Ивин Вячеслав Вадимович
канд. экон. наук, доцент
Школа экономики и менеджмента
ФГАОУ ВО «Дальневосточный федеральный университет»
г. Владивосток, Приморский край
доцент
ФГАОУ ВО «Владивостокский государственный
университет экономики и сервиса»
г. Владивосток, Приморский край

DOI 10.21661/r-473123

ПРИМЕНЕНИЕ ЯЗЫКА R И СРЕДЫ RSTUDIO ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

Аннотация*: в данной статье даётся краткая характеристика высокоуровневого объектно-ориентированного языка программирования и среды для статистических вычислений и визуализации исходных и расчётных данных R; обосновывается его эффективность применения в учебном процессе при подготовке бакалавров и научном исследовании для магистрантов и аспирантов не только как инструмента обработки данных, но и как среды для обучения программированию и анализу данных.*

Ключевые слова*: анализ данных, язык R, среда RStudio, обучение программированию, высшее образование.*

Статистический анализ данных является неотъемлемой частью любого научного исследования или бизнес-проекта. Качественная обработка данных повышает шансы опубликовать статью в высокорейтинговом журнале и вывести исследование на более высокий, например, международный, уровень. Существует достаточно широкий спектр компьютерных программ, способных обеспечить эффективный качественный и количественный анализ данных, однако большинство из них проприетарные, лицензия на которые стоит больших денег (от нескольких сотен до нескольких тысяч долларов США и выше). Однако,

существуют и альтернативные программы обработки данных, например, язык *R* и ряд интегрированных сред разработки (*IDE*), поддерживающих этот язык, за которые не надо платить, а их надёжность и популярность конкурируют с лучшими коммерческими статистическими пакетами [8].

На данный момент насчитываются десятки качественных пакетов обработки статистических данных, среди которых явными лидерами считаются *SPSS*, *SAS* и *MatLab*. Однако, 2010 году *R* вошёл в список победителей конкурса британского журнала *Infoworld* в номинации на лучшее открытое программное обеспечение для разработки приложений [1], а в 2013 году, несмотря на высокую конкуренцию, *R* стал самым используемым программным продуктом для статистического анализа в научных публикациях [5]. Кроме того, в последнее десятилетие *R* становится всё более востребованным и в бизнес-секторе: такие компании-гиганты, как *Google*, *Facebook*, *Ford*, *New York Times* и др. активно используют его для сбора, анализа и визуализации данных [4; 6], а в корпорации *Boeing* язык *R* является основным инструментом анализа и обработки данных [2].

Значительная часть европейских и американских университетов в последние годы активно переходит к использованию *R* в учебной и научно-исследовательской деятельности вместо дорогостоящих коммерческих разработок [9, с. 5].

*Я*зык *R* – это мощный высокоуровневый объектно-ориентированный язык программирования и среда для статистических вычислений и визуализации исходных и расчётных данных, который позволяет решить множество задач в области обработки данных; это бесплатная программа с открытым кодом (*GNU GPL*), предназначенная для работы под управлением наиболее часто используемых операционных систем (*Microsoft Windows*, *Mac OS*, *Linux* и *Unix*) и поддерживающая тысячи специализированных модулей и утилит. Одной из важнейших особенностей языка *R* является эффективная реализация векторных операций, позволяющая использовать весьма компактную запись при обработке данных большого объёма.

Всё это делает *R* высоко эффективным средством для извлечения полезной информации из «гор сырых данных», в т.ч. и из больших данных (*Big Data*), и,

соответственно, удобным и эффективным инструментом для обучения технологии анализа, обработки и визуализации данных.

Таким образом, в настоящее время язык *R* является одним из ведущих статистических инструментов в мире. Он активно применяется в различных сферах деятельности, например, в генетике, молекулярной биологии и биоинформатике, науках об окружающей среде (экология, метеорология и др.), экономических и сельскохозяйственных дисциплинах. Также *R* всё больше используется в обработке медицинских данных, вытесняя с рынка такие коммерческие пакеты, как *SAS* и *SPSS* [3].

Для удобства работы пользователя с *R* разработан ряд графических интерфейсов, в том числе *RStudio*, *Rgui*, *JGR*, *RKWard*, *SciViews-R*, *Statistical Lab*, *R Commander*, *Rattle* и др.

Кроме того, в ряде текстовых и кодовых редакторов предусмотрены специальные режимы для работы с *R*, в частности в *ConTEXT*, *Emacs* (*Emacs Speaks Statistics*), *jEdit*, *Kate*, *Syn*, *TextMate*, *Tinn-R*, *Vim*, *Bluefish*, *WinEdt* (с пакетом *RWinEdt*), *Gedit* (с пакетом *rgedit*/*gedit-r-plugin*). Для среды разработки *Eclipse* существует специализированный *R*-плагин; доступ к функциям и среде выполнения *R* возможен из *Python* с использованием пакета *RP*y; работать с *R* можно из эконометрического пакета *Gretl*.

Существует несколько программ-оболочек для удобства работы с *R*, внешний вид и функциональность которых могут сильно отличаться. Наиболее популярными вариантами таких программ-оболочек являются *Rgui*, *RStudio* и *R*, запущенный в терминале *Linux* / *UNIX* в виде командной строки.

В качестве основных достоинств среды *R* можно отметить:

– бесплатность и кроссплатформенность;

– богатый арсенал используемых статистических методов и инструментов;

– качественная векторная графика;

– более 12 000 проверенных пакетов;

– гибкость в использовании:

– позволяет создавать / редактировать скрипты и пакеты;

– взаимодействует с другими языками программирования, такими как *C*/*C++*, *Java* и *Python*;

– может работать с форматами данных для *SAS*, *SPSS* и *STATA*;

– импортирует данные в формат *TeX* (*LaTeX*);

– активное сообщество пользователей и разработчиков;

– регулярные обновления, хорошая документация и техническая поддержка.

В качестве недостатков следует отметить:

– небольшой объём информации на русском языке (хотя за последние годы появилось несколько обучающих курсов и интересных книг);

– относительная сложность в использовании для пользователя, незнакомого или малознакомого с программированием. Частично это можно сгладить, работая в *GUI Rcmdr*, но для нестандартных решений всё же необходимо использовать консоль и командную строку.

Освоение современных методов анализа данных является необходимым условием для достижения любой амбициозной цели будь то в науке, бизнесе или образовании, поэтому применение языка программирования и среды *R* – очень важно для профессиональной подготовки будущих бакалавров, специалистов и магистров, а также для послевузовского образования в различных сферах деятельности, связанных с анализом и обработкой данных.

На наш взгляд интегрированная среда разработки *IDE RStudio является наиболее комфортной и дружественной для учащихся*, *т.к.* имеет более удобный (по сравнению с консолью и другими интегрированными средами разработки для языка *R*) интерфейс, цветовую «подсветку» и автоматическое завершение кода, а также ряд других функций, максимально упрощающих освоение и, соответственно, работу с *R*, что и обусловило внедрение *R* в учебный процесс для направлений подготовки «Экономика», «Менеджмент», «Торговое дело» и др., основная образовательная программа которых предусматривает изучение современных средств и методов проведения статистического анализа данных.

По мимо освоения *R* в рамках учебных дисциплин «Инструментальные средства анализа и обработки данных», «Введение в обработку больших данных», «Анализ больших данных» и др. студенты знакомится с технологиями подготовки, хранения и систематизации данных; учатся использовать эконометрические и математические методы и модели для обработки и анализа больших данных; приобретают практические навыки статистической обработки данных и работы с графикой с помощью языка *R* (с применением *RStudio* и др. *IDE*), входящего в рейтинг самых востребованных и перспективных языков программирования, а также приобретают навыки использования результатов анализа (больших) данных при принятии решений в различных предметных областях.

IDE RStudio представляет собой бесплатную интегрированную среду разработки для языка *R*. Благодаря ряду своих особенностей этот активно развивающийся программный продукт делает работу с *R* удобной для учебного процесса с подготовленными слушателями. Среда разработки позволяет использовать не только стандартные процедуры, описывающие совокупность основных подходов, инструментов и методов статистической обработки данных, но и процедуры и функции из дополнительных пакетов библиотеки (*CRAN Packages* [7]), в т. ч. предназначенные для обработки структурированных и неструктурированных данных больших объёмов и значительного многообразия для получения воспринимаемых человеком результатов. На момент написания статьи международным сообществом исследователей было размещено в библиотеке более 12 000 дополнительных пакетов на языке *R*, охватывающих всевозможные разделы эконометрического анализа, прогнозирования временных рядов, многомерной статистики, причём каждый из пакетов сопровождается документацией по установленному стандарту.

Консоль *RStudio* (*Console*) предоставляет целый ряд опций, делающих работу с *R* простой и продуктивной (*рисунок 1*). Освоение этих опций, наряду с возможностями, доступными в панелях *Source* (Редактор кода) и *History* (История), может оправдать затраченное на обучение время.

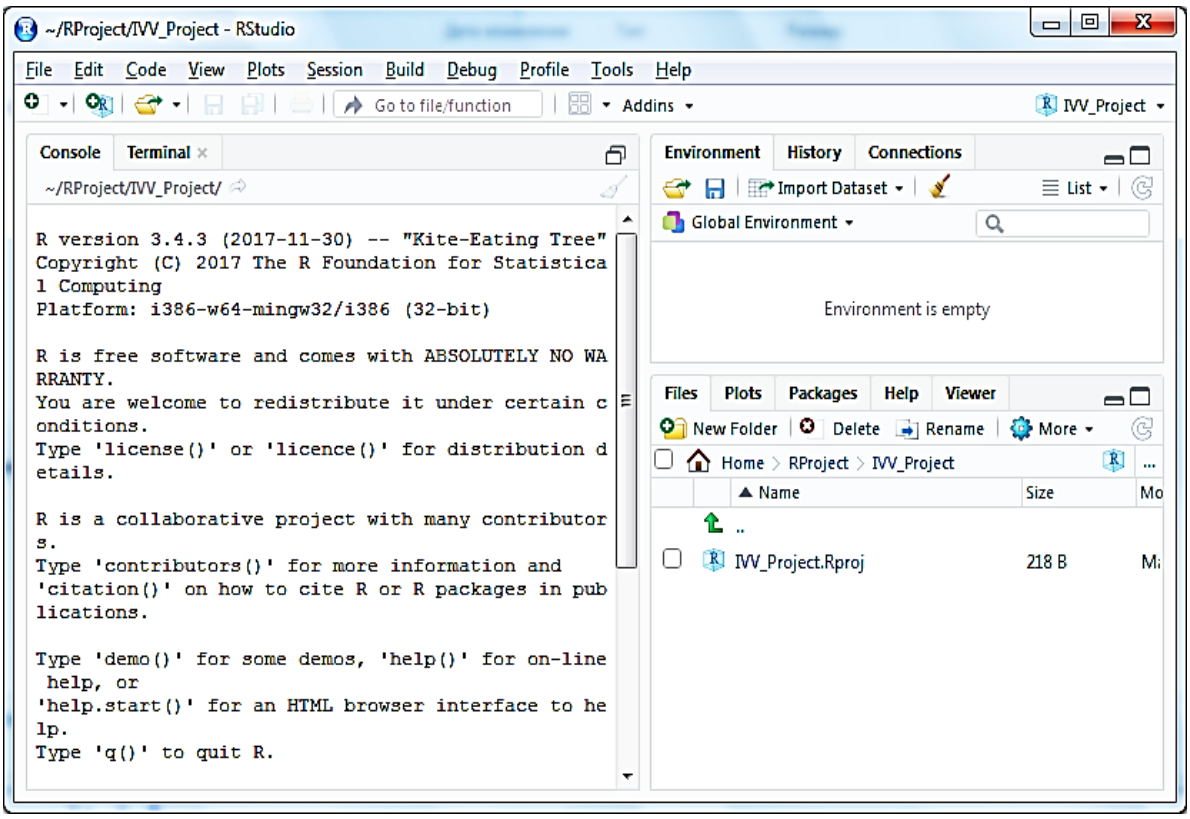

Рис. 1. Интерфейс *IDE RStudio*

Редактор кода *RStudio* включает ряд опций для продуктивной работы, в частности «подсветку» кода и другие специализированные опции по работе с кодом следующих типов файлов: *R*-скрипты, документы *Sweave*, документы *TeX*, автоматическое завершение кода, одновременное редактирование нескольких файлов, поиск и замену определённых частей кода. Кроме того, в *RStudio* имеются гибкие возможности по выполнению кода непосредственно из окна редактора. Для многих учащихся это является предпочтительным способом работы с *R*.

RStudio поддерживает выполнение кода непосредственно из окна Редактора – выполняемые команды посылаются в Консоль, где появляется также результат их выполнения (рисунок 2). *RStudio* включает ряд опций, обеспечивающих быструю навигацию по *R*-коду. Во время работы *RStudio* создаёт базу данных всех команд, которые пользователь вводит в Консоль. Имеется возможность просмотра этой базы данных при помощи панели *History* (История).

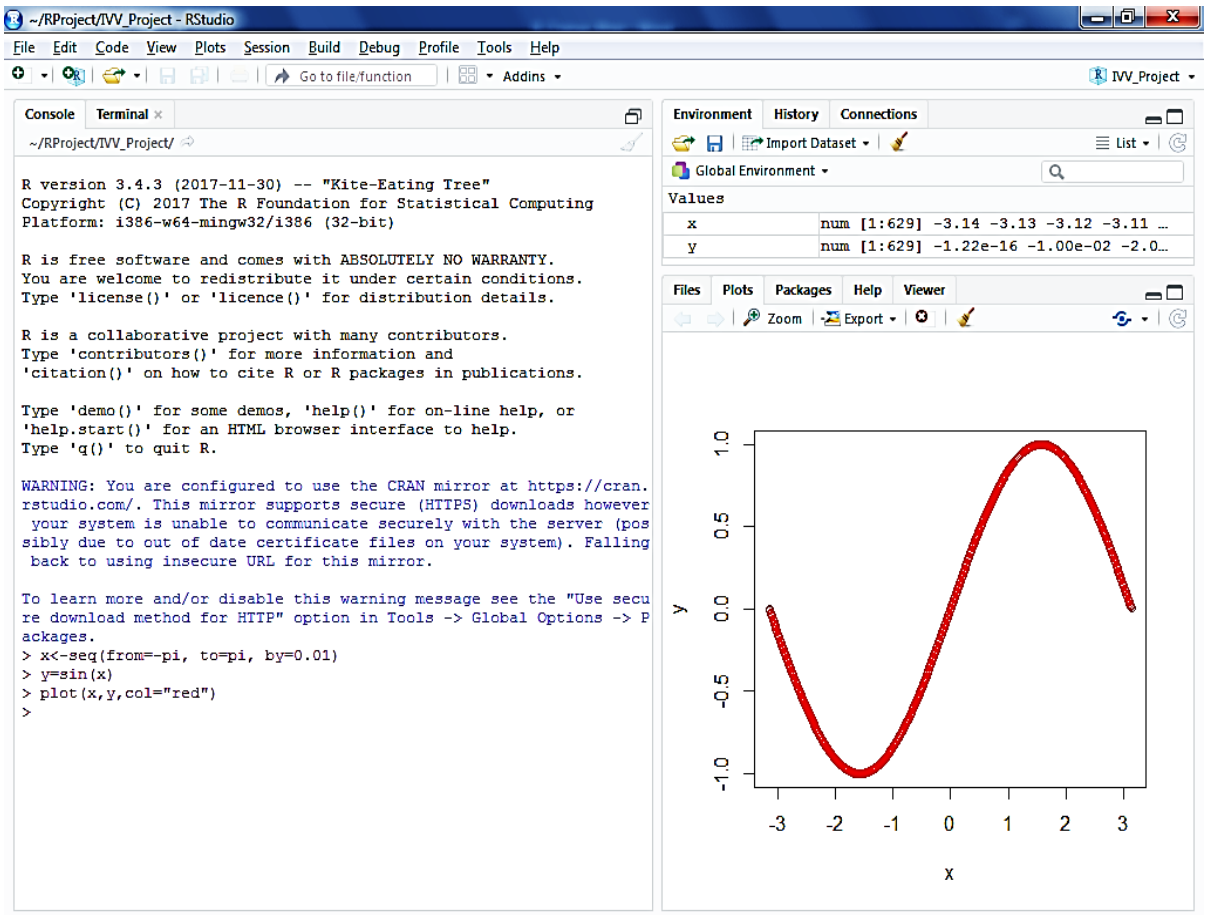

Рис. 2. Выполнение кода непосредственно из окна Редактора *IDE RStudio*

Если все файлы, имеющие отношение к определённому проекту, хранятся в одной папке, имеет смысл сделать её исходной для работы. *RStudio* автоматически будет делать рабочей папкой ту, в которой хранится открываемый файл. *RStudio* позволяет организовать работу в соответствующие контексту проекты так, что каждый проект будет иметь свою собственную рабочую директорию, рабочее пространство, историю и скрипты. Проекты *RStudio* ассоциированы с рабочими директориями *R*, следовательно, проект можно создать: в новой директории; в существующей директории, где уже хранятся скрипты с *R*-кодом и данные; путём копирования файлов, хранящихся в одной из онлайн-систем контроля версий. Для создания нового проекта служит команда *New Project* (Новый Проект), доступная из закладки *Projects* главного меню и из панели инструментов (в дальнем правом углу рабочего окна программы). Имеется несколько опций для настройки поведения каждого конкретного проекта в *RStudio*. Эти опции доступны по команде *Project Options* из раздела *Project* главного меню программы.

Таким образом, язык *R, IDE RStudio* и библиотека *CRANE* по научному уровню и диапазону возможностей составляют конкуренцию коммерческим продуктам и могут быть использованы для укрепления междисциплинарных связей, в написании выпускной квалификационной работы и научно-исследовательской работе студентов всех уровней подготовки.

В то же время выявлено требование по отличному уровню освоения базового курса статистики на этапе бакалавриата и по наличию навыков программирования, включая применение объектно-ориентированной технологии в написании исходных кодов. Ограничения приводят к выводу об индивидуальном подходе к применению этого программного обеспечения, с учётом возможностей и потребностей учащегося магистратуры. Тем не менее, в случае оправданного использования, появляется возможность использовать потенциал научного роста обучаемого, введения готовых или самостоятельно написанных исходных кодов в выпускную квалификационную работу, ознакомления магистранта с мировым уровнем достижений в области интеллектуального анализа данных.

Краткое перечисление этих возможностей позволяет сделать вывод о доступности применения среды *RStudio* в учебном процессе и научном исследовании для магистрантов, имеющих базовую подготовку в области программирования и общей теории статистики.

Список литературы

1. R-анонс – важные объявления проекта R [Электронный ресурс]. – Режим доступа: https://stat.ethz.ch/pipermail/r-announce/2018/000626.html

2. Bhalla D. Companies using R [Электронный ресурс]. – Режим доступа: http://www.listendata.com/2016/12/companies-using-r.html

3. Galili T. Tutorials for learning R [Электронный ресурс]. – Режим доступа: https://www.r-bloggers.com/how-to-learn-r-2/

4. List of Big Companies using R [Электронный ресурс]. – Режим доступа: http://www.makemeanalyst.com/companies-using-r/

5. Muenchen R.A. The Popularity of Data Analysis Software [Электронный ресурс]. – Режим доступа: http://www.r4stats.com/articles/popularity/

6. Statistical & Financial Consulting by Stanford PhD [Электронный ресурс]. – Режим доступа: http://www.stanfordphd.com/Statistical_Software.html

7. Доступ к сетевому архиву «The Comprehensive R Archive Network» [Электронный ресурс]. – Режим доступа: http://cran.r-project.org

8. Сёмочкина И.Ю. Применение языка R и среды RStudio для математической обработки данных / И.Ю. Сёмочкина, О.В. Прокофьев [Электронный ресурс]. – Режим доступа: http://www.penzgtu.ru/fileadmin/filemounts/confcit/articles/spring-25_2017/semochkina.pdf

9. Статистический анализ данных в системе R: Учебное пособие / А.Г. Буховец, П.В. Москалев, В.П. Богатова, Т.Я. Бирючинская; под ред. проф. А.Г. Буховца. – Воронеж: ВГАУ, 2010. – 124 с.