

Анализ данных для социальных исследователей

ДМИТРИЙ АЛЕКСЕЕВИЧ КОЗЬМИН

МАГИСТР ПОЛИТИЧЕСКИХ НАУК

31 МАРТА 2025 Г.

Для тех, кто хочет освоить применение методов количественного анализа «с нуля». В процессе курса с помощью языка программирования R будут изучены такие методы, как Т-тест, Анова, Хи-квадрат, корреляционный анализ, регрессионный анализ (МНК и логит/пробит регрессионные модели), чтобы в конце курса написать собственное исследование с применением этих методов.

Курс

Курс предполагает как лекционные, так и семинарские занятия. На лекционных занятиях лектором будет предоставлена информация посредством оральной коммуникации с использованием технологии презентации PowerPoint. На семинарских занятиях предполагается совместная работа со студент:ками в программе RStudio для непосредственного применения методов количественного анализа на готовых скриптах. В качестве домашних работ скрипты будет необходимо писать самостоятельно. Лекция и семинар в сумме будут составлять два академических часа, то есть 90 минут. Семинарская часть наступает непосредственно сразу после лекционной части.

Лекция 1. Введение в содержание курса: количественные методы в социальных науках, обсуждение основных инструментов для анализа данных.

Семинар 1. Установка R и RStudio

Лекция 2. Методы описательной статистики: типы данных, импорт данных разных разрешений с помощью пакета "readr", извлечение отдельных данных из массива).

Семинар 2. Работа с базой данных ОВД-Инфо, рассмотрение количества задержанных в зависимости от года и региона. Влияние других факторов на количество задержанных.

Лекция 3. Хи-квадрат: альтернативная и нулевая статистические гипотезы, расчёт

хи-квадрат в R, Сравнение выборок: статистические тесты: параметрические – непараметрические; двусторонние – левосторонние – правосторонние

Семинар 3. Работа с базой данных политических предпочтений избирателей РФ.

Тестирование гипотез, предложенных присутствующими на семинаре с помощью метода анализа данных "Хи-квадрат".

Лекция 4. Т-тест: Параметрические тесты: t-тест для независимых и парных выборок.

Непараметрические тесты: тест Вилкоксона (Манна-Уитни) для независимых и парных выборок. Тест Шапиро-Уилкса для проверки нормальности распределения. Иллюстрация работы статистических тестов.

Семинар 4. Работа с базой данных документаций о выданных грантах, собранной медиа "Говорит НеМосква", проверка гипотезы о связи между близостью к государственным организациям и общей суммой полученных средств, где единицей наблюдения будет являться организация, которой выделяют грант.

Лекция 5. Корреляционный анализ: корреляция и ковариация, коэффициент корреляции Пирсона. Интерпретация значений коэффициента корреляции. Значимость коэффициента корреляции. Коэффициент корреляции Спирмена. Построение корреляционных матриц в R.

Семинар 5. Работа с базой данных проекта всемирного исследования ценностей WVS, проверка гипотез корреляции между субъективным благополучием и голосованием на выборах. Проверка направления связи и силы связи.

Лекция 6. МНК-регрессия: принцип, оформление, интерпретация: Отличие регрессии от корреляции. Зависимая и независимая переменные. Метод наименьших квадратов (МНК).

Парная линейная регрессия: уравнение. Интерпретация регрессионной выдачи.

Коэффициент детерминации (R^2)

Семинар 6. Работа с базой данных об эффективности университетов, собранной Центром Перспективных Управленческих Решений. Проверка влияния доли сотрудников с кандидатской степенью на показатели эффективности университета. Проверка иных гипотез, предложенных присутствующими на семинаре.

Лекция 7. Биномиальная регрессия: Обобщённые линейные модели. Логистическая регрессия. Уравнение бинарной логистической регрессии. Параметры оценки логистических моделей. Выдача логистической регрессии, её интерпретация.

Предсказанные вероятности и отношения шансов.

Семинар 7. Работа с базой данных опроса 13 волны проекта "Хроники". Поиск факторов, влияющих на позицию россиян по различным чувствительным вопросам, которые изучает этот проект. Проверка иных гипотез, предложенных присутствующими на семинаре.

Лекция 8. Порядковая логистическая регрессия: Предсказанные вероятности и отношения шансов для порядковой логистической регрессии, консультации по итоговым работам

Семинар 8. Работа с базой данных опроса Исследовательского центра Ханны Арендт.

Проверка гипотез оригинального исследования и тестирование новых гипотез при помощи регрессионной выдачи

Лекция 9. Диагностика регрессионных моделей: Проверка модели на мультиколлинеарность, гетероскедастичность, эндогенность, автокорреляцию и нормальное распределение регрессионных остатков. Визуализация диагностических тестов.

Семинар 9. Работа с базами данных, которые будут предложены участниками семинара.

Система оценки: чтобы получить сертификат, необходимо будет выполнить домашние задания. После каждого семинара (кроме первого) студентам нужно будет выполнять небольшие домашние задания в виде скриптов, где будет необходимо самостоятельно применить изученные методы на готовых данных. Чтобы получить рекомендацию, в конце курса будет необходимо предоставить своё собственное исследование на любую интересующую тематику с применением одного или нескольких изученных методов количественного анализа. В этой работе должно содержаться обоснование выбора данных и метода исследования, операционализация переменных, визуализация результатов исследования в виде графика или таблицы. Предполагаемый объём работы: 3-4 тысячи слов. Также будет необходимо предоставить скрипт R и базу данных для воспроизводимости исследования.

Пререквизиты к слушателям

Желание сделать собственное исследование в области социальных наук.

Список литературы

- 1) Кабаков Р. И. R в действии. Анализ и визуализация данных в программе R / Пер. с англ. П. А. Волковой. М.: ДМК Пресс, 2014. 768 с.
- 2) Шипунов А. Б. и др. Наглядная статистика: используем R! М: ДМК Пресс, 2012. 296с.
- 3) Bakija J. 2013. A Non-Technical Introduction to Regression, P. 1-10. URL: web.williams.edu/Economics/wp/Bakija-Non-Technical-Introduction-to-Regression.pdf
- 4) Brambor T, Clark W. R. and Golder M. 2006. Understanding Interaction Models: Improving Empirical Analyses. Political Analysis, Vol. 14, No. 1, 63-82.
- 5) Field A., Miles J. and Field Z. 2012. Discovering Statistics Using R. SAGE Publications 992p.
- 6) Geddes B. 2003. Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics. University of Michigan Press, P. 89-106
- 7) King G. 1986. How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. American Journal of Political Science, Vol. 30, No. 3, 666-687.

Дополнительная информация

Цех социологии и гендерных исследований

Контакты для обратной связи

+393200229492 ; dmicozy@gmail.com