

Домашнее задание 2. Проведение собственного исследования с точки зрения эконометрики.

Формат:

- Подготовить презентацию со слайдами, оформить код на языке python и сделать доклад на 10-15 минут.
- Дата: 4 декабря.
- Сдавать можно в группах по 1-3 человека.
- Для подтверждения всех своих тезисов нужно прикладывать вырезки с таблицами или метриками из python (например, гипотеза о незначимости коэффициента X на 5% уровне значимости отвергается, потому что $p_value = 0.01 < 0.05$). Утверждения без отсылок к исходной статье не засчитываются. Устные заявления без отсылок не засчитываются.
- В день презентации до 17:00 прислать мне на почту презентацию и код. Слайды во время устного рассказа должны совпадать с присланной презентацией.
- Библиотеку statsmodels использовать необязательно, можно взять любую, которая вам больше нравится.
- При отсутствии по уважительным и доказанным в учебной части причинам формат сдачи обсуждается индивидуально.

Задание:

Собрать или найти любой датасет, который вам интересен, и провести полноценное исследование формата, как вы делали презентации в домашней работе 1.

Постановка задачи может звучать, например, так: исследователь собрал датасет из разных приложений со своего телефона и хочет проверить гипотезу: как количество пройденных в день шагов влияет на уровень сна (по 100 балльной шкале).

В статье в качестве одного из методов численного анализа должна присутствовать классическая линейная регрессия (методы анализа временных рядов, панельных данных, бинарного выбора, и другие не подходят).

Не нужно подделывать результаты: если подразумеваете наличие мультиколлинеарности, докажите это тестами и скажите, к чему она приведет в модели.

Не нужно собирать модель до тех пор, пока она будет идеальная и без проблем, если они там, очевидно, есть.

В презентации нужно ответить на все вопросы из списка (если вопрос для вашего исследования нерелевантен, нужно так об этом и сказать):

- 1) Выбрать любой готовый датасет для исследования или собрать самому. Идеи: любые датасеты с Kaggle, Росстат, Центральный банк, hh, cian и тд. Рассказать о нем. Уникальный самостоятельный датасет оценивается дополнительным баллом. (0-1 балл + (1 extra))
- 2) Сформулировать словесно гипотезы касаясь влияния чего-то, которые вы хотите проверить. Например, как количество пройденных в день шагов влияет на уровень сна (по 100 балльной шкале). (0-1 балл)
- 3) Описать статистические характеристики ваших данных (среднее, гистограммы и тп). Показать графики и обсудить, что вы на них видите. (0-1 балл)
- 4) Оценить регрессию с помощью метода МНК. (0-1 балл)
- 5) Дайте интерпретацию оцененным коэффициентам в регрессии. Совпали ли знаки оцененных коэффициентов с выдвинутыми гипотезами? (0-1 балл)
- 6) Обсудите значимость коэффициентов регрессии и их p-value? (0-2 балла)
- 7) Оценить метрики качества регрессии (0-1 балл)
- 8) Рассмотреть проблему мультиколлинеарности. Если она есть, то будете что-то делать с ней или нет? (0-2 балла)
- 9) Рассмотреть проблему гетероскедастичности. Если она есть, то будете что-то делать с ней или нет? (0-2 балла)
- 10) Рассмотреть проблему автокорреляции ошибок. Если она есть, то будете что-то делать с ней или нет? (0-1 балл)
- 11) Рассмотреть другие наборы переменных, если есть подозрения, что они нужны (квадраты, логарифмы, лаги факторов) (0-1 балл)
- 12) Построить тестовые прогнозы с доверительными интервалами (0-1 балл)
- 13) Дать заключение исследованию (0-1 балл)