos que diz respeito à segunda etapa dos esforços de uma empresa em compilar um banco de dados de informações de pesquisas. Cada caso corresponde a uma unidade doméstica diferente das cidades selecionadas na primeira etapa e a região, província, distrito, cidade, subdivisão e identificação da unidade são registrados. As informações de amostragem dos dois primeiros estágios do design também são incluídas.

6. Kennedy, R., C. Riquier, e 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.

- **demo_cs.sav.** Esse é um arquivo de dados hipotéticos que contém informações de pesquisas coletadas usando um plano de amostragem complexo. Cada caso corresponde a uma unidade doméstica diferente e várias informações demográficas e de amostragem são registradas.
- **diabetes_costs.sav.** Esse é um arquivo de dados hipotéticos que contém informações que são mantidas por uma empresa de seguros sobre beneficiários do seguro que têm diabetes. Cada caso corresponde a um beneficiário de seguro diferente.
- **dietstudy.sav.** Este arquivo de dados hipotéticos contém os resultados de um estudo da "dieta Stillman" [7]. Cada caso corresponde a um assunto separado e registra os pesos em libras antes e depois da dieta e os níveis de triglicérides em mg/100 ml.
- **dmdata.sav.** Esse é um arquivo de dados hipotéticos que contém informações demográficas e de compras para uma empresa de marketing direto. *dmdata2.sav* contém informações para um subconjunto de contatos que recebeu um teste de envio e *dmdata3.sav* contém informações sobre os contatos restantes que não receberam o teste de envio.
- **dvdplayer.sav.** Esse é um arquivo de dados hipotéticos que diz respeito ao desenvolvimento de um novo DVD player. Usando um protótipo, a equipe de marketing coletou dados do grupo em foco. Cada caso corresponde a um usuário pesquisado separado e registra algumas informações demográficas sobre ele e suas respostas às perguntas sobre o protótipo.
- **Employee data.sav.** Esse é um arquivo de dados hipotéticos que contém informações específicas do funcionário (nível de educação, categoria de emprego, salário atual, experiência anterior, e assim por diante).
- **german_credit.sav.** Este arquivo de dados é obtido do conjunto de dados "Crédito alemão" no Repositório de Bancos de Dados de Aprendizado de Máquina [8] na Universidade da Califórnia, Irvine.
- **grocery_1month.sav.** Este arquivo de dados hipotéticos é o arquivo de dados *grocery_coupons.sav* com as compras semanais "agregadas" para que cada caso corresponda a um cliente separado. Algumas das variáveis que foram alteradas semanalmente desaparecem como resultado e o montante gasto registrado é, agora, a soma dos montantes gastos durante as quatro semanas do estudo.
- **grocery_coupons.sav.** Esse é um arquivo de dados hipotéticos que contém dados da pesquisa coletados por uma cadeia de supermercados interessada nos hábitos de compra de seus clientes. Cada cliente é seguido durante quatro semanas e cada caso corresponde às informações de registros e semanais do cliente separado sobre onde e como o cliente compra, incluindo quanto foi gasto em supermercados durante essa semana.
- **guttman.sav.** Bell [9] apresentou uma tabela para ilustrar os grupos sociais possíveis. Guttman [10] usou uma parte dessa tabela, na qual cinco variáveis descrevendo coisas como interação social, sentimentos em pertencer a um grupo, proximidade física dos membros e formalidade do relacionamento foram cruzadas com sete grupos sociais teóricos, incluindo multidões (por exemplo, pessoas em um jogo de futebol), públicos (por exemplo, pessoas em um teatro ou leitura em sala de aula), público (por exemplo, públicos de jornais ou televisão), aglomerações (como uma multidão, mas com interação muito mais intensa), grupos primários (familiares), grupos secundários (voluntários) e a comunidade moderna (confederação informal resultante da grande proximidade física e da necessidade de serviços especializados).
- **health_funding.sav.** Esse é um arquivo de dados hipotéticos que contém dados sobre financiamento de saúde (quantia por 100 da população), taxas de doenças (taxa por 10.000 da população) e visitas aos provedores de assistência médica (taxa por 10.000 da população). Cada caso representa uma cidade diferente.

---

7. Rickman, R., N. Mitchell, J. Dingman, e J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association,* 228:, 54-58.
8. Blake, C. L., e C. J. Merz. 1998. "UCI Repository of machine learning databases". Disponível em http://www.ics.uci.edu/~mlearn/MLRepository.html.
9. Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. Nova 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.

- **hivassay.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de um laboratório farmacêutico em desenvolver uma análise rápida para detectar a infecção por HIV. Os resultados da análise são oito tonalidades profundas de vermelho, com tonalidades mais profundas indicando maior probabilidade de infecção. Um teste de laboratório foi realizado em 2.000 amostras de sangue, metade das quais estava infectada com HIV e metade estava limpa.
- **hourlywagedata.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos salários por hora de enfermeiros de posições de escritório e hospitalar e com níveis variados de experiência.
- **insurance_claims.sav.** Esse é um arquivo de dados hipotéticos que diz respeito a uma empresa de seguros que deseja construir um modelo para sinalizar reclamações suspeitas e potencialmente fraudulentas. Cada caso representa uma reclamação separada.
- **insure.sav.** Esse é um arquivo de dados hipotéticos que diz respeito a uma empresa de seguros que está estudando os fatores de risco que indicam se um cliente terá que fazer uma reclamação sobre um contrato de seguro de vida de termo de 10 anos. Cada caso no arquivo de dados representa um par de contratos, um dos quais registrou uma reclamação e o outro não registrou, com correspondência de idade e sexo.
- **judges.sav.** Esse é um arquivo de dados hipotéticos que diz respeito às pontuações fornecidas por juízes treinados (mais um entusiasta) a 300 apresentações de ginástica. Cada linha representa uma apresentação separada; os juízes visualizaram as mesmas apresentações.
- **kinship_dat.sav.** Rosenberg e Kim [11] definiram analisar 15 termos de parentesco (tia, irmão, primo, filha, pai, neta, avô, avó, neto, mãe, sobrinho, sobrinha, irmã, filho, tio). Eles pediram a quatro grupos de estudantes universitários (dois femininos, dois masculinos) que classificassem esses termos com base em semelhanças. Foi solicitado que dois grupos (um feminino, um masculino) classificassem duas vezes, com a segunda classificação baseada em um critério diferente da primeira. Assim, um total de seis "fontes" foi obtido. Cada fonte corresponde a uma matriz de proximidade de 15 x 15, cujas células são iguais ao número de pessoas em uma fonte menos o número de vezes que os objetos foram particionados juntos nessa fonte.
- **kinship_ini.sav.** Este arquivo de dados contém uma configuração inicial para uma solução tridimensional para *kinship_dat.sav*.
- **kinship_var.sav.** Este arquivo de dados contém variáveis independentes *sexo*, *ger*(ação) e *grau* (de separação) que podem ser usadas para interpretar as dimensões de uma solução para *kinship_dat.sav*. Especificamente, elas podem ser usadas para restringir o espaço da solução para uma combinação linear dessas variáveis.
- **marketvalues.sav.** Este arquivo de dados diz respeito a vendas de moradias no desenvolvimento de um novo alojamento em Algonquin, Ill., durante os anos de 1999–2000. Estas vendas são uma questão de registro público.
- **nhis2000_subset.sav.** A National Health Interview Survey (NHIS) é uma pesquisa grande, baseada na população civil dos EUA. As entrevistas são realizadas através de comunicação direta em uma amostra nacionalmente representativa das famílias. Informações demográficas e observações sobre comportamentos de funcionamento e status são obtidas para membros de cada família. Esse arquivo de dados contém um subconjunto de informações da pesquisa de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Documentação e arquivo de dados de uso público. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Acessado 2003.
- **ozone.sav.** Os dados incluem 330 observações sobre seis variáveis meteorológicas para prever a concentração de ozônio das variáveis restantes. Pesquisadores anteriores [12], [13], entre outras coisas, encontraram falta de linearidade entre essas variáveis, o que prejudica as abordagens de regressão padrão.

---

11. Rosenberg, S., e 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., e J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association,* 80, 580-598.

13. Hastie, T., e R. Tibshirani. 1990. *Generalized additive models*. Londres: Chapman e Hall.

- **pain_medication.sav.** Este arquivo de dados hipotéticos contém os resultados de um teste clínico para medicação anti-inflamatória para tratar a dor artrítica crônica. De interesse específico é o tempo que leva para o medicamento fazer efeito e como ele se compara a uma medicação existente.
- **patient_los.sav.** Este arquivo de dados hipotéticos contém os registros do tratamento de pacientes que foram internados no hospital por suspeita de enfarte do miocárdio (MI ou "ataque cardíaco"). Cada caso corresponde a um paciente separado e registra muitas variáveis relacionadas à sua permanência no hospital.
- **patlos_sample.sav.** Este arquivo de dados hipotéticos contém os registros de tratamento de uma amostra de pacientes que recebeu trombolíticos durante o tratamento para enfarte do miocárdio (MI ou "ataque cardíaco"). Cada caso corresponde a um paciente separado e registra muitas variáveis relacionadas à sua permanência no hospital.
- **poll_cs.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de especialistas em pesquisa em determinar o nível de suporte público para uma nota antes da legislatura. Os casos correspondem aos eleitores registrados. Cada caso registra o estado, município e bairro em que o eleitor vive.
- **poll_cs_sample.sav.** Este arquivo de dados hipotéticos contém uma amostra dos eleitores listados em *poll_cs.sav*. A amostra foi obtida de acordo com o design especificado no arquivo de plano *poll.csplan* e este arquivo de dados registra as probabilidades de inclusão e as ponderações da amostra. Observe, no entanto, que, como o plano de amostragem usa um método de probabilidade-proporcional-ao-tamanho (PPS), há também um arquivo que contém as probabilidades de seleção de conjunto (*poll_jointprob.sav*). As variáveis adicionais correspondentes aos dados demográficos do eleitor e sua opinião sobre a nota proposta foram coletadas e incluíram o arquivo de dados depois que a amostra foi obtida.
- **property_assess.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de um assessor de estado em manter as avaliações do valor de propriedade atualizadas em relação aos recursos limitados. Os casos correspondem às propriedades vendidas no estado no ano passado. Cada caso no arquivo de dados registra o município no qual a propriedade está, o assessor que visitou pela última vez a propriedade, o tempo desde essa avaliação, a valorização feita naquele momento e o valor de venda da propriedade.
- **property_assess_cs.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de um assessor de estado em manter as avaliações de valor da propriedade atualizadas em relação aos recursos limitados. Os casos correspondem às propriedades no estado. Cada caso no arquivo de dados registra o estado, município e bairro em que a propriedade está, o tempo desde a última avaliação e a valorização feita naquele momento.
- **property_assess_cs_sample.sav.** Este arquivo de dados hipotéticos contém uma amostra das propriedades listadas em *property_assess_cs.sav*. A amostra foi obtida de acordo com o design especificado no arquivo de plano *property_assess.csplan* e esse arquivo de dados registra as probabilidades de inclusão e as ponderações da amostra. A variável adicional *Valor atual* foi coletada e incluída no arquivo de dados depois que a amostra foi obtida.
- **recidivism.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de uma agência de aplicação da lei do governo em entender as taxas de reincidência em sua área de jurisdição. Cada caso corresponde a um ofensor anterior e registra suas informações demográficas, alguns detalhes de seu primeiro crime e o tempo até sua segunda prisão, se tiver ocorrido dentro de dois anos após a primeira prisão.
- **recidivism_cs_sample.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de uma agência de aplicação da lei do governo em entender as taxas de reincidência em sua área de jurisdição. Cada caso corresponde a um ofensor anterior, solto de sua primeira prisão durante o mês de junho de 2003 e registra suas informações demográficas, alguns detalhes de seu primeiro crime e os dados de sua segunda prisão, se tiver ocorrido no final de junho de 2006. Os ofensores foram selecionados de departamentos de amostragem de acordo com o plano de amostragem especificado em *recidivism_cs.csplan*; como ele faz uso de um método de probabilidade-proporcional-ao-tamanho (PPS), há também um arquivo que contém as probabilidades de seleção de conjunto (*recidivism_cs_jointprob.sav*).

- **rfm_transactions.sav.** Um arquivo de dados hipotéticos que contém dados da transação de compra, incluindo a data da compra, o(s) item(ns) comprado(s) e o valor monetário de cada transação.
- **salesperformance.sav.** Esse é um arquivo de dados hipotéticos que diz respeito à avaliação de dois novos cursos de treinamento de vendas. Sessenta funcionários, divididos em três grupos, receberão treinamento padrão. Além disso, o grupo 2 obtém treinamento técnico; o grupo 3, um tutorial prático. Cada funcionário foi testado no final do curso de treinamento e sua pontuação foi registrada. Cada caso no arquivo de dados representa um estagiário separado e registra o grupo ao qual ele foi designado e a pontuação que recebeu no exame.
- **satisf.sav.** Esse é um arquivo de dados hipotéticos que diz respeito a uma pesquisa de satisfação realizada por uma empresa de varejo em 4 locais de loja. 582 clientes foram pesquisados e cada caso representa as respostas de um único cliente.
- **screws.sav.** Este arquivo de dados contém informações sobre as características de parafusos, cavilhas, porcas e tachas [14].
- **shampoo_ph.sav.** Esse é um arquivo de dados hipotéticos que diz respeito ao controle de qualidade em uma fábrica de produtos para cabelos. Em intervalos de tempo regulares, seis lotes de saída separados são medidos e seu pH é registrado. O intervalo de destino é 4.5–5.5.
- **ships.sav.** Um conjunto de dados apresentado e analisado em outro lugar [15] que diz respeito aos danos aos navios de carga causados pelas ondas. As contagens de incidentes podem ser modeladas como ocorrendo em uma taxa de Poisson dado o tipo do navio, o período de construção e o período de serviço. Os meses agregados de serviço para cada célula da tabela formada pela classificação cruzada de fatores fornecem os valores para a exposição ao risco.
- **site.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de uma empresa em escolher novos sites para sua empresa em expansão. Eles contrataram dois consultores para avaliar separadamente os sites, quem, além de um relatório estendido, resumiu cada site como um possível cliente "bom", "justo" ou "fraco".
- **smokers.sav.** Este arquivo de dados é resumido do 1998 National Household Survey of Drug Abuse e é uma amostra da probabilidade dos lares americanos. (http://dx.doi.org/10.3886/ICPSR02934) Assim, a primeira etapa em uma análise deste arquivo de dados deve ser ponderar os dados para refletir as tendências da população.
- **stocks.sav** Este arquivo de dados hipotéticos contém preços e volume de estoque para um ano.
- **stroke_clean.sav.** Este arquivo de dados hipotéticos contém o estado de um banco de dados médico após ter sido limpo seguindo os procedimentos em Statistics Base Edition.
- **stroke_invalid.sav.** Este arquivo de dados hipotéticos contém o estado inicial de um banco de dados médico e contém vários erros de entrada de dados.
- **stroke_survival.** Este arquivo de dados hipotéticos diz respeito aos tempos de sobrevivência para pacientes que estão saindo de um programa de reabilitação pós-derrame isquêmico diante de vários desafios. Pós-derrame, a ocorrência de enfarte do miocárdio, derrame isquêmico ou derrame hemorrágico é observado e o momento do evento registrado. A amostra é truncada à esquerda porque inclui apenas pacientes que sobreviveram até o fim do programa de reabilitação administrado pós-derrame.
- **stroke_valid.sav.** Este arquivo de dados hipotéticos contém o estado de um banco de dados médico depois que os valores foram verificados usando o procedimento Validar Dados. Ele contém também casos potencialmente anormais.
- **survey_sample.sav.** Este arquivo de dados contém dados da pesquisa, incluindo dados demográficos e várias medidas de atitude. Ele é baseado em um subconjunto de variáveis do 1998 NORC General Social Survey, embora alguns valores de dados tenham sido modificados e variáveis fictícias adicionais tenham sido incluídas para fins de demonstração.

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

15. McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2ª ed. Londres: Chapman & Hall.

- **tcm_kpi.sav.** Esse é um arquivo de dados hipotéticos que contém valores dos principais indicadores de desempenho semanais para um negócio. Ele também contém dados semanais para várias métricas controláveis durante o mesmo período de tempo.
- **tcm_kpi_upd.sav.** Este arquivo de dados é idêntico a *tcm_kpi.sav*, mas contém dados para quatro semanas extras.
- **telco.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de uma empresa de telecomunicações para reduzir a perda de clientes na base de clientes. Cada caso corresponde a um cliente separado e registra várias informações demográficas e de uso de serviço.
- **telco_extra.sav.** Este arquivo de dados é semelhante ao arquivo de dados *telco.sav*, mas o "aforamento" e as variáveis de gastos do cliente transformadas por log foram removidos e substituídos por variáveis de gasto do cliente transformadas por log padronizadas.
- **telco_missing.sav.** Este arquivo de dados é um subconjunto do arquivo de dados *telco.sav*, mas alguns dos valores de dados demográficos foram substituídos por valores omissos.
- **testmarket.sav.** Este arquivo de dados hipotéticos diz respeito aos planos de uma cadeia de fast food para incluir um novo item no menu. Há três campanhas possíveis para promover o novo produto, portanto, o novo item é introduzido em locais em vários mercados selecionados aleatoriamente. Uma promoção diferente é usada em cada local e as vendas semanais do novo item são registradas para as primeiras quatro semanas. Cada caso corresponde a um local semanal separado.
- **testmarket_1month.sav.** Este arquivo de dados hipotéticos é o arquivo de dados *testmarket.sav* com as vendas semanais "agregadas" para que cada caso corresponda a um local separado. Algumas das variáveis que foram alteradas semanalmente desaparecem como resultado e as vendas registradas são, agora, a soma das vendas durante as quatro semanas do estudo.
- **tree_car.sav.** Esse é um arquivo de dados hipotéticos que contém dados demográficos e de preço de compra do veículo.
- **tree_credit.sav.** Esse é um arquivo de dados hipotéticos que contém dados demográficos e de histórico de empréstimo bancário.
- **tree_missing_data.sav** Esse é um arquivo de dados hipotéticos que contém dados demográficos e de histórico de empréstimo bancário com um grande número de valores omissos.
- **tree_score_car.sav.** Esse é um arquivo de dados hipotéticos que contém dados demográficos e de preço de compra do veículo.
- **tree_textdata.sav.** Um arquivo de dados simples com apenas duas variáveis destinadas principalmente a mostrar o estado padrão de variáveis antes da designação do nível de medição e dos rótulos de valor.
- **tv-survey.sav.** Esse é um arquivo de dados hipotéticos que diz respeito a uma pesquisa realizada por um estúdio de televisão que está considerando se deve estender a execução de um programa bem-sucedido. Foi perguntado a 906 entrevistados se eles assistiriam ao programa sob várias condições. Cada linha representa um entrevistado separado; cada coluna é uma condição separada.
- **ulcer_recurrence.sav.** Este arquivo contém informações parciais de um estudo projetado para comparar a eficácia de duas terapias para evitar a recorrência de úlceras. Ele fornece um bom exemplo de dados com intervalo censurado e foi apresentado e analisado em outro lugar [16].
- **ulcer_recurrence_recoded.sav.** Este arquivo reorganiza as informações em *ulcer_recurrence.sav* para permitir que você modele a probabilidade de eventos para cada intervalo do estudo em vez de simplesmente a probabilidade de eventos no final do estudo. Ele foi apresentado e analisado em outro lugar [17].
- **verd1985.sav.** Este arquivo de dados diz respeito a uma pesquisa [18]. As respostas de 15 pessoas a 8 variáveis foram registradas. As variáveis de interesse são divididas em três conjuntos. O conjunto 1 inclui *idade* e *estado civil*, o conjunto 2 inclui *animal de estimação* e *notícias*, e o conjunto 3 inclui *música* e

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

em tempo real. *Animal de estimação* é escalado como nominal múltiplo e *idade* é escalado como ordinal; todas as outras variáveis são escaladas como nominal único.

- **virus.sav.** Esse é um arquivo de dados hipotéticos que diz respeito aos esforços de um provedor de serviços de Internet (ISP) em determinar os efeitos de um vírus em suas redes. Eles têm controlado a porcentagem (aproximada) de tráfego de e-mail infectado em suas redes ao longo do tempo, do momento da descoberta até a ameaça ser contida.
- **wheeze_steubenville.sav.** Este é um subconjunto de um estudo longitudinal dos efeitos da poluição do ar na saúde de crianças [19]. Os dados contêm medidas binárias repetidas de status de respiração difícil para crianças de Steubenville, Ohio, nas idades de 7, 8, 9 e 10 anos, juntamente com um registro fixo de se a mãe era ou não fumante durante o primeiro ano do estudo.
- **workprog.sav.** Esse é um arquivo de dados hipotéticos que diz respeito a um programa de trabalho do governo que tenta colocar pessoas desfavorecidas em empregos melhores. Uma amostra dos participantes potenciais do programa foi seguida, alguns dos quais foram selecionados aleatoriamente para inscrição no programa, enquanto outros não foram. Cada caso representa um participante separado do programa.
- **worldsales.sav** Este arquivo de dados hipotéticos contém a receita de vendas por continente e produto.

19. Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, e 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.

Avisos

Essas informações foram desenvolvidas para produtos e serviços oferecidos nos Estados Unidos. Esse material pode estar disponível a partir da IBM em outros idiomas. No entanto, pode ser necessário possuir uma cópia do produto ou da versão do produto nesse idioma para acessá-lo.

É possível que a IBM não ofereça produtos, serviços ou recursos discutidos neste documento em outros países. Consulte um representante IBM local para obter informações sobre produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser utilizados. Qualquer produto, programa ou serviço funcionalmente equivalente, que não infrinja nenhum direito de propriedade intelectual da IBM poderá ser utilizado em substituição a este produto, programa ou serviço. Entretanto, a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade do Cliente.

A IBM pode ter patentes ou solicitações de patentes pendentes relativas a assuntos tratados nesta publicação. O fornecimento desta publicação não lhe garante direito algum sobre tais patentes. Pedidos de licença podem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
Rio de Janeiro, RJ
Brasil

Para pedidos de licença relacionados a informações de DBCS (Conjunto de Caracteres de Byte Duplo), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie pedidos de licença, por escrito, para:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS NÃO SE LIMITANDO ÀS GARANTIAS IMPLÍCITAS DE NÃO-VIOLAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias explícitas ou implícitas em certas transações; portanto, esta instrução pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar o(s) produto(s) e/ou programa(s) descritos nesta publicação, sem aviso prévio.

Qualquer referência nestas informações a websites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a esses websites. Os materiais contidos nesses websites não fazem parte dos materiais para esse produto IBM e o uso desses websites é de inteira responsabilidade do Cliente.

A IBM por usar ou distribuir as informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o Cliente.

Licenciados deste programa que desejam obter informações sobre o mesmo com o objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) o uso mútuo de informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
Rio de Janeiro, RJ
Brasil

Tais informações podem estar disponíveis, sujeitas a termos e condições apropriadas, incluindo em alguns casos o pagamento de uma taxa.

O programa licenciado descrito nesta publicação e todo o material licenciado disponível são fornecidos pela IBM sob os termos do Contrato com o Cliente IBM, do Contrato Internacional de Licença do Programa IBM ou de qualquer outro contrato equivalente.

Os exemplos de dados de desempenho e do Cliente citados são apresentados apenas para propósitos ilustrativos. Resultados de desempenho reais podem variar dependendo das configurações específicas e das condições operacionais.

Informações relativas a produtos não IBM foram obtidas junto aos fornecedores dos respectivos produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou esses produtos e não pode confirmar a precisão de desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Perguntas sobre os recursos de produtos não IBM devem ser endereçadas aos fornecedores desses produtos.

Instruções relativas à direção futura ou intento da IBM estão sujeitas a mudança ou retirada sem aviso e representam metas e objetivos apenas.

Estas informações contêm exemplos de dados e relatórios utilizados nas operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de assuntos, empresas, marcas e produtos. Todos esses nomes são fictícios e qualquer semelhança com pessoas ou empresas reais é mera coincidência.

LICENÇA DE COPYRIGHT:

Estas informações contêm programas de aplicativos de amostra na linguagem fonte, ilustrando as técnicas de programação em diversas plataformas operacionais. O Cliente pode copiar, modificar e distribuir estes programas de amostra sem a necessidade de pagar à IBM, com objetivos de desenvolvimento, utilização, marketing ou distribuição de programas aplicativos em conformidade com a interface de programação de aplicativo para a plataforma operacional para a qual os programas de amostra são criados. Esses exemplos não foram testados completamente em todas as condições. Portanto, a IBM não pode garantir ou implicar a confiabilidade, manutenção ou função destes programas. Os programas de amostra são fornecidos "NO ESTADO EM QUE SE ENCONTRAM", sem garantia de qualquer tipo. A IBM não será responsabilizada por quaisquer danos decorrentes do uso dos programas de amostra.

Cada cópia ou parte destes programas de amostra ou qualquer trabalho derivado deve incluir um aviso de copyright com os dizeres:

Marcas comerciais

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em muitos países no mundo todo. Outros nomes de produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. ma lista atual de marcas comerciais da IBM está disponível na web em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml.

Adobe, o logotipo Adobe, PostScript e o logotipo PostScript são marcas registradas ou marcas comerciais da Adobe Systems Incorporated nos Estados Unidos e/ou em outros países.

Intel, o logotipo Intel, Intel Inside, o logotipo Intel Inside, Intel Centrino, o logotipo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou de suas subsidiárias nos Estados Unidos e em outros países.

Linux é uma marca registrada da Linus Torvalds nos Estados Unidos, e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou em outros países.

UNIX é uma marca registrada da The Open Group nos Estados Unidos e em outros países.

Java e todas as marcas comerciais e logotipos baseados em Java são marcas comerciais ou marcas registradas da Oracle e/ou suas afiliadas.

Índice Remissivo

Impresso no Brasil