

СОВРЕМЕННЫЕ ПРОГРАММЫ ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА СОЦИАЛЬНО-ЭКОНОМИЧЕСКОЙ ИНФОРМАЦИИ

Чегерова Татьяна Ивановна, Чегеров Вадим Геннадьевич

Учреждение образования «Могилевский государственный университет
имени А. А. Кулешова» (Могилев, Республика Беларусь)

В статье проводится обзор возможностей и сравнительный анализ наиболее популярных статистических пакетов.

Принятие обоснованных управленческих решений должно опираться на достоверные результаты статистического анализа эмпирических данных. Для этого в настоящее время существует множество различных статистических пакетов, самые мощные из которых: R (S-PLUS, RStudio), Matlab, SAS, SPSS, Stata, Statistica и другие. Каждый из этих пакетов имеет большую функциональность, достаточное количество инструментов для реализации этой функциональности. Программы постоянно совершенствуются и выбор наиболее подходящей для конкретных задач зависит от множества факторов, в том числе необходимость специальных знаний в области программирования и статистики, дружественность интерфейса, представление результатов анализа данных [1]. Не последнее место в этом перечне занимает стоимость приобретения лицензионной версии. По функциональности статистические пакеты можно классифицировать как:

универсальные (SPSS, STATA, Statistica, EVies, Gretl, R)
профессиональные (SAS, BMDP)
специализированные (BioStat, MedCalc, DATASCOPE).

Некоторые несложные виды статистического анализа можно выполнить используя надстройки к Excel. MS Excel не является статистическим пакетом, но он входит в MS Office, включает много статистических функций и дает возможность подключить встроенный пакет Анализа данных. Этих возможностей бывает достаточно на начальных стадиях исследования. Для небольших исследований, когда не требуется проводить кластеризации данных, а лишь необходимо установить некоторые зависимости, дать статистическое описание исследуемым переменным, данный пакет будет экономически выгодным. Кроме того, в последние годы стали появляться надстройки к Excel, обладающие расширенными возможностями статистической обработки данных, такие как Attestat, Real Statistics. Недостатком является то, что данные в Excel не представляют собой «аккуратные данные», что совершенно необходимо для статистического анализа. Практически всегда эти данные требуют большой предварительной работы для приведения их к «аккуратному» виду, когда в каждой колонке хранится один признак данных в одинаковом формате, а каждая строка является полной характеристикой исследуемого объекта по совокупности его признаков. Кроме того, имеются существенные ограничения по объему данных.

Среди практиков очень популярен SPSS, так как не требует специального обучения, имеет удобный пользовательский интерфейс, нажатие нескольких кнопок позволяет выполнить сложные статистические расчеты. Стандартный статистический анализ реализуется полностью с удобным выводом и неплохой визуализацией. Если же задачи исследования лежат в области моделирования процессов, распознавания образов или обработки сигналов, то такие возможности у SPSS весьма ограничены.

Очень большие возможности для аналитика дают Stata и SAS. Процедуры Stata можно вызывать, нажимая кнопки в меню или запуская простые сценарии. В части меню Stata напоминает SPSS. Отдельные сильные стороны Stata по сравнению со всеми другими пакетами: обработка данных опроса (стратифицированные выборки, кластеризация), надежные оценки и тесты, методы продольных данных, многомерные временные ряды. И SAS, и Stata являются языками программирования, поэтому они позволяют строить аналитику на основе стандартных про-

цедур. SAS имеет большие преимущества по сравнению с другими пакетами: большие наборы данных, скорость, красивая графика, гибкость форматирования вывода, процедуры временных рядов, процессы подсчета. Наряду с этими преимуществами есть и недостатки, в частности то, что аналитика сопровождается массивным выводом информации, занимающим иногда десятки страниц. Это доставляет определенные неудобства исследователю при необходимости получения оперативного результата. Все упомянутые выше пакеты являются платными.

Если необходимо расширить рамки стандартного анализа, то реализация задачи будет успешнее в R, так как по сути своей он является консольным языком программирования с относительно простым синтаксисом [2]. Главными идеологами и авторами R являются авторы самого известного в мире статистического пакета SPSS Росс Айхэка и Роберт Джентлемен. Безусловным преимуществом R является то, что он бесплатный. Кроме того в анализе данных очень важную роль играет визуализация данных и в этом плане R предоставляет большие возможности, поскольку графических библиотек существует уже более тысячи. В целом же число постоянно поддерживаемых библиотек составляло на начало 2022 года более 16000 (на GitHub и Cran- крупнейших платформах для разработки программного обеспечения в мире). Это значит, что при появлении новой идеи по обработке некоего класса данных исследователь сам реализует ее в виде библиотеки R. Главное преимущество – открытый код. Если для целей исследования не подходит работа авторской библиотеки, имеется возможность переписать ее под себя. Лучшие библиотеки, пользующиеся наибольшей популярностью, пишутся на C++, или потом переписываются на C++ для быстроты работы.

Возможности R как для анализа данных, так и для графической интерпретации результатов анализа ограничены только знаниями исследователя и желанием изучать новые библиотеки и инструменты анализа. Сообщество ученых – практиков, использующих R, не имеет никаких национальных или языковых ограничений. Если не можете сами найти решение – есть возможность задать вопрос на форуме, отвечают всегда, обычно несколько человек, вместе можно обсудить варианты ответов и выбрать лучшее решение. При таком большом количестве библиотек, которые постоянно пополняются новыми, невозможно знать обо всех, но всегда найдется специалист, имеющий необходимый опыт, обладающий интересными знаниями и предложивший наиболее простое и красивое решение.

Список источников

1. Цыпин, А. П. Статистические пакеты программ в социально-экономических исследованиях / А. П. Цыпин, А. С. Сорокин // Азимут научных исследований: экономика и управление. – 2016. – Т. 5. – № 4 (17). – С. 379–384.
2. R: A Language and Environment for Statistical Computing. Reference Index / The R Development Core Team. Version 4.1.2. – 2021. – 3795 р.