

Метода анализа неоднородных данных и паттерн-анализ

Организуем рабочее пространство

17 января 2025

В каком направлении мы движемся

- приобрести уверенность в обращении с классической линейной регрессионной моделью
- научиться выявлять неоднородный характер взаимосвязи
- выйти хорошо вооруженными (подходящими методами для работы с гетерогенной выборкой) и практиковать разборчивость (понимать, когда и какой метод использовать)
- изучить методы анализа паттернов

Наш план:

- повторить основы статистики, в частности, модель классической линейной регрессии
- изучить и научиться применять регрессионные модели с переменными взаимодействия
- изучить и научиться применять модели с фиксированными эффектами и модели со случайными эффектами, модели со смешанными эффектами
- изучить и научиться применять методы анализа паттернов

Полезные контакты:

Преподаватели

Сальникова Дарья Вячеславовна (darysalnikova@yandex.ru;
dsalnikova@hse.ru;

web-page: <https://www.hse.ru/staff/salnikova>)

Мячин Алексей Леонидович (amyachin@hse.ru;

web-page: <https://www.hse.ru/org/persons/69178868>)

Наш ассистент

Матвей Бислер (mabisler@edu.hse.ru)

Что Вас ожидает

Формат занятий

Лекции (новый материал + обсуждения) + семинары (решение задач, практика в Python)

Типы заданий

- Семинарская активность
- Домашние задания
- Проверочные работы
- Эссе
- Экзамен

Правила прохождения дистанции (1)

Заветная формула

$$\text{Итог} = 0.1 \times \text{ДЗ} + 0.1 \times \text{Проверочная1} + 0.1 \times \text{Проверочная2} + 0.1 \times \text{Проверочная3} + 0.2 \times \text{ЭССЕ} + 0.1 \times \text{Активность} + 0.3 \times \text{Экзамен}$$
$$\text{Экзамен} = 0.6 \times \text{Часть1} + 0.4 \times \text{Часть2}$$

Домашние задания

Оценка рассчитывается как результат деления количества баллов, набранных за ДЗ студентом, на максимальное количество баллов за все ДЗ. Опоздание в пределах одного дня (24 часа) – штраф 1 балл. При опоздании более чем на 2 дня (48 ч. после дедлайна) за данное ДЗ выставляется оценка 0. НО для получения содержательных комментариев можно сдать ВСЕГДА (конечно, не позднее окончания курса).

Правила прохождения дистанции (2)

Семинарская активность

Семинарская активность оценивается по 10-балльной шкале. Баллы за семинарскую активность не возмещаются вне зависимости от причины пропуска семинаров.

Проверочные работы

В формате closed-book. В случае пропуска студентом проверочной работы по уважительной подтвержденной причине вес работы переносится на экзамен, в противном случае за нее выставляется оценка «0».

Правила прохождения дистанции (3)

Экзамен

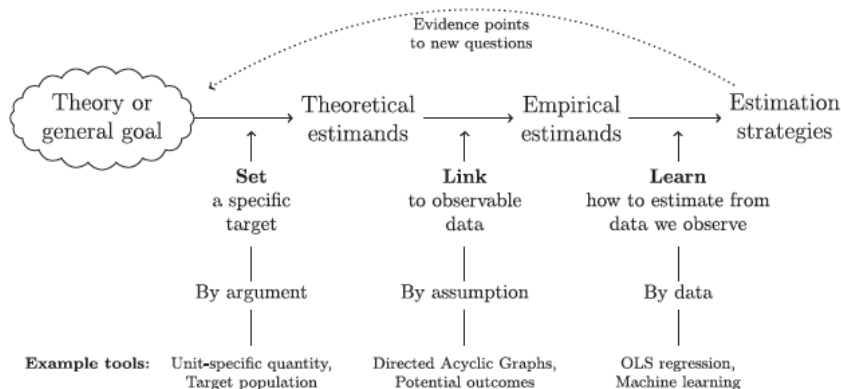
В экзаменационной работе могут встретиться задания повышенной сложности или задания, только частично разобранные на семинарах, верное решение которых предполагается для возможности получения оценок 9 и 10. Процент таких заданий не превышает 20%.

Что и когда сдавать?

ЭССЕ: 23.59 5 июня 2025

- Файл с текстом эссе. Лучше всего в .pdf
- Python-notebook (проверьте на реплицируемость без ошибок)
- База данных

Источник: Lundberg et al., 2021



Как организовать свой творческий порыв (1)

ЭССЕ (примерно 20000 – 30000 знаков с пробелами)

Как организовать свой творческий порыв (1)

ЭССЕ (примерно 20000 – 30000 знаков с пробелами)

- Abstract (кратко о чем Ваша работа, основные результаты, можно упомянуть о данных и методах)

Как организовать свой творческий порыв (1)

ЭССЕ (примерно 20000 – 30000 знаков с пробелами)

- Abstract (кратко о чем Ваша работа, основные результаты, можно упомянуть о данных и методах)
- Введение
 - ▶ Ключевой вопрос
 - ▶ Кто еще задавался таким вопросом? Кратко – об их находках.
 - ▶ Чем ответы предыдущих исследователей Вам не угодили? (неполные / есть ошибки – неточности)

Как организовать свой творческий порыв (1)

ЭССЕ (примерно 20000 – 30000 знаков с пробелами)

- Abstract (кратко о чем Ваша работа, основные результаты, можно упомянуть о данных и методах)
- Введение
 - ▶ Ключевой вопрос
 - ▶ Кто еще задавался таким вопросом? Кратко – об их находках.
 - ▶ Чем ответы предыдущих исследователей Вам не угодили? (неполные / есть ошибки – неточности)
- Гипотезы (формулируются эмпирически тестируемые предположения, НЕ ПУТАТЬ со статистическими гипотезами, множить гипотезы не стоит – не более 3)

Как организовать свой творческий порыв (2)

ЭССЕ (примерно 20000 – 30000 знаков с пробелами)

Как организовать свой творческий порыв (2)

ЭССЕ (примерно 20000 – 30000 знаков с пробелами)

- Данные и методы
 - ▶ Откуда данные? Опишите источники
 - ▶ Опишите данные: структура данных, размер выборки, описательные статистики, как измерены Ваши переменные и погрузите в контекст исследования: какой концепт они отражают
 - ▶ Обоснование методов, обсуждение альтернативных вариантов
 - ▶ Спецификация моделей, какую задачу решает оценивание той или иной модели

Как организовать свой творческий порыв (3)

ЭССЕ (примерно 20000 – 30000 знаков с пробелами)

Как организовать свой творческий порыв (3)

ЭССЕ (примерно 20000 – 30000 знаков с пробелами)

- Результаты

- ▶ В табличной форме, соответствующей зарубежным публикационным стандартам
- ▶ Описание результатов в «сухой» форме (просто то, что получилось, без критической оценки – оставьте это до раздела Discussion)
- ▶ Robustness checks (проверки результатов на устойчивость)

Как организовать свой творческий порыв (3)

ЭССЕ (примерно 20000 – 30000 знаков с пробелами)

- Результаты

- ▶ В табличной форме, соответствующей зарубежным публикационным стандартам
- ▶ Описание результатов в «сухой» форме (просто то, что получилось, без критической оценки – оставьте это до раздела Discussion)
- ▶ Robustness checks (проверки результатов на устойчивость)

- Обсуждение результатов и заключение

- ▶ Резюме полученных основных результатов и проверок на устойчивость
- ▶ Насколько соответствуют результаты предшествующей литературе, если есть расхождения, то почему
- ▶ Ограничения Вашего исследования