Введение в количественные методы для социальных наук

КОЗЬМИН ДМИТРИЙ АЛЕКСЕЕВИЧ

МАГИСТР ПОЛИТИЧЕСКИХ НАУК, ПОЛИТОЛОГ, АНАЛИТИК ДАТА-ОТДЕЛА ВАЖНЫХ ИСТОРИЙ

26 АВГУСТА 2024 Г.

Цель курса: научить работе с количественными данными с нуля с целью проведения исследований. В процессе курса планируется изучение таких методов анализа данных, как Хи-квадрат, корреляционный анализ, регрессионный анализ (МНК и логит/пробит регрессионные модели), а также их визуализация и описательная статистика. В качестве инструмента анализа курс опирается на язык программирования R. Итоговой работой будет написание собственного исследования с применением изученных методов.

Курс

Структура курса:
Курс предполагает лекции для изучения методов и семинарские занятия для применения методов в программе RStudio на готовых скриптах. (В качестве домашних работ скрипты будет необходимо писать самостоятельно)

- Лекция 1. Содержание курса. Основная и дополнительная литература. Количественные методы в социальных науках. Специализированные статистические программы: виды, примеры, особенности. Установка и запуск статистической среды R и RStudio. Интерфейс R и RStudio. Команды, объекты и функции в R.
- Семинар 1. Запуск RStudio. Основные арифметические действия. Объекты и функции. Оператор присваивания. Загрузка и запуск пакетов. Ввод данных: функция c(). Сохранение скриптов. Открытие сохранённых скриптов. Выход из программы.

- Лекция 2. Типы данных. Базы данных: глобальные и авторские. Генеральная совокупность и выборка. Описательная статистика: меры центральной тенденции и меры разброса. Нормальное распределение и центральная предельная теорема.

- Семинар 2. Импорт данных в R из Excel: функция read.csv(). Кнопка Import Dataset в RStudio. Импорт данных в R из других статистических программ: пакет foreign, функции read.spss(), read.dta(). Работа с загруженными данными: функции attach() и detach(), команда $, извлечение отдельных данных из массива. Таблицы. Расчёт описательной статистики.

- Лекция 3. Роль визуализации данных в научном исследовании. Принципы визуализации данных. Типы диаграмм: диаграмма рассеяния, диаграмма распределения (гистограмма), диаграмма размахов (боксплот), скрипичная диаграмма, столбчатая диаграмма, круговая диаграмма. Примеры диаграмм из политологических исследований.
- Семинар 3. Функции для создания диаграмм: plot(), hist(), boxplot(), vioplot(), barplot(), pie(). Функции для корректировки диаграмм: par(), title(), lines(), legend(). Сохранение диаграмм в разных форматах и с разным разрешением. Пакет ggplot2: краткий обзор.

- Лекция 4. Статистические гипотезы: альтернативная и нулевая. Статистические ошибки: первого и второго родов. Статистическая значимость. Статистика хи-квадрат. Использование статистики хи-квадрат в социальных науках.
- Семинар 4. Расчёт статистики хи-квадрат. Работа с данными.

- Лекция 5. Биномиальный тест. Сравнение выборок: статистические тесты (параметрические – непараметрические; двусторонние – левосторонние – правосторонние). 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 5. Расчет биномиального теста, t-теста, теста Вилкоксона (Манна-Уитни), теста Шипиро-Уилкса. Работа с данными.

- Лекция 6. Корреляция и ковариация. Коэффициент корреляции Пирсона. Интерпретация значений коэффициента корреляции. Значимость коэффициента корреляции. Коэффициент корреляции Спирмена. Построение корреляционных матриц в R.
- Семинар 6. Реализация корреляционного анализа в R. Работа с данными.

- Лекция 7. Отличие регрессии от корреляции. Зависимая и независимая переменные. Метод наименьших квадратов (МНК). Парная линейная регрессия: уравнение. Интерпретация регрессионной выдачи. Коэффициент детерминации (R 2).
- Семинар 7. Реализация парной линейной регрессии в R. Работа с данными.

- Лекция 8. Множественная линейная регрессия: уравнение. Статистика бетакоэффициентов. Стандартизация бета-коэффициентов. F-статистика.
- Семинар 8. Реализация множественной линейной регрессии в R. Работа с данными.

- Лекция 9. Категориальные и порядковые переменные в регрессионном анализе. Однофакторный дисперсионный анализ (one-way ANOVA). Эффекты взаимодействия между переменными. Визуализация эффектов взаимодействия.
- Семинар 9. Расчёт регрессионных моделей с категориальными и порядковыми переменными. Проведение однофакторного дисперсионного анализа. Расчёт и визуализация эффектов взаимодействия. Работа с данными.

- Лекция 10. Оформление результатов регрессионной выдачи: примеры. Пакет stargazer. Композиционное построение количественных исследований.
- Семинар 10. Рассмотрение композиционного построения исследования, использующего в качестве основного метода анализа данных множественную линейную регрессию.

- Лекция 11. Предпосылки МНК-регрессии. Технические проблемы регрессионных моделей: мультиколлинеарность, гетероскедастичность, выбросы, влиятельные наблюдения. Диагностика и способы решения технических проблем регрессионных моделей.
- Семинар 11. Работа с функциями: vif(), ncvTest(), spreadLevelPlot(), qqplot(), outlierTest(), infuencePlot() и др. Работа с данными.

- Лекция 12. Обобщённые линейные модели. Логистическая регрессия. Уравнение бинарной логистической регрессии. Параметры оценки логистических моделей. Выдача логистической регрессии, её интерпретация. Предсказанные вероятности и отношения шансов.
- Семинар 12. Реализация логистической регрессии в R. Работа с данными.

Система оценки:
- 50% Домашние задания: после каждого семинара (кроме первого)
слушателям будет необходимо выполнять небольшие домашние задания в виде скриптов с применением изученных методов на готовых данных.
- 50% Итоговая работа: в конце курса будет необходимо предоставить своё собственное исследование на любую интересующую тематику с применением одного или нескольких изученных методов количественного анализа. Требования к итоговой работе: теоретическая рамка исследования, обоснование выбора данных и метода исследования, операционализация переменных, визуализация результатов исследования в виде графика или таблицы, интерпретация результатов. Предполагаемый объём работы: 3-4 тысячи слов + скрипт R и набор данных для воспроизводимости исследования.

Пререквизиты к слушателям

Для освоения курса слушателям достаточно владеть арифметическими действиями и желание применить изученные методы в собственном исследовании Желательно: понимание основ статистики, уровень английского не ниже B1

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.

Контакты для обратной связи

dmicozy@gmail.com