Программа учебной дисциплины «Анализ и визуализация данных в R»

Утверждена

Академическим советом ООП

Протокол № 38 от 21 июня 2018 г.

Автор	Руднев М.Г.
Число кредитов	4
Контактная работа (час.)	40
Самостоятельная работа (час.)	112
Курс	4
Формат изучения дисциплины	Без использования онлайн-курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целью освоения курса "Анализ и визуализация данных в R" является формирование компетенций, связанных с решением задач по сбору, анализу и визуализации количественных данных в социологии. Курс направлен на освоение новых технологий при использовании известных методов анализа данных, а также метода моделирования структурными уравнениями, его главная задача – в ознакомлении студентов с программной средой R. Специальное внимание уделяется принципам и технике визуализации количественных данных. Ключевая ценность, реализуемая в этом курсе – прозрачность и воспроизводимость количественных исследований.

Курс является введением в основные понятия и команды языка R, представляет реализацию как знакомых студентам методов анализа данных в этой среде (описательные статистики, анализ главных компонент, кластерный и регрессионный анализы), так и новые методы: путевой анализ и конфирматорный факторный анализ. В курсе представлены основные принципы работы в R, операции по сбору, анализу и презентации количественной информации. В результате освоения этого курса студенты будут способны реализовать и документировать процесс исследования от сбора данных до (автоматизированной) публикации отчетов.

Большое внимание уделяется изучению примеров решения конкретных задач по материалам исследовательских проектов. Программа предусматривает проведение семинарских занятий и лабораторных работ, подготовка к которым осуществляется студентами самостоятельно по рекомендованной литературе. Помимо этого, предусматривается выполнение и последующая проверка обязательных домашних работ (решение задач).

В результате освоения дисциплины студент должен:

знать:

- основные принципы работы языка R;

- синтаксис и базовые функции R;

- функционал пакетов ggplot2;

- принципы построения, оценки качества, сравнения и модификации структурно-ковариационных моделей;

уметь:

- обрабатывать и анализировать данные для подготовки аналитических решений, экспертных заключений и рекомендаций;
- использовать социологические методы исследования для изучения актуальных социальных проблем, для идентификации потребностей и интересов социальных групп;
- отличать эффективную визуализацию информации от проблематичной;
- строить модель конфирматорного факторного анализа;

владеть:

- способностью использовать основные законы естественнонаучных дисциплин в профессиональной деятельности, применять методы математического анализа и моделирования, теоретического и экспериментального исследования;
- способностью и готовностью использовать знание методов и теорий социальных и гуманитарных наук при осуществлении экспертной, консалтинговой и аналитической деятельности.

Изучение дисциплины "Анализ и визуализация данных в R" базируется на следующих дисциплинах:

- Теория вероятностей и математическая статистика
- Анализ социологических данных

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- знать основные законы классической и современной физики;
- знать простейшие методы решения физических задач;
- обладать навыками работы с измерительными приборами.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в R.

Общая характеристика языка R. Базовые команды, пакеты в R. Знакомство с RStudio и R commander. Объекты и функции. Виды объектов. Понятие класса объекта. Типы хранения данных: векторы, двухмерные таблицы, матрицы, массивы, списки. Типы переменных: числовые, строчные, факторы. Виды пропущенных данных: NA, NaN. Способы работы с пропущенными данными. Понятие среды, ссылки на функции из разных пакетов, создание собственной среды. Циклы for, while, repeat. Создание собственной функции.

Тема 2. Чтение, преобразование и экспорт данных в R.

Основные функции из пакетов foreign, haven, car и dplyr. Основные идеи html, markdown и LaTEX. Пакет stargazer. Имитация данных в R.

Тема 3. Анализ данных в R.

Линейные и логистические бинарные регрессии в lm и glm. Анализ главных компонент в prcomp и princom. Кластерный анализ в kmeans и hclust. Многомерное шкалирование в mds.

Команда sapply и mapply. Дебаггинг. Оптимизация кода. Создание автоматических отчетов, знакомство с rmarkdown.

Тема 4. Визуализация данных в R base и ggplot2.

Мотивы визуализации. Виды графиков. Связь между моделью анализа и графиками. Синтаксис ggplot2: qplot, geom, aes. Использование пространства координат: одно-, двух-, трех- мерные, сферические, географические системы координат. Использование символов и цветов. Создание анимированных графиков в пакете animation.

Тема 5. Основы структурно-ковариационного моделирования. Путевой анализ.

Путевой анализ. Путевые диаграммы. Путевой коэффициент. Опосредованное воздействие (медиация) и взаимодействие (модерация). Рекурсия. Ограничения параметров.

Идентификация структурной модели. Правила нахождения возможности идентифицировать произвольную структурную модель. Переопределение модели. Ошибки структурной модели. Зависимость ошибок. Модификационные индексы.

Оценки согласия модели. Оценки, основанные на Хи-квадрат, информационные критерии. Возможности по отвержению и выбору модели.

Тема 6. Конфирматорный факторный анализ.

Модели измерения латентных переменных: разведывательный и подтверждающий факторный анализ. Анализ главных компонент. Частные наименьшие квадраты. Формативные и рефлективные измерительные инструменты.

Этапы построения и модификации измерительной модели.

MTMM модели, модели со структурой средних, факторы высшего порядка

Построение моделей эксплораторного и конфирматорного факторного анализа в пакетах factanal и lavaan. Сравнение моделей в lavaan.

III. ОЦЕНИВАНИЕ

Студентам предлагается 3 письменных домашних задания, включающих сбор данных, построение заданного типа модели, и создание автоматизированного отчета, предполагающее использование программы R. Задачи включают в себя самостоятельное построение моделей на индивидуальных для каждого студента данных.

1. Загрузка и подготовка данных в R. Построение модели регрессионного анализа.
2. Визуализация описательных и аналитических статистик. Выгрузка автоматизированного отчета.
3. Построение модели конфирматорного факторного анализа.

Порядок формирования оценок по дисциплине

Преподаватель оценивает самостоятельную работу студентов: правильность и своевременность выполнения домашних работ, задания для которых выдаются на семинарских занятиях. За нарушение срока сдачи работы на 1 неделю оценка за нее снижается на 50%, на 2 недели – на 100%.

Итоговая оценка по дисциплине Оценка складывается из двух компонентов:

- Три домашних задания по разделам «Анализ данных», «Визуализация данных» и «Струтурные уравнения» – 80%.
- Письменный тест на знание теоретических понятий – 20%

Итоговая оценка за курс округляется математически (например, 5,5 округляется до 6, а 5,4-до 5). Исключение составляют итоговые оценки по дисциплине: оценки менее 4-х баллов. В этом случае даже 3,9 округляется до 3.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

1. Напишите значение (значения), которое возвращает следующая строка кода R: sapply(1:5, function(x) x + 10)
2. Исправьте ошибку, из-за которой может не работать следующая строка кода R? Names(data) = c('var1', 'var2', 'var3')
3. Напишите по памяти код точечного графика с разноцветными точками и подписями с использованием функций из пакета ggplot2.
4. Каким символом в rmarkdown обозначаются заголовки?
5. В исходную структурную модель вы решили добавить переменную "возраст". С помощью какой статистики (статистик) согласия можно сравнивать эту и исходную модель?
6. Чем медиация отличается от модерации? Приведите пример. Как моделируются модерация и медиация переменных?
7. Что такое параметры модели? Чем они отличаются от переменных? Что такое ограничение и фиксация параметров? Для чего они могут применяться?
8. Нарисуйте путевую диаграмму структурной модели с наблюдаемыми переменными. Поясните смысл каждого элемента.
9. Зачем необходима оценка согласия модели? Опишите принцип традиционных оценок согласия модели.

V. РЕСУРСЫ

5.1 Основная литература

Язык и среда программирования R: Учебное пособие / Золотарюк А.В. - М.:НИЦ ИНФРА-М, 2018. - 183 с. Режим доступа: http://znanium.com/catalog/product/997099

Brown, T. Confirmatory Factor Analysis for Applied Research. Second edition. Guilford Press. New York, London. 2014. Доступна в подписке, URL: https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1768752

Дополнительная литература

Saquib, Nazmus. Mathematica Data Visualization, Packt Publishing Ltd, 2014. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1800642.

Руднев М. Г. Инвариантность измерения базовых ценностей по методике Шварца среди русскоязычного населения четырех стран // Социология: 4М. 2013. № 37. С. 7-38. Электронный ресурс, URL: https://maksimrudnev.com/wp-content/uploads/2017/07/2013-d180d183d0b4d0bdd0b5d0b2-d0b8d0bdd0b2d0b0d180d0b0d0bdd182d0bdd0bed181d182d18c-d181d180d0b5d0b4d0b8-d180d183d181d181d0bad0bed18f.pdf

5.2 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	R + Rstudio	*Бесплатная программа*
2.	Open Office / LibreOffice	*Бесплатная программа*

5.3 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	*1.* Lavaan Online Tutorial. Chapters “A CFA example”, “Mean Structures”, “Modification indices”, “Extracting information”.	URL: http://lavaan.ugent.be/tutorial/index.html
2.	• База данных Европейского социального исследования ESS.	URL: http://EuropeanSocialSurvey.org
	• Chang, W. (2012). R Graphics Cookbook: Practical Recipes for Visualizing Data. O'Reilly Media, Inc.	URL: http://r-cookbook.com

	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Открытое образование	URL: https://openedu.ru/

5.5 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для самостоятельных занятий по дисциплине оснащены ПЭВМ с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ