Настраиваемые таблицы 26 IBM SPSS

Примечание

Прежде чем использовать эту информацию и продукт, описанный в ней, прочтите сведения в разделе “Замечания” на стр. 19.

Информация о продукте

Это издание применимо к версии 26, выпуск 0, модификация 0 IBM® SPSS Statistics и ко всем последующим версиям и модификациям до тех пор, пока в новых изданиях не будет указано иное.

Содержание

Настраиваемые таблицы

В SPSS Statistics Standard Edition или в модуль Настраиваемые таблицы включены следующие функции настраиваемых таблиц.

Интерфейс Настраиваемые таблицы

Интерфейс построителя таблиц

Пользовательские таблицы создаются при помощи удобного интерфейса, который позволяет в процессе создания таблицы видеть, как она будет выглядеть. Процедура также, в отличие от традиционных процедур с "диалоговыми окнами", обеспечивает высокий уровень гибкости, включая возможность изменять размер окна и размер панелей в окне.

Построение таблиц

На вкладке Таблица можно выбрать переменные и итожащие статистики, которые должны появиться из интерфейса Настраиваемые таблицы.

Анализ > Таблицы > Настраиваемые таблицы

Список переменных. Переменные файла данных показаны в левой части окна. Процедура Настраиваемые таблицы различает два уровня измерения переменных и различным образом обрабатывает переменные разных уровней измерения:

Категориальное. Это данные с ограниченным числом уникальных значений или категорий (например, пол или религия). Категориальные переменные могут быть текстовыми или числовыми, в которых категории закодированы числовыми кодами (например, 0 = *Женский* , а 1 = *Мужской*). Также эти данные называются качественными данными. Категориальные переменные могут быть либо **номинальные**, либо **порядковые**

- *Номинальная*. Переменную можно рассматривать как номинальную, когда ее значения представляют категории без естественного упорядочения, например, подразделение компании, где работает наемный сотрудник. Примеры номинальных переменных включают регион, почтовый индекс или религию.
- *Порядковая*. Переменную можно рассматривать как порядковую, когда ее значения представляют категории с некоторым естественным для них упорядочением, например, уровни удовлетворенности обслуживанием от крайней неудовлетворенности до крайней удовлетворенности. Примеры порядковых переменных включают баллы, представляющие степень удовлетворенности или уверенности, или баллы, оценивающие предпочтение.

Категориальные переменные задают в таблице категории (в строках, столбцах или слоях), и статистикой, выводимой по умолчанию, является частота (количество наблюдений в каждой категории). К примеру, таблица для категориальной переменной пола, выводимая с параметрами по умолчанию, будет содержать в себе только количество мужчин и женщин.

Масштаб. Это данные, измеренные на интервальной шкале или на шкале отношений, для которых существует и порядок значений, и расстояния между значениями. Например, зарплата 7195 рублей больше зарплаты 5398 рублей, а расстояние между этими зарплатами - 1797 рублей. Такие данные также называют непрерывными.

Количественные переменные обычно подытоживаются для категорий категориальных переменных, и статистикой, выводимой по умолчанию, является среднее значение. К примеру, таблица с переменной дохода по категориям пола будет содержать в себе средние значения дохода для мужчин и женщин.

Также можно подытоживать количественные переменные сами по себе, без использования категориальных переменных для задания групп. Это очень полезно в случае **состыковывания** итожащих статистик для нескольких количественных переменных.

Наборы множественных ответов

Настраиваемые таблицы также позволяют работать со специальными переменными, которые называются **наборами множественных ответов** . Наборы множественных ответов не являются переменными в обычном смысле. Их нельзя увидеть в Редакторе данных, и они не распознаются другими процедурами. В наборах множественных ответов для ввода ответов на вопросы, на которые можно дать больше одного ответа, используются несколько переменных. Наборы множественных ответов обрабатываются как категориальные переменные, и большинство операций, которые можно выполнять с категориальными переменными, можно делать также и с наборами множественных ответов.

Значок рядом с каждой переменной позволяет идентифицировать тип переменной.

Категории. Когда вы выделяете в списке переменных категориальную переменную, категории переменной отображаются в окне Информация о переменной. Когда вы задаете данную переменную в таблице, эти категории выводятся на панели холста. Если переменная не имеет заданных категорий, то в окне Информация о переменной и на панели холста выводятся только две категории, задаваемые по умолчанию: *Категория 1* и *Категория 2*.

Категории, показанные в построителе таблиц, основаны на **метках значений**, т.е. описательных метках, присваиваемых различным значениям данных (к примеру, числовые значения 0 и 1 могут иметь, соответственно, метки значений *мужской* и *женский*). Пользователь может определить метки переменных на панели Информация о переменных в редакторе данных.

Панель холста. Построение таблицы осуществляется посредством переноса переменных в строки и столбцы панели холста. Панель холста позволяет увидеть, как будет выглядеть таблица. Реальные значения данных не показаны в ячейках таблицы на панели холста, но при помощи панели холста вы всегда можете увидеть холст той таблицы, которую собираетесь построить. Для категориальных переменных построенная таблица может содержать больше категорий, чем таблица на панели макета, в том случае, если в файле данных переменные имеют значения, для которых не заданы метки значений.

Основные правила и ограничения процесса построения таблиц

- Для категориальных переменных итожащие статистики основываются на переменной на внутреннем уровне вложения в измерении статистик.
- Определение измерения статистик (строки или столбцы) для категориальных переменных основано на порядке переноса переменных на панель холста. К примеру, если первая переменная переносится в поле строк, то именно измерение строк станет измерением статистик.
- Количественные переменные могут подытоживаться только внутри категорий переменных на внутреннем уровне вложения как в измерении строк, так и в измерении столбцов. (Вы можете располагать количественную переменную на любом уровне в таблице, но подытоживание будет происходить на внутреннем уровне вложения).
- Количественные переменные не могут подытоживаться внутри других количественных переменных. Вы можете состыковывать несколько количественных переменных или подытоживать количественные переменные по категориям категориальных переменных. Но вы не можете вкладывать одну количественную переменную в другую или располагать одну количественную переменную в измерении строк, а другую количественную переменную - в измерении столбцов.
- Если какая-либо переменная в активном наборе данных содержит более 12 000 определенных меток, пользователь не может использовать построитель таблиц для создания таблиц. Если нет необходимости включать переменные с превышением этого порога в пользовательские таблицы, можно определить и применить наборы, из которых эти переменные будут исключены. Если необходимо использовать какие-либо переменные с более чем 12000 определенными метками значений, воспользуйтесь командой `CTABLES` для создания таких таблиц.

Как построить таблицу

1. Выберите в меню:

   Анализ > Таблицы > Настраиваемые таблицы
2. Перенесите одну или несколько переменных в поле строк и/или поле столбцов на панели холста.
3. Нажмите **Создать**, чтобы создать таблицу.

Для удаления переменной из панели холста

1. Выберите (щелкните) переменную на панели холста.
2. Щелкните правой кнопкой мыши и выберите в выпадающем меню **Удалить переменную**.

Вложение переменных

Вложение, подобно таблицам сопряженности, предназначено для выявления связи между двумя категориальными переменными. Его особенность состоит в том, что одна переменная вкладывается в другую в одном измерении таблицы. Например, можно вложить переменную *Пол респондента* в переменную *Возрастная категория* в измерении строк, чтобы количество мужчин и женщин было показано для каждой возрастной категории.

Вы также можете вложить количественную переменную в категориальную переменную. Например, вы можете вложить переменную *Доход респондента* в переменную *Пол респондента* для вывода отдельных средних значений (медиан или других итожащих статистик) дохода для мужчин и для женщин.

Чтобы вложить переменные

1. Перенесите категориальную переменную в поле строк или поле столбцов панели холста.
2. Перетащите категориальную или количественную переменную на категориальную переменную сроки или столбца.
3. Выберите из меню **Вложить выше всех переменных**, **Вложить слева** или **Вложить справа**.

Таблица 1. Вложенные категориальные переменные

Переменная 1	**Переменная 2**	**Сводная статистика**
Категория 1	Категория 1	12
	Категория 2	34
	Категория 3	56
Категория 2	Категория 1	12
	Категория 2	34
	Категория 3	56

Примечание: Настраиваемые таблицы не поддерживают послойную обработку (сравнение групп) при обработке файла разбиения. Чтобы добиться того же результата, как при послойной обработке файлов разбиения, расположите переменные файла разбиения в таблице в самых ближних внешних слоях вложения.

Изменить статистики

В окне Изменить статистики можно:

- Добавлять и удалять итожащие статистики из таблицы.

Какие статистики (и другие опции) доступны на панели Изменить статистики, зависит от уровня измерения переменной, для которой задаются статистики. Источник статистик (переменная, для которой будут задаваться статистики) определяется при помощи следующих правил:

- **Уровень измерения** . Если таблица (или подтаблица в составной таблице) содержит количественную переменную, то статистики будут основываться на данной количественной переменной.

- **Порядок выбора переменных** . Определение измерения статистик (строки или столбцы) для категориальных переменных основано на порядке переноса переменных на панель холста. К примеру, если первая переменная переносится в поле строк, то именно измерение строк станет измерением статистик.
- **Вложение** . Для категориальных переменных статистики основываются на переменной на внутреннем уровне вложения в измерении статистик.

Итожащие статистики для категориальных переменных: Основными статистиками, доступными для категориальных переменных, являются количества и проценты. Для итогов и подытогов вы можете задать другие итожащие статистики. Задаваемые итожащие статистики включают в себя меры положения центра распределения (такие как среднее и медиана) и дисперсии (такие как стандартное отклонение), которые могут быть применены к некоторым порядковым категориальным переменным.

Частота. Количество наблюдений в каждой ячейке или количество ответов в наборе множественных ответов. Если используется взвешивание, данное значение представляет собой взвешенное количество.

- Если используется взвешивание, это значение представляет собой взвешенное количество.
- Взвешенное количество одинаково для взвешивания с глобальным набором данных (**Данные > Взвесить наблюдения...**).

Невзвешенная частота. Невзвешенное количество наблюдений в каждой ячейке таблицы. Это значение отличается от количества наблюдений только в том случае, если учитывается взвешивание.

Скорректированное количество. Скорректированное количество, используемое при вычислениях эффективных базовых весов. Если вы не используете переменную эффективных базовых весов, то скорректированное количество совпадает с количеством.

Проценты по строке. Проценты в каждой строке. Сумма процентов в каждой строке подтаблицы (простые проценты) равна 100%. Проценты по строкам обычно полезны в том случае, когда переменная *столбцов* является категориальной.

Проценты по столбцу. Проценты по каждому столбцу. Сумма процентов в каждом столбце подтаблицы (простые проценты) равна 100%. Проценты по столбцам обычно полезны в том случае, когда переменная *строк* является категориальной.

Проценты по подтаблице. Проценты в каждой ячейке основаны на общем количестве наблюдений в подтаблице. Все проценты в ячейках в подтаблице основаны на одном и том же общем количестве наблюдений, и их сумма в подтаблице равняется 100%. В таблицах с вложениями, подтаблицы задаются переменной, предшествующей самому внутреннему уровню вложения. Например, в таблице, представляющей переменную *Семейное положение* внутри переменной *Пол респондента* внутри переменной *Возрастная категория* , переменная *Пол респондента* задает подтаблицы.

Проценты по таблице. Проценты в каждой ячейке основаны на общем количестве наблюдений в таблице. Все процентные значения основаны на одном и том же общем количестве наблюдений, и их сумма равняется 100% (простые проценты).

Доверительные интервалы

- Нижний и верхний доверительные пределы доступны для счетчиков, процентов, средних, медианы, процентилей и суммы.
- Текстовая строка "&[Confidence Level]" в метке включается в метки столбцов для доверительных уровней в таблице.
- Среднеквадратичная ошибка доступна для счетчиков, процентов, средних и суммы.
- Доверительные интервалы и среднеквадратичная ошибка недоступны для наборов множественных ответов.

Уровень

Доверительный уровень для доверительных интервалов, выраженный в процентах. Значение должно быть от 1 до 99.

Несколько наборов ответов

Для наборов множественных ответов можно получить проценты, основанные на количестве наблюдений, количестве ответов или количествах. Дополнительную информацию смотрите в разделе “Итожащие статистики для наборов множественных ответов”.

База для расчета процентов: Проценты могут быть рассчитаны тремя различными способами, в зависимости от того, каким образом обрабатываются пропущенные значения:

Простые проценты. Проценты основаны на количестве наблюдений, использованных при построении таблицы, и их сумма всегда равна 100%. Если какая-либо категория исключена из таблицы, наблюдения, относящиеся к данной категории, исключаются из общей базы. Наблюдения с системными значениями отсутствия всегда исключаются из базы. Наблюдения с пользовательскими значениями отсутствия исключаются в том случае, если пользовательские категории отсутствия исключены из таблицы (по умолчанию), и включаются в том случае, если пользовательские категории отсутствия включены в таблицу. Любые проценты, которые не имеют в имени *% валидных* или *Итоговый %* являются простыми процентами.

Итоговые проценты. К базе простых процентов добавляются наблюдения с системными и пользовательскими значениями отсутствия. Сумма процентов может составлять менее 100%.

Допустимые проценты. Наблюдения с пользовательскими значениями отсутствия исключаются из базы простых процентов даже в том случае, если пользовательские категории отсутствия включены в таблицу.

Примечание: В отличие от пользовательских категорий отсутствия наблюдения из вручную исключенных категорий всегда исключаются из базы.

Итожащие статистики для наборов множественных ответов: Для наборов множественных ответов доступными являются следующие дополнительные статистики.

Процент ответов по строке/столбцу/ слою. Процент, основанный на ответах.

% ответов по столбцу/строке/слою (база: наблюдения). В числителе находится количество ответов, в знаменателе - количество респондентов.

% ответов по столбцу/строке/слою (база: ответы). В числителе находится количество, в знаменателе - итоговое количество ответов.

Процент ответов по строке/столбцу в слое. Процент по подтаблицам. Процент, основанный на ответах.

% ответов по столбцу/строке в слое (база: наблюдения). Проценты по подтаблицам. В числителе находится количество ответов, в знаменателе - количество респондентов.

% ответов по столбцу/строке в слое (база: ответы). Проценты по подтаблицам. В числителе находится количество, в знаменателе - итоговое количество ответов.

Ответы. Количество ответов.

Процент ответов по подтаблице/таблице. Процент, основанный на ответах.

% ответов по подтаблице/таблице (база: наблюдения). В числителе находится количество ответов, в знаменателе - количество респондентов.

% по подтаблице/таблице (база: ответы). В числителе находится количество, в знаменателе - итоговое количество ответов.

Итожащие статистики для количественных переменных и настраиваемых итогов категориальных переменных: В дополнение к количествам и процентам, доступным для категориальных переменных, для количественных переменных и задаваемых итогов/подытогов категориальных переменных доступны также следующие итожащие статистики. Эти итожащие статистики не доступны для наборов множественных ответов или текстовых (алфавитно-цифровых) переменных.

Среднее значение. Арифметическое среднее; сумма, деленная на число наблюдений.

Медиана. Значение, выше и ниже которого попадает половина наблюдений; 50-й процентиль.

Режим. Наиболее часто встречающееся значение. Если существует несколько модальных значений, используется наименьшее.

Минимум. Наименьшее значение.

Максимум. Наибольшее значение.

Пропущенные. Количество значений отсутствия (как пользовательских, так и системных).

Процентиль. Можно добавить в таблицу 5-й, 25-й, 75-й, 95-й и/или 99-й процентиль.

Диапазон. Разность между максимумом и минимумом.

Стандартное отклонение. Мера разброса вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.

Сумма. Сумма значений.

Проценты суммы. Проценты, основанные на суммах. Доступны для отдельных строк и столбцов (в подтаблицах), строк и столбцов в целом (по подтаблицам), слоев, подтаблиц и таблиц.

Всего. Количество присутствующих значений, пользовательских значений отсутствия и системных значений отсутствия. Не включает в себя наблюдения из категорий, исключенных вручную и отличающихся от пользовательских категорий отсутствия.

Скорректированное общее число. Скорректированное общее число, используемое при вычислениях эффективных базовых весов. Если вы не используете переменную эффективных базовых весов (вкладка Опции), скорректированное общее число совпадает с общим числом. Этот статистический показатель недоступен для наборов множественных ответов.

Допустимые N. Число присутствующих значений. Не включает в себя наблюдения из категорий, исключенных вручную и отличающихся от пользовательских категорий отсутствия.

Скорректированное число допустимых. Скорректированное число допустимых, используемое при вычислениях эффективных базовых весов. Если вы не используете переменную эффективных базовых весов (вкладка Опции), скорректированное число допустимых совпадает с общим числом допустимых.

Дисперсия. Мера разброса относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной (и является квадратом стандартного отклонения).

Доверительные интервалы

- Нижний и верхний доверительные пределы доступны для счетчиков, процентов, средних, медианы, процентилей и суммы.
- Текстовая строка "&[Confidence Level]" в метке включается в метки столбцов для доверительных уровней в таблице.
- Среднеквадратичная ошибка доступна для счетчиков, процентов, средних и суммы.
- Доверительные интервалы и среднеквадратичная ошибка недоступны для наборов множественных ответов.

Уровень

Доверительный уровень для доверительных интервалов, выраженный в процентах. Значение должно быть от 1 до 99.

Составные таблицы

Каждая подтаблица, задаваемая стыкуемой переменной, рассматривается как отдельная таблица, и итожащие статистики подсчитываются соответствующим образом.

Категории и итоги

Пользовательские таблицы позволяют:

- Переупорядочивать категории.
- Вставлять итоги.
- В случае с переменными без меток значений можно только сортировать категории и вставлять итоги.

Порядок открытия опций категорий и итогов

1. Перенесите категориальную переменную или набор множественных ответов на панель холста.
2. Щелкните правой кнопкой мыши по переменной на панели холста и выберите во всплывающем меню одну из опций категорий и итогов.

Чтобы отсортировать категории:

1. Щелкните правой кнопкой мыши по переменной на панели холста, выберите во всплывающем меню пункт **Сортировать категории**, а затем выберите способ сортировки:
   - По значениям
   - По меткам
   - По количеству
   - По меньшему

Итоги

1. Щелкните правой кнопкой мыши по переменной на панели холста, выберите во всплывающем меню **Показать итог**, а затем выбирают, где показать итог:
   - Категория выше
   - Категория ниже

Если выбранная переменная вложена в другую переменную, то итоги будут вставляться для каждой подтаблицы.

Настраиваемые таблицы: Статистики критериев

Функция Статистические критерии предоставляет критерии значимости для пользовательских таблиц

Эти критерии недоступны для таблиц, в которых метки категорий убраны из измерений таблицы по умолчанию, или для вычисленных категорий.

Критерии средних и пропорций по столбцам

Критерии средних по столбцам доступны для количественных переменных. Критерии пропорций по столбцам доступны для категориальных переменных.

Сравнение средних между столбцами

Парные критерии равенства средних по столбцам. У таблицы должна быть категориальная переменная в столбцах и количественная переменная на внутреннем уровне строк. Таблица должна включать среднее в качестве итожащей статистики.

Для отдельных категориальных переменных можно оценить дисперсию для всех категорий или только для сравниваемых категорий. Для переменных множественных ответов дисперсия для критерия средних всегда основана только на сравниваемых категориях.

Сравнить пропорции столбцов

Парные критерии равенства пропорций столбцов. В таблице должна быть по крайней мере одна категориальная переменная и по столбцам, и по строкам. Таблица должна содержать количества или процентные доли по столбцам.

Уровень значимости

Уровень значимости для критериев средних и пропорций по столбцам.

- Значение должно быть от 0 до 1.
- Если вы задаете для уровня значимости, буквы верхнего регистра используются для идентификации значений значимости, не превосходящих меньшего уровня. Буквы нижнего регистра используются для идентификации значений значимости, не превосходящих большего уровня.
- Если вы выбрали **Использовать нижние индексы в стиле APA**, второе значение игнорируется.

Настроить p-значения на множественность сравнений

Поправка **Бонферрони** применяется для групповой вероятности ошибки (family-wise error rate, FWER). Метод **Беньямини-Хохберга** служит для контроля доли ложных отклонений гипотез (false discovery rate, FDR). Эти методы менее консервативны по сравнению с поправкой Бонферрони.

Вывести значимые различия

Для критериев средних и пропорций по столбцам вы можете вывести значимые результаты в отдельной таблице или в основной таблице.

В отдельной таблице

Результаты применения критериев значимости показаны в отдельной таблице. Если отличия двух значений существенны, ячейка, соответствующая большему значению, содержит ключ, который идентифицирует столбец с меньшим значением.

Вывести значения значимости

Значения значимости выводятся в скобках после каждого значения ключа в ячейке. Эта опция доступна, только когда результаты значимости выводятся в отдельной таблице.

В главной таблице

Результаты критерия значимости выводятся в основной таблице. Каждая категория столбцов в таблице идентифицируется буквенным ключом. Для каждой значимой пары ключ категории с меньшим средним или пропорцией по столбцу выводится в категории с большим средним или пропорцией по столбцу.

- Когда вы ставите указатель на ключ в ячейке метки столбца в опорной таблице, все ячейки в таблице с этим ключом значимости выделяются. Для таблицы с несколькими переменными в измерении столбца выделяются только ячейки в этой подтаблице.
- Чтобы выбрать все ячейки в таблице (или подтаблице) с этим ключом значимости, щелкните правой кнопкой мыши по ячейке метки столбца и выберите **Выбрать > Выбрать все ячейки с этим ключом значимости**.

Использовать нижние индексы в стиле APA

Идентифицировать значимые различия при помощи форматирования в стиле APA, где используются буквенные нижние индексы. Если два значения существенно различны, для этих значений выводятся различные буквенные нижние индексы. Эти нижние индексы не являются сносками. Когда действует этот параметр, заданный стиль сносок в текущем шаблоне таблиц переопределяется, и сноски отображаются в виде чисел в верхнем индексе. Чтобы выделить все ячейки в одной строке с одним и тем же ключом значимости, щелкните правой кнопкой мыши по ячейке с ключом значимости и выберите **Выбрать ячейки с подобной значимостью**.

Проверка независимости (критерий хи-квадрат)

Критерий хи-квадрат независимости для таблиц, в которых по столбцам и по строка существует хотя бы одна категориальная переменная.

Использовать подытоги вместо подытоженных категорий

Каждый подытог заменяет свои категории для проверки значимости. Если снят, то проверки значимости применяются лишь к тем подытогам, категории которых скрыты.

Включить в проверки переменные множественных ответов

Категории наборов множественных ответов включаются в критерии значимости. Если опция не выбрана, категории наборов множественных ответов не включаются в критерии значимости.

Файлы для примеров

Файлы примеров, установленные вместе с продуктом, находятся во вложенной папке *Образцы* папки, в которой установлена система. В подкаталоге Samples есть отдельная папка для каждого из следующих языков: английский, французский, немецкий, итальянский, японский, корейский, польский, русский, упрощенный китайский, испанский и традиционный китайский.

Не все файлы примеров доступны на каждом языке. Если файл примера не доступен на конкретном языке, языковая папка содержит версию этого файла примера на английском языке.

Описание

Ниже дано краткое описание файлов, используемых в различных примерах в данной документации.

- **accidents.sav.** Это файл с гипотетическими данными, относящимися к страховой компании, изучающей факторы риска возраста и пола для дорожно-транспортных происшествий в заданном районе. Каждое наблюдение соответствует перекрестной классификации возрастной категории и пола.
- **adl.sav.** Это файл гипотетических данных, относящихся к усилиям по определению преимуществ предлагаемого вида лечения больных с инсультом. Врачи разбили случайным образом пациентов-женщин с инсультом на две группы. Первая группа получала стандартную физиотерапию, а вторая получала дополнительную эмоциональную терапию. Через три месяца после лечения была произведена оценка в виде порядковых переменных способности каждой пациентки выполнять действия повседневной жизни.
- **advert.sav.** Это файл гипотетических данных, относящихся к усилиям розничных торговцев установить зависимость между деньгами, затрачиваемыми на рекламу, и результатами продаж. Для этого они собрали данные о прошлых продажах и связанные с ними затраты на рекламу
- **aflatoxin.sav.** Это файл гипотетических данных, относящихся к тестированию урожая кукурузы на афлатоксин, яд, концентрация которого колеблется в больших пределах между партиями урожая и в пределах одной партии урожая. Обработчик зерен получил 16 образцов от каждой из 8 партий урожая и измерил уровни афлатоксина в частях на миллиард (PPB).
- **anorectic.sav.** Работая над стандартизацией симптоматологии аноректичного/булимичного поведения, исследователи [1] исследовали 55 подростков с известными нарушениями питания. Каждый пациент обследовался четыре раза за четыре года, что дало в сумме 220 обследований. При каждом обследовании

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.

пациентов оценивали по каждому из 16 симптомов. Оценки симптомов пропущены для пациента 71 на 2 обследовании, пациента 76 на втором обследовании и пациента 47 на третьем обследовании, что дает 217 валидных наблюдений.

- **anticonvulsants.sav.** В медицинских исследованиях с помощью обобщенной линейной смешанной модели можно установить, например, может ли новый противосудорожный препарат снизить частоту эпилептических припадков. Повторные измерения на одном и том же пациенте обычно обнаруживают высокую положительную корреляцию, поэтому в данном случае подходит смешанная модель со случайными эффектами. Поле назначения (число припадков) принимает положительные целые значения, то есть подойдет обобщенная линейная смешанная модель с распределением Пуассона и логарифмической связью.
- **bankloan.sav.** Это файл гипотетических данных, относящихся к усилиям банка снизить частоту неуплат задолженностей по кредитам. Файл содержит финансовую и демографическую информацию о 850 бывших и потенциальных клиентах. Первые 700 наблюдений - это клиенты, которые ранее получали кредиты. Последние 150 наблюдений представляют собой потенциальных клиентов, которых банку нужно классифицировать как хорошие или плохие риски кредитования.
- **bankloan_binning.sav.** Это файл гипотетических данных, содержащий финансовую и демографическую информацию о 5000 бывших клиентах.
- **bankloan_cs.sav.** Это файл гипотетических данных, по которым банк рассчитывает идентифицировать характеристики клиентов, предрасположенных к задержке в погашении долгов, а также использовать эти характеристики, чтобы оценить степень риска при выдаче кредитов.
- **bankloan_cs_noweights.sav.** Это файл гипотетических данных, по которым банк рассчитывает идентифицировать характеристики клиентов, предрасположенных к задержке в погашении долгов, а также использовать эти характеристики, чтобы оценить степень риска при выдаче кредитов. Веса выборки не включены в файл.
- **behavior.sav.** В классическом примере [2]52 студентов попросили оценить комбинации 15 ситуаций и 15 поведений по 10-балльной шкале от 0="очень подходит" to 9="абсолютно не подходит." С усреднением по индивидуумам значения принимались как различия.
- **behavior_ini.sav.** Этот файл данных содержит исходную конфигурацию двумерного решения для *behavior.sav* .
- **brakes.sav.** Это файл гипотетических данных, относящихся к управлению качеством на заводе, выпускающем дисковые тормоза для высококлассных автомобилей. Файл данных содержит измерения диаметра 16 дисков от каждой из 16 производственных машин. Целевой диаметр для дисков составляет 322 миллиметра.
- **breakfast.sav.** В классическом исследовании [3], 21 магистров бизнеса - выпускников школы Уортон и их супруг попросили оценить 15 объектов завтрака в порядке предпочтения, от 1="наиболее предпочтительного" до 15="наименее предпочтительного." Их предпочтения регистрировались по шести различным сценариям, от "Общего предпочтения" до "Легкая закуска, только с напитками."
- **breakfast-overall.sav.** Этот файл данных содержит предпочтения объектов завтрака для первого сценария, только "Общее предпочтение".
- **broadband_1.sav.** Это файл гипотетических данных, содержащий количество абонентов национальных широкополосных услуг. Этот файл данных содержит номера ежемесячных абонентов по 85 регионам за четырехлетний период.
- **broadband_1.sav.** Этот файл данных идентичен файлу *broadband_1.sav* , но содержит данные для трех дополнительных месяцев.
- **cable_survey.sav.** Руководители провайдера кабельной связи для телевидения, телефонии и интернета хотят лучше узнать своих потенциальных заказчиков. Они проводят опрос среди 2000 жителей в своем районе предоставления услуг и предлагают им выбрать вариант ответа для каждого из трех видов связи: (1) не пользуются такими услугами; (2) подписаны на услуги других провайдеров; (3) пользуются

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.

услугами компании. При этом опросе дополнительно собирается некоторая демографическая информация, такая как пол, категория возрастной группы (4 уровня), категория образования (3 уровня), категория дохода (3 уровня), категория типа жилья (4 уровня), категория времени проживания по текущему адресу (3 уровня), число жителей в доме и так далее.

- **car_insurance_claims.sav.** Набор данных, представленный и проанализированный в другой работе [4], касается исков за повреждение автомобилей. Среднюю сумму иска можно смоделировать как имеющую гамма-распределение, использующее функцию инверсной связи для соотношения среднего зависимой переменной с линейной комбинацией возраста обладателя страховки, типа автомобиля и возраста автомобиля. Количество зарегистрированных исков можно использовать в качестве масштабирующего веса.
- **car_sales.sav.** Этот файл данных содержит гипотетические оценки продаж, прайс-листы и технические характеристики различных марок и моделей автомобилей. Прайс-листы и технические характеристики были получены поочередно от *edmunds.com* и сайтов производителей.
- **car_sales_uprepared.sav.** Это измененная версия файла *car_sales.sav* , в которую не входят все трансформированные версии полей.
- **carpet.sav.** В известном примере [5] компания, заинтересованная в маркетинге нового средства чистки ковров, хочет изучить влияние пяти факторов на потребительские предпочтения - дизайн упаковки, фирменное название, цена, знак *Идеальный дом* и гарантия возврата средств. Есть три уровня факторов для дизайна упаковки, каждый из которых отличается размещением щетки аппликатора; три фирменных названия (*K2R* , *Glory* и *Bissell*); три ценовых уровня и два уровня (либо да, либо нет) для каждого из последних двух факторов. Десять потребителей ранжировали 22 профиля, определяемых этими факторами. Переменная *Предпочтение* содержит ранг среднего ранжирования для каждого профиля. Низкое ранжирование соответствует высокому предпочтению. Эта переменная отражает общую меру предпочтения для каждого профиля.
- **carpet_prefs.sav.** Этот файл данных основан на том же примере, что и описанный для *carpet.sav* , но он содержит фактическое ранжирование, полученное от каждого из 10 потребителей. Потребителей попросили ранжировать 22 профилей изделия от наиболее предпочтительного до наименее предпочтительного. Переменные с *PREF1* по *PREF22* содержат идентификаторы ассоциированных профилей, как они определены в *carpet_plan.sav* .
- **catalog.sav.** Этот файл данных содержит цифры гипотетических ежемесячных продаж трех товаров, продаваемых компанией, торгующей по каталогу. Сюда также включены данные для пяти возможных переменных предикторов.
- **catalog_seasfac.sav.** Этот файл данных такой же, что и *catalog.sav* , за исключением того, что в него добавлен набор сезонных факторов, рассчитанных с помощью процедуры Сезонная декомпозиция, вместе с сопровождающими переменными дат.
- **cellular.sav.** Это файл гипотетических данных, относящихся к компании сотовой связи, старающейся уменьшить отток абонентов. Баллы предрасположенности к оттоку абонентов применяются к учетным записям в ранге от 0 до 100. Учетные записи, набирающие 50 и выше баллов, возможно, собираются сменить провайдера.
- **ceramics.sav.** Это файл гипотетических данных, относящихся к усилиям производителя определить, имеет ли новый высококачественный сплав более высокую жаростойкость, чем стандартный сплав. Каждое наблюдение представляет собой отдельный тест одного из сплавов. Регистрируется температура, при которой отказывает подшипник.
- **cereal.sav.** Это файл гипотетических данных, относящихся к опросу 880 людей об их предпочтениях за завтраком, с указанием их возраста, пола, семейного положений и ведут ли они активный образ жизни (исходя из того, делают ли они физические упражнения по крайней мере два раза в неделю). Каждое наблюдение представляет собой отдельного респондента.
- **clothing_defects.sav.** Это файл гипотетических данных, относящихся к процессу управления качеством на швейной фабрике. Из каждой партии, изготовленной на фабрике, инспекторы берут образец одежды и считают количество забракованной одежды.

---

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.

- **coffee.sav.** Этот файл данных относится к воспринимаемым образам шести марок кофе со льдом [6] . Для каждого из 23 атрибутов образа кофе со льдом люди подбирали все марки, которые описывал данный атрибут. Шесть марок были обозначены как AA, BB, CC, DD, EE и FF, чтобы сохранить конфиденциальность.
- **contacts.sav.** Это файл гипотетических данных, относящихся списков контактов для группы корпоративных торговых представителей по продаже компьютеров. Каждое контактное лицо классифицировалось по отделению компании, в котором этот человек работает, и по должности, которую он занимает. Также регистрировался объем последней продажи, время с момента последней продажи и величина компании контактного лица.
- **credit_card.sav.** Гипотетическое изучение использование кредитных карт по ежемесячным тратам за два года, причем затраты с первичной карты разбиты по типу транзакций (продовольственные товары, другая розница, развлечения, путешествия и другое). Каждая запись в наборе данных соответствует данному месяцу затрат и типу транзакций, поэтому по каждой карте требуется 2 года × 12 месяцев в год × 5 типов транзакций = 120 записей.
- **creditpromo.sav.** Это файл гипотетических данных, относящихся к усилиям банка снизить частоту невозврата кредитов. Для этого были случайным образом отобраны 500 владельцев кредитных карточек. Половина из них получила рекламное объявление с предложением пониженной процентной ставки на покупки, которые будут сделаны в течение следующих трех месяцев. Половина получила стандартное сезонное рекламное объявление.
- **cross_sell.sav.** У компании заказов по почте есть книжный клуб и клуб любителей компакт-дисков. Каждый месяц компания делает специальные предложения для членов клубов. Компания хочет создать модель полных продаж по ежемесячным специальным предложениям на основании всех покупок книг и компакт-дисков и типа предложений для членов клубов. В этой ситуации подходит модель двухэтапной регрессии методом наименьших квадратов, так как потраченные на специальные предложения деньги - это деньги, не потраченные на книги или компакт-диски, поэтому существует контур обратной связи между откликом и этими двумя предикторами.
- **customer_dbase.sav.** Это файл гипотетических данных, относящихся к усилиям компании использовать информацию в своем хранилище данных, чтобы сделать особые предложения клиентам, которые вероятнее всего откликнутся. Подмножество базы данных клиентов было выбрано случайным образом, этим клиентам было сделано особое предложение и зарегистрирована их реакция.
- **customer_information.sav.** Это файл гипотетических данных, содержащий почтовые сведения о пользователе, например имя и адрес.
- **customer_subset.sav.** Подмножество из 80 наблюдений из файла *customer_dbase.sav*.
- **debate.sav.** Это файл гипотетических данных, относящихся к парным ответам на опрос участников политических дебатов до и после этих дебатов. Каждое наблюдение представляет собой отдельного респондента.
- **debate_aggregate.sav.** Этот файл гипотетических данных, агрегирующий ответы в *debate.sav* . Каждое наблюдение представляет собой перекрестную классификацию предпочтения до и после дебатов.
- **demo.sav.** Это файл гипотетических данных, относящихся к базе данных клиентов, сделавших покупки, с целью рассылки ежемесячных предложений. Регистрируется, отреагировал ли клиент на предложение, а также различная демографическая информация.
- **demo_cs_1.sav.** Это файл гипотетических данных, относящихся к первому этапу усилий компании составить базу данных с информацией опросов. Каждое наблюдение соответствует отдельному городу, также регистрируется идентификация региона, провинции, района и города.
- **demo_cs_2.sav.** Это файл гипотетических данных, относящихся ко второму этапу усилий компании составить базу данных с информацией опросов. Каждое наблюдение представляет собой отдельную семью из городов, отобранных на первом этапе. Также регистрируется идентификация региона, провинции, района, города, городского района. Также включена информация о выборках на первых двух этапах проекта.

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.

- **demo_cs.sav.** Это файл гипотетических данных, содержащий информацию опроса, собранную с помощью проекта комплексной выборки. Каждое наблюдение соответствует отдельной семье. Регистрируется также различная демографическая информация и информация о выборках.
- **diabetes_costs.sav.** Это файл гипотетических данных, который содержит информацию, поддерживаемую страховой компанией о держателях полисов, страдающих диабетом. Каждое наблюдение соответствует одному держателю полиса.
- **dietstudy.sav.** Этот файл гипотетических данных содержит результаты исследования "диеты Стилмана" [7]. Каждое наблюдение соответствует отдельному субъекту и регистрирует его или ее вес до и после в фунтах и уровень триглицерида в мг/100 мл.
- **dmdata.sav.** Это гипотетические данные, которые содержат сведения о демографии и покупках для компании, занимающейся прямым маркетингом. Файл *dmdata2.sav* содержит информацию для подмножества адресов, по которым получена пробная рассылка, а *dmdata3.sav* содержит информацию об остальных адресах, по которым пробная рассылка не получена.
- **dvdplayer.sav.** Это файл гипотетических данных, относящихся к разработке нового DVD-плеера. Маркетинговая команда собрала с помощью прототипа данные о целевой группе. Каждое наблюдение соответствует отдельному отслеживаемому пользователю и фиксирует некоторую демографическую информацию о нем и его ответах на вопросы о прототипе.
- **Employee data.sav.** Это файл гипотетических данных, содержащий конкретную информацию о нанятых сотрудниках (уровень образования, текущая зарплата, предыдущий опыт и так далее).
- **german_credit.sav.** Этот файл данных взят из набора данных "German credit" в репозитории баз данных машинного обучения [8] Калифорнийского университета, город Ирвин.
- **grocery_1month.sav.** Этот файл гипотетических данных - файл данных *grocery_coupons.sav* со "свернутыми" еженедельными покупками, так что каждое наблюдение соответствует отдельному клиенту. Некоторые еженедельно изменявшиеся переменные в результате исчезают, и теперь регистрируемая потраченная сумма представляет собой сумму затрат на покупки, сделанные в течение четырех недель исследования.
- **grocery_coupons.sav.** Это гипотетический файл данных, содержащий данные опроса, проведенного сетью магазинов бакалейных товаров, заинтересованной в покупательских привычках своих клиентов. Каждый покупатель отслеживался в течение четырех недель, и каждое наблюдение соответствует отдельной паре клиент-неделя и регистрирует информацию о том, где и как покупает клиент, включая сумму, потраченную на бакалейные товары в течение этой недели.
- **guttman.sav.** Белл [9] представил таблицу для иллюстрации возможных социальных групп. Гуттман [10] использовал часть этой таблицы, в которой пять переменных, описывающих такие вещи, как социальное взаимодействие, чувство принадлежности к группе, физическая близость членов группы и формальность отношений были пересечены с семью теоретическими социальными группами, включая толпы (например, публика на футбольном матче), аудитории (например, люди в театре или на лекции), публику (например, аудитория газет или телевидения), сборища (подобные толке, но с гораздо более сильным взаимодействием), первичные группы (близкие), вторичные группы (добровольцы) и современное общество (слабая конфедерация, возникающая в результате физической близости и потребности в специализированных услугах).
- **health_funding.sav.** Это файл гипотетических данных, содержащий данные о финансировании здравоохранения (сумма на 100 человек населения), заболеваемости (заболеваемость на 10000 человек населения) и посещениях провайдеров медицинских услуг (частота на 100000 человек). Все примеры представляют разные города.

---

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.

8. Blake, C. L., and C. J. Merz. 1998. "UCI Repository of machine learning databases." Доступна в 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.

- **hivassay.sav.** Это файл гипотетических данных, относящихся к усилиям фармацевтической лаборатории разработать быстрый тест для обнаружения ВИЧ-инфекции. Результаты теста - восемь все более насыщенных оттенков красного, причем более темный цвет указывает на большую вероятность инфекции. Лабораторные исследования проводились на 2000 образцах крови, половина из которых была ВИЧ-инфицирована, а другая половина - чистая.
- **hourlywagedata.sav.** Этот файл гипотетических данных, относящихся к почасовой оплате медицинских сестер, работающих в офисе и госпитале и имеющих различный опыт.
- **insurance_claims.sav.** Этот файл гипотетических данных, относящихся к страховой компании, которой необходимо построить модель для того, чтобы отмечать подозрительные, потенциально обманные требования. Каждое наблюдение представляет собой отдельное требование.
- **insure.sav.** Это гипотетический файл данных, относящихся к страховой компании, которая изучает факторы риска того, что клиент подаст заявку на выплату по 10-летнему контракту страхования жизни. Каждое наблюдение в файле представляет собой пару контрактов, в одном из которых зарегистрирована заявка, а в другом - нет, подобранных по полу и возрасту.
- **judges.sav.** Это файл гипотетических данных, относящихся к баллам, присужденным опытными судьями (плюс один энтузиаст) 300 гимнастическим упражнениям. Каждая строка представляет собой отдельное упражнение. Судьи видели одни и те же упражнения.
- **kinship_dat.sav.** Розенберг и Ким [11] проанализировали 15 терминов родства (тетя, брат, двоюродный брат, дочь, отец, внучка, дедушка, бабушка, внук, племянник, племянница, сестра, сын, дядя). Они попросили четыре группы студентов колледжа (две женских и две мужских) рассортировать эти термины на основе подобия. Две группы (одну женскую и одну мужскую) попросили отсортировать два раза, причем вторая сортировка основывалась на критерии, отличающемся от критерия первой сортировки. Таким образом, всего получено шесть “источников”. Каждый источник соответствует матрице близости размером 15 x 15, значения в ячейках которой равно разности между количеством людей в источнике и числом случаев, когда объекты оказывались в одном разделе в этом источнике.
- **kinship_ini.sav.** Этот файл данных содержит начальную конфигурацию двумерного решения для *behavior.sav* .
- **kinship_var.sav .** Этот файл данных содержит независимые переменные *gender*, *gener* (поколение) и *degree* (степень разделения), которые можно использовать для интерпретации измерений в решении для *kinship_dat.sav*. В частности, их можно использовать для ограничения пространства решений линейной комбинацией этих переменных.
- **marketvalues.sav.** Этот файл данных относится к продаже жилья в новостройках Алгонкин, Иллинойс за период после 1999–2000. Сведения об этих продажах общедоступны.
- **nhis2000_subset.sav.** Проект National Health Interview Survey (NHIS) - это большое, основанное на населении исследование гражданского населения США. Опросы проводились лицом к лицу в национально репрезентативной выборке домохозяйств. Для членов каждой семьи получали демографическую информацию и наблюдения за поведением в отношении здоровья и состоянием здоровья. Этот файл данных содержит подмножество информации из исследования 2000 года. Национальный центр статистики здравоохранения. National Health Interview Survey, 2000. Файл данных и документация для общественного пользования. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Доступ получен в 2003 году.
- **ozone.sav.** Данные включают 330 наблюдений шести метеорологических переменных для прогнозирования концентрации озона по оставшимся переменным. Среди прочих, предыдущие исследователи [12], [13], обнаружили среди этих переменных нелинейности, затрудняющие применение стандартных методов регрессионного анализа.

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.

- **pain_medication.sav.** Этот файл гипотетических данных содержит результаты клинических испытаний противовоспалительного препарата для лечения хронической боли в суставах. Особый интерес представляет время, необходимое для того, чтобы лекарство начало действовать, и как это сравнимо с существующим препаратом.
- **patient_los.sav.** Этот файл гипотетических данных содержит истории болезни пациентов, поступивших в госпиталь с подозрением на инфаркт миокарда (ИМ или "сердечный приступ"). Каждое наблюдение соответствует отдельному пациенту и фиксирует много переменных, связанных с пребыванием пациента в госпитале.
- **patlos_sample.sav.** Этот файл гипотетических данных содержит истории болезни выборки пациентов, получавших тромболитики во время лечения инфаркта миокарда (ИМ или "сердечного приступа"). Каждое наблюдение соответствует отдельному пациенту и фиксирует много переменных, связанных с пребыванием пациента в госпитале.
- **poll_cs.sav.** Это файл гипотетических данных, относящихся к усилиям специалистов по опросам определить уровень общественной поддержки законопроекта перед тем, как он пройдет утверждение. Наблюдения соответствуют зарегистрированным избирателям. Каждое наблюдение соответствует округу, городу и окружению, в котором живет избиратель.
- **poll_cs_sample.sav.** Этот файл гипотетических данных содержит выборку избирателей, перечисленных в *poll_cs.sav* . Выборка осуществлялась в соответствии с проектом, заданным в файле плана *poll.csplan* , и данный файл данных содержит вероятности включения и веса выборки. Заметьте, однако, что из-за того, что в плане выборки используется метод вероятности, пропорциональной размеру (PPS), есть также файл, содержащий вероятности совместной выборки (*poll_jointprob.sav*). Дополнительные переменные, соответствующие демографическим данным избирателя и его мнению о предлагаемом законопроекте, собирались и добавлялись в файл данных после того, как была сделана выборка.
- **property_assess.sav.** Это файл гипотетических данных, относящийся к усилиям асессора округа поддерживать обновление оценок недвижимости при ограниченных ресурсах. Наблюдения соответствуют недвижимости, проданной в округе за последний год. Каждое наблюдение в файле данных регистрирует участок, на котором расположена недвижимость, имя асессора, который посещал недвижимость последним, время, прошедшее с момента последней оценки, оценку, сделанную в то время и продажную стоимость недвижимости.
- **property_assess_cs.sav.** Это файл гипотетических данных, относящийся к усилиям асессора штата поддерживать обновление оценок недвижимости при ограниченных ресурсах. Наблюдения соответствуют недвижимости в штате. Каждое наблюдение в файле данных регистрирует округ, участок и окружение, в котором расположена недвижимость, время, прошедшее с момента последней оценки и сделанную тогда оценку.
- **property_assess_cs_sample.sav.** Этот файл гипотетических данных содержит выборку недвижимости, перечисленной в *property_assess_cs.sav* . Выборка была сделана в соответствии с проектом, заданным в файле плана *property_assess.csplan* , и данный файл данных содержит вероятности включения и веса выборки. Дополнительную переменную *Текущее значение* собирали и добавляли в файл данных после того, как была сделана выборка.
- **recidivism.sav.** Это файл гипотетических данных, относящихся к усилиям государственного судебного исполнительного органа понять уровень рецидивизма в своей области юрисдикции. Каждое наблюдение соответствует лицу, ранее совершившему правонарушение, и регистрирует демографическую информацию о нем, некоторые подробности первого правонарушения и время, прошедшее до повторного ареста, если он произошел в течение двух лет после первого ареста.
- **recidivism_cs_sample.sav.** Это файл гипотетических данных, относящихся к усилиям государственного судебного исполнительного органа понять уровень рецидивизма в своей области юрисдикции. Каждое наблюдение соответствует лицу, ранее совершившему правонарушение, освобожденному после первого ареста в течение июня 2003 года, и регистрирует демографическую информацию о нем, некоторые подробности первого правонарушения и дату второго ареста, если он произошел до конца июня 2006 года. Правонарушители выбирались из выборочных отделений в соответствии с планом выборки, заданным в *recidivism_cs.csplan*; из-за того, что в плане выборки используется метод вероятности, пропорциональной размеру (PPS), есть также файл, содержащий вероятности совместной выборки (*recidivism_cs_jointprob.sav*).

- **rfm_transactions.sav.** Это файл гипотетических данных, содержащий данные о покупках, включая дату покупки, перечень приобретенных товаров, а также сумму покупки для каждой такой транзакции.
- **salesperformance.sav.** Это файл гипотетических данных, относящихся к оценке двух новых курсов по обучению продажам. Шестьдесят сотрудников, разбитых на три группы, получали стандартное обучение. Кроме того, в группе 2 проводилось техническое обучение, а в группе 3 - практические занятия. По окончании курсов обучения каждый сотрудник был протестирован и его оценка зафиксирована. Каждое наблюдение в файле данных представляет собой отдельного обучаемого и регистрирует группу, которой от был назначен, и оценку, которую он получил на экзамене.
- **satisf.sav.** Это файл гипотетических данных, относящихся к опросу об удовлетворении, проведенному компанией розничной продажи в 4 местах расположения магазинов. Всего было опрошено 582 клиента, и каждое наблюдение представляет собой ответы одного клиента.
- **screws.sav.** Этот файл данных содержит информацию о технических характеристиках винтов, болтов, гаек и гвоздей [14].
- **shampoo_ph.sav.** Это файл гипотетических данных, относящихся к управлению качеством на фабрике, производящей средства для ухода за волосами. Через равные промежутки времени измерялись шесть отдельных выходных партий и регистрировался их pH. Диапазон назначения равен 4.5–5.5.
- **ships.sav.** Набор данных, представленный и проанализированный в другой работе [15], относящийся к повреждениям, причиненным грузовым судам волнами. Количество инцидентов можно смоделировать в виде распределения Пуассона при заданных типе судна, времени постройки и продолжительности эксплуатации. Агрегированные месяцы эксплуатации для каждой ячейки таблицы, образованной перекрестной классификацией факторов, дают значения подверженности риску.
- **site.sav.** Это файл гипотетических данных, относящихся к усилиям компании выбрать новые площадки для своего расширяющегося бизнеса. Она наняла двух консультантов, чтобы они по-отдельности оценили площадки. Эти консультанты, в дополнение к подробному отчету, оценили каждую площадку как "хорошую," "подходящую" или "плохую".
- **smokers.sav.** Этот файл данных представляет собой выборку из 1998 National Household Survey of Drug Abuse и является вероятностной выборкой американских домохозяйств. (http://dx.doi.org/10.3886/ICPSR02934) Таким образом, первый шаг в анализе этого файла данных должен состоять во взвешивании, отражающем тенденции совокупности.
- **stocks.sav** Этот файл гипотетических данных содержит биржевые цены и объемы продаж за один год.
- **stroke_clean.sav.** Этот файл гипотетических данных содержит состояние медицинской базы данных после того, как ее почистили с помощью процедур в Statistics Base Edition.
- **stroke_invalid.sav.** Этот файл гипотетических данных содержит первоначальное состояние медицинской базы данных, а также несколько ошибок в записях данных.
- **stroke_survival.** Это файл гипотетических данных, относящихся к срокам дожития для пациентов, выходящих из реабилитационной программы Post-Ischemic Stroke Face a Number of Challenges. Регистрировались пост-инсультное состояние, возникновение инфаркта миокарда, ишемический инсульт или геморрагический инсульт и время события. Выборка усечена слева, поскольку она включает только пациентов, выживших к концу программы пост-инсультной реабилитации.
- **stroke_valid.sav.** Этот файл гипотетических данных содержит состояние медицинской базы данных после проверки значений с помощью процедуры Проверить данные. Она по-прежнему содержит потенциально ненормальные наблюдения.
- **survey_sample.sav.** Этот файл данных содержит данные опроса, включая демографические данные и различные показатели, характеризующие отношение. Он основывается на подмножестве переменных из 1998 NORC General Social Survey, но для демонстрационных целей были изменены некоторые значения данных и добавлены дополнительные фиктивные переменные.
- **tcm_kpi.sav.** Это файл гипотетических данных, который содержит еженедельные значения ключевых показателей эффективности для бизнеса. Он содержит также еженедельные данные для многих контролируемых показателей за тот же период времени.

---

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.

- **tcm_kpi_upd.sav.** Этот файл данных идентичен файлу *tcm_kpi.sav*, но содержит данные для 4 дополнительных недель.
- **telco.sav.** Это файл гипотетических данных, относящихся к усилиям телекоммуникационной компании уменьшить отток абонентов в своей базе данных клиентов. Каждое наблюдение соответствует отдельному клиенту и регистрирует различную демографическую информацию и информацию о пользовании услугами.
- **telco_extra.sav.** Этот файл данных аналогичен файлу данных *telco.sav* , но переменные "срок пребывания" и log-преобразованные затраты клиента были удалены и заменены стандартизованными переменными log-преобразованные затраты клиента.
- **telco_missing.sav.** Этот файл данных является подмножеством файла данных *telco.sav*, но некоторые значения демографических данных были заменены значениями отсутствия.
- **testmarket.sav.** Это файл гипотетических данных, относящихся к планам сети быстрого питания добавить новое блюдо в свое меню. Есть три возможных компании по продвижению нового продукта, и новое блюдо в торговых точках на нескольких выбранных случайным образом рынках. В каждой точке использовался различный метод продвижения, а еженедельные продажи нового блюда регистрировались для первых четырех недель. Каждое наблюдение соответствует отдельной точке-неделе.
- **testmarket_1month.sav.** Этот файл гипотетических данных - файл данных *testmarket.sav* со "свернутыми" еженедельными продажами, причем каждое наблюдение соответствует отдельному расположению. Некоторые еженедельно изменявшиеся переменные в результате исчезают, и теперь регистрируемые продажи представляют собой сумму продаж в течение четырех недель исследования.
- **tree_car.sav.** Это файл гипотетических данных, содержащий демографические данные и данные продажных цен автомобилей.
- **tree_credit.sav.** Это файл гипотетических данных, содержащий демографические данные и данные истории банковских займов.
- **tree_missing_data.sav** Это файл гипотетических данных, содержащий демографические данные и данные хронологии банковских займов с большим количеством пропущенных значений.
- **tree_score_car.sav.** Это файл гипотетических данных, содержащий демографические данные и данные продажных цен автомобилей.
- **tree_textdata.sav.** Простой файл данных только с двумя переменными, предназначенный прежде всего для того, чтобы показать состояние по умолчанию переменных перед назначением уровня измерения и меток переменных.
- **tv-survey.sav.** Это файл гипотетических данных, относящихся опросу, проведенному телестудией о том, нужно ли расширять успешную программу. 906 респондентов спросили, будут ли они смотреть программу при различных условиях. Каждая строка представляет собой отдельного респондента, а каждый столбец - отдельное условие.
- **ulcer_recurrence.sav.** Этот файл содержит частичную информацию из исследования, предназначенного для сравнения эффективности двух методов лечения для предотвращения повторного возникновения язвы желудка. Он представляет собой хороший пример интервал-цензурированных данных и был представлен и проанализирован в другой работе [16].
- **ulcer_recurrence_recoded.sav.** Этот файл реорганизует информацию в *ulcer_recurrence.sav* , чтобы можно было моделировать вероятность события для каждого интервала исследования, а не просто вероятность события в конце исследования. Он был представлен и проанализирован в другой работе [17].
- **verd1985.sav.** Этот файл данных относится к опросу [18]. Регистрировались ответы 15 субъектов на 8 переменных. Эти переменные были разбиты на три набора. В набор 1 входят *age* и *marital*, в набор 2 входят *pet* и *news* и в набор 3 входят *music* и *live*. Переменная *pet* является множественной полиномиальной, *age* является порядковой. Все остальные переменные являются одиночными номинальными.

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.** Это файл гипотетических данных, относящихся к усилиям провайдера Интернет-услуг (ISP) определить влияние вируса на свои сети. Они отследили процент (приблизительный) зараженного e-mail трафика по своим сетям по времени от момента обнаружения до устранения угрозы.
- **wheeze_steubenville.sav.** Это подмножество длительного исследования влияния загрязнения воздуха на здоровье детей [19]. Данные содержат повторяющиеся двоичные измерения состояния свистящего дыхания у детей из Штойбенвилля, штат Огайо, в возрасте 7, 8, 9 и 10 лет, вместе с фиксированной регистрацией того, курила ли мать в течение первого года исследования.
- **workprog.sav.** Это файл гипотетических данных, относящихся к государственной программе работ, которая старается предоставить неимущим людям лучшую работу. Отслеживалась выборка потенциальных участников программы, причем некоторые из них были выбраны для участия в программе, а другие - нет. Каждое наблюдение представляет собой отдельного участника программы.
- **worldsales.sav** Этот гипотетический файл данных содержит товарооборот по континентам и товарам..

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.

Замечания

Эта публикация разрабатывалась для продуктов и услуг, предлагаемых в США. Этот материал может быть доступен от IBM на других языках. Однако для его получения может понадобиться приобрести продукт или версию продукта на нужном языке.

IBM может не предоставлять в других странах продукты, услуги и аппаратные средства, описанные в данном документе. За информацией о продуктах и услугах, предоставляемых в вашей стране, обращайтесь к местному представителю IBM. Ссылки на продукты, программы или услуги IBM не означают и не предполагают, что можно использовать только указанные продукты, программы или услуги IBM. Разрешается использовать любые функционально эквивалентные продукты, программы или услуги, если при этом не нарушаются права IBM на интеллектуальную собственность. Однако ответственность за оценку и проверку работы любого продукта, программы или сервиса, не произведенного корпорацией IBM, лежит на пользователе.

IBM может располагать патентами или рассматриваемыми заявками на патенты, относящимися к предмету данного документа. Предъявление данного документа не предоставляет какую-либо лицензию на эти патенты. Вы можете послать письменный запрос о лицензии по адресу:

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

По поводу лицензий, связанных с использованием наборов двухбайтных символов (DBCS), обращайтесь в отдел интеллектуальной собственности IBM в вашей стране или направьте запрос в письменной форме по адресу:

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

КОРПОРАЦИЯ INTERNATIONAL BUSINESS MACHINES ПРЕДОСТАВЛЯЕТ ДАННУЮ ПУБЛИКАЦИЮ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ЯВНЫХ ИЛИ ПОДРАЗУМЕВАЕМЫХ ГАРАНТИЙ, ВКЛЮЧАЯ, НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ, ПОДРАЗУМЕВАЕМЫЕ ГАРАНТИИ ОТСУТСТВИЯ НАРУШЕНИЙ, КОММЕРЧЕСКОЙ ПРИГОДНОСТИ ИЛИ СООТВЕТСТВИЯ КАКОЙ-ЛИБО КОНКРЕТНОЙ ЦЕЛИ. В некоторых странах для ряда сделок не допускается отказ от явных или предполагаемых гарантий; в таком случае данное положение к вам не относится.

Эта информация может содержать технические неточности и типографские ошибки. В представленную здесь информацию периодически вносятся изменения; эти изменения будут включаться в новые издания данной публикации. Фирма IBM может в любое время без уведомления вносить изменения и усовершенствования в продукты и программы, описанные в этой публикации.

Любые ссылки в данной информации на сайты, не принадлежащие IBM, приводятся только для удобства и никоим образом не означают поддержки этих сайтов. Материалы на этих сайтах не входят в число материалов по данному продукту IBM, и весь риск пользования этими сайтами несете вы сами.

IBM может использовать или распространять предоставленную вами информацию любым способом, как фирма сочтет нужным, без каких-либо обязательств перед вами.

Если обладателю лицензии на данную программу понадобится информация о возможности: (i) обмена данными между независимо разработанными программами и другими программами (включая данную) и (ii) совместного использования таких данных, он может обратиться по адресу:

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

Такая информация может быть доступна при соответствующих условиях и соглашениях, включая в некоторых случаях взимание платы.

Описанную в данном документе лицензионную программу и все прилагаемые к ней лицензированные материалы IBM предоставляет на основе положений Соглашения между IBM и Заказчиком, Международного Соглашения о Лицензиях на Программы IBM или любого эквивалентного соглашения между IBM и заказчиком.

Упоминаемые данные о производительности и примеры клиентов представлены только для иллюстративных целей. Фактические результаты производительности могут быть иными в зависимости от определенных конфигураций и конкретных условий.

Информация, касающаяся продуктов других компаний (не IBM) была получена от поставщиков этих продуктов, из опубликованных ими заявлений или из прочих общедоступных источников. IBM не проводила тестирования этой продукции и не может подтвердить или опровергнуть информацию о точности ее работы и совместимости, а также другие заявления относительно продуктов других производителей (не IBM). Вопросы относительно возможностей продуктов других компаний (не IBM) следует адресовать поставщикам этих продуктов.

Утверждения, касающиеся намерений и планов IBM, могут быть изменены без предварительного предупреждения; они приведены здесь только для обозначения целей и задач IBM.

Эти сведения содержат примеры данных и отчетов, используемых в повседневных деловых операциях. Чтобы проиллюстрировать их настолько полно, насколько это возможно, данные примеры включают имена индивидуумов, названия компаний, брендов и продуктов. Все эти имена и названия вымышлены и любое их сходство с реальными именами и названиями компаний полностью случайно.

ЛИЦЕНЗИЯ НА КОПИРОВАНИЕ:

Эта информация содержит примеры исходных текстов прикладных программ, которые иллюстрируют приемы программирования на различных платформах. Разрешается копировать, изменять и распространять эти примеры программ в любой форме без оплаты фирме IBM для целей разработки, использования, сбыта или распространения прикладных программ, соответствующих интерфейсу прикладного программирования операционных платформ, для которых эти примера программ написаны. Эти примеры не были всесторонне проверены во всех возможных условиях. Поэтому IBM не может гарантировать их надежность, пригодность и функционирование. Примеры программ предоставляются "КАК ЕСТЬ", без каких-либо гарантий. IBM не несет никакой ответственности за какой либо ущерб, причиненный в результате использования этих программ.

Каждая копия или каждая часть этих примеров программ или работы, основанной на них, должна содержать следующее замечание об авторских правах:

Товарные знаки

IBM, логотип IBM, и ibm.com являются товарными знаками или зарегистрированными товарными знаками компании International Business Machines Corp., зарегистрированными во многих странах мира. Прочие наименования продуктов и услуг могут быть товарными знаками, принадлежащими IBM или другим компаниям. Текущий список товарных знаков IBM можно найти в Интернете в разделе "Copyright and trademark information" ("Информация об авторских правах и товарных знаках") по адресу www.ibm.com/legal/copytrade.shtml.

Adobe, логотип Adobe, PostScript и логотип PostScript являются либо зарегистрированными товарными знаками, либо товарными знаками корпорации Adobe Systems в Соединенных Штатах и/или других странах.

Intel, логотип Intel, Intel Inside, логотип Intel Inside, Intel Centrino, логотип Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium и Pentium являются товарными знаками или зарегистрированными товарными знаками компании Intel или ее дочерних компаний в Соединенных Штатах и других странах.

Linux является зарегистрированным товарным знаком Linus Torvalds в Соединенных Штатах и других странах.

Microsoft, Windows, Windows NT и логотип Windows являются товарными знаками корпорации Microsoft в Соединенных Штатах и других странах.

UNIX является зарегистрированным товарным знаком The Open Group в Соединенных Штатах и других странах.

Java и все основанные на Java товарные знаки и логотипы являются товарными знаками Oracle и/или его филиалов.

Индекс

Напечатано в Дании