СЕРВУЛЯ Ф. студ. групп. МЭД 08
Научн. руков.: Овсянников В.П., к. т. н., доц.
ГВУЗ "Донецкий национальный технический университет"
г. Донецк

АЛГОРИТМ АНАЛИЗА ДАННЫХ ХАРАКТЕРИЗУЮЩИХ РАБОТУ УГОЛЬНОГО ПРЕДПРИЯТИЯ СРЕДСТВАМИ ПРОГРАММНОГО КОМПЛЕКСА SPSS

Приведен алгоритм применения программного комплекса SPSS для статистического анализа данных, характеризующих работу угольного предприятия.

Актуальность. Программный комплекс SPSS развивается уже на протяжении 40 лет. Основным его достоинством является самый широкий охват существующих статистических методов. Известно, что SPSS предоставляет огромные возможности анализа и прогнозирования не только в сфере психологии, социологии, биологии и медицины, но и в области маркетинговых исследований и управлении качеством продукции.

Программа SPSS позволяет анализировать данные, строить прогнозы, оценивать взаимосвязи и зависимости. SPSS позволяет сделать анализ сезонности, выявлять степень влияния различных факторов — таких, как затраты на рекламу, активность продавцов, активность конкурентов и т.п., на уровень продаж. Поэтому использование программного комплекса SPSS для анализа работы угольного предприятия позволяет получить дополнительную информацию, которую невозможно обнаружить, применяя для обработки данных офисные программы, например Excel [1].

Цель исследований: получить алгоритм анализа данных характеризующих работу угольного предприятия и оценить эффективность применения для этих целей программного комплекса SPSS.

Основная часть. Пример таблицы данных характеризующих работу угольного предприятия приведен в [2].

Определение общих параметров выборки средствами программного комплекса SPSS состоит из ряда взаимосвязанных шагов. Это:

- определение реального количества данных;
- определение структуры выборки;
- установление доверительного уровня статистической надежности выборки;
- расчет статистической ошибки и определение репрезентативности выборки.

Для решения всех задач по анализу данных составляется схема кодировки таблицы.

Схема кодировки представляет собой таблицу соответствия вопросов и вариантов ответа анкеты внутреннему представлению переменных в базе данных SPSS. Впоследствии ввод данных в компьютер и их кодирование производятся согласно данной формализованной структуре. Различные типы данных кодируются в схеме кодировки (и в базе данных SPSS) по-разному. Существует три основных типа кодирования данных.

1. Одновариантные данные, у которых есть только один вариант значений, кодируются одной переменной (например, ql). Тип шкалы в данном случае может быть любым.

2. Многовариантные данные, кодируются несколькими одновариантными переменными (например, q3_l, q3_2). Тип шкалы одновариантных переменных может быть только номинальным (дихотомическим).

3. Данные с интервальной шкалой (для числовых данных, например q5_t), либо номинальным (для нечисловых данных, например q4_t).

Все эти типы данных применяются при кодировании таблицы из [2].

Ввод данных в компьютер является четвертым шагом первого (подготовительного) этапа статистического анализа данных. Он неразрывно связан со следующим шагом — кодированием переменных. Существуют две взаимосвязанные и взаимообусловленные процедуры.

Известны три основных способа формирования базы данных в формате SPSS (перечислены в порядке убывания популярности).

1. Импорт базы данных из других программных источников (Microsoft Access, Microsoft Excel, текстовых файлов и других).

2. Ввод данных непосредственно в SPSS при помощи специализированного программного обеспечения (SPSS Data Entry).

3.Ручной ввод данных в SPSS.

Поскольку данные о работе угольного предприятия представлены в формате Microsoft Excel то импорт базы данных является наиболее естественной процедурой.

Следует отметить, что SPSS поддерживает импорт из любых источников данных, совместимых с технологией ODBC (соответствующие драйверы для них должны быть предварительно установлены в Microsoft Windows). Например, чтобы добавить возможность импорта из базы данных Microsoft Paradox (файлы типа *.db), необходимо щелкнуть на кнопке Add Data Source в диалоговом окне Database Wizard. На экране появится стандартное окно Microsoft Windows Администратор источников данных ODBC. В этом диалоговом окне представлен список уже установленных в SPSS источников данных. Чтобы добавить новый источник, отсутствующий в данном перечне, следует щелкнуть на кнопке Добавить.

После того как в файл SPSS помещена таблица с данными по

исследованию, следует перейти к очередному этапу формирования базы данных — кодированию переменных.

Если данные вводятся в SPSS методом импорта, то будут получены только имена переменных и их значения. В этом случае кодирование переменных является обязательным шагом и должно проводиться сразу после процедуры импорта

Следующим шагом будет анализ различий. Цель анализа различна — выявление групп шахт, статистически значимо различающихся между собой. Все статистические процедуры, относящиеся к группе процедур, которые позволяют выявить такие различия (t-тесты и дисперсионный анализ), сравнивают респондентов на основании средних значений переменных. Иными словами, провести различие можно на основании двух или более числовых переменных.

Далее необходимо выполнить дисперсионный анализ. Как известно, при анализе данных маркетинговых исследований достаточно сравнить только две группы данных, то есть установить различия между анализируемыми шахтами. Однако, часто у исследователей возникает необходимость проанализировать не две, а три или более категории признаков. В этом случае следует прибегнуть к использованию дисперсионного анализа, который позволяет анализировать одновременно любое число групп.

При помощи ассоциативного анализа становится возможным анализировать данные не только по отдельности, а в зависимости от других данных. Этот вид анализа иногда называют построением разрезов, поскольку он позволяет определить не только наличие связи между данными, но и силу связи между переменными и то, каким образом ведет себя одна переменная при изменении другой (возрастает или убывает).

В процессе ассоциативного анализа выявляются следующие типы зависимостей.

■ **Немонотонные** зависимости свидетельствуют только о наличии определенной связи между двумя переменными, но не позволяют судить о направлении или силе связи. Пример немонотонной зависимости: мужчины в основном покупают рыбные консервы в продовольственных магазинах, а женщины — на рынках.

■ **Монотонные** зависимости — это зависимости, по которым можно узнать не только наличие, но и направление связи. Пример монотонной зависимости: мужчины покупают пиво чаще, чем женщины. Монотонные зависимости бывают двух видов:

возрастающие — первая переменная возрастает при возрастании второй;

убывающие — первая переменная убывает при возрастании второй.

■ **Линейные** зависимости характеризуются уравнением функции y = a + Ьx (график линейной функции). Связь между двумя переменными в

данном случае является линейной, то есть на основании этой зависимости мы можем сказать, насколько изменится одна переменная при изменении второй.

■ **Нелинейные**. Примерами нелинейных связей между двумя переменными являются: экспоненциальная, логарифмическая, степенная, полиномиальная зависимости — то есть в данном случае связь присутствует и изменяется по какому-либо известному математическому закону.

Следующий шаг алгоритма — корреляционный анализ Корреляционный анализ предназначен для выявления наличия, а также определения направления и силы линейной связи между несколькими переменными, имеющими интервальный или порядковый тип шкалы. Необходимо отметить, что дихотомические переменные также могут принимать участие в корреляционном анализе. С точки зрения SPSS они рассматриваются как порядковые переменные.

Существует два основных типа коэффициентов корреляции, рассчитываемых в зависимости от вида шкалы переменных, участвующих в анализе.

1. Для переменных с интервальной шкалой применяется коэффициент корреляции Пирсона. Он позволяет охарактеризовать линейную связь между двумя переменными по указанным параметрам [2]: наличию (есть/нет), направлению (убывает/возрастает) и силе (очень слабая/слабая/умеренная/сильная).

2. Если хотя бы одна из пары исследуемых переменных имеет порядковую или дихотомическую шкалу, используются ранговые коэффициенты корреляции Спирмана или Кендала. Чаще всего эти коэффициенты применяются в маркетинговых исследованиях в тех случаях, когда необходимо установить степень соответствия двух ранжированных списков. Например, если имеются схемы выбора какого-либо продукта различными целевыми группами респондентов (в виде ранжированных по важности параметров) и необходимо установить, насколько точно они соответствуют друг другу (или различаются).

Последний шаг рассматриваемого алгоритма — линейный регрессионный анализ и статистическое прогнозирование Линейная регрессия является наиболее часто используемым видом регрессионного анализа. Известны три основные задачи, решаемые в маркетинговых исследованиях при помощи линейного регрессионного анализа.

1. Определение того, какие частные параметры товара (в данном случае угля) оказывают влияние на общее впечатление потребителей от данного продукта. Установление направления и силы данного влияния. Расчет, каким будет значение результирующего параметра при тех или иных значениях частных параметров.

2. Выявление того, какие частные характеристики товара влияют на

общее впечатление потребителей от данного продукта (построение схемы выбора продукта потребителями). Установление соотношения между различными частными параметрами по силе и направлению влияния на общее впечатление.

3. Графическое прогнозирование поведения одной переменной в зависимости от изменения другой (используется только для двух переменных). Как правило, целью проведения регрессионного анализа в данном случае является не столько расчет уравнения, сколько построение тренда (то есть аппроксимирующей кривой, графически показывающей зависимость между переменными). По полученному уравнению можно предсказать, каким будет значение одной переменной при изменении (увеличении или уменьшении) другой

Выводы. Таким образом, рассмотрены основные этапы анализа данных характеризующих работу угольного предприятия с использованием для этих целей программного комплекса SPSS и практически доказана большая эффективность таких расчетов по сравнению с результатами, полученными при применении для этих же целей Microsoft Excel.

Библиографический список

1 **Бююль Ахим, Цёфель Петер** SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: Пер. с нем. / Ахим Бююль, Петер Цёфель — СПб. : ООО «ДиаСофтЮП», 2005- 608 с

2 **Овсянников В.П. Штенге А.А.** Анализ данных, характеризующих работу угольного предприятия, методом главных компонент.

Дрейпер Н., Смит Г. Прикладной регрессионный анализ, (в 2-х т.). — М.: Финансы и статистика, 198