1 Данные

В этом домашнем задании вам предстоит работать с данными о лесных пожарах. Вы будете оценивать зависимость сгоревшей площади (area) от различных погодных условий. Данные и их описание можно найти по ссылке:
https://archive.ics.uci.edu/ml/datasets/Forest+Fires

2 Правила

В качестве решения принимаются следующие виды файлов:

1. Jupyter notebook
2. Rmd-файл с чанками кода, скомпилированный в PDF
3. R-файл с ответам в виде комментариев

Задачи одинаково решаемы на любом из двух языков программирования. Сдача скрипта без комментариев и/или плагиат кода караются нулём баллов. Использование Stata карается нулём баллов. Всего за работу можно получить 10 баллов максимум. Бонусами можно покрыть недочёты в других номерах, но имейте в виду, что они могут быть сложнее обычных задач. Куда деть избыток бонусов мы пока не придумали.

Все таблицы и графики должны быть подписаны. Все оси у графиков тоже. Графики без осей/комментариев и прочих обязательных атрибутов оцениваться не будут. На защите диплома навряд ли кто-то станет проверять ваши числа в таблицах, а вот отсутствие подписи будет резать глаз. Если в задаче не указана визуализация, но вы считаете, что она необходима, ни в чём себе не отказывайте.

Не забывайте фиксировать seed для воспроизводимости результатов.

3 Задания

1. 0 баллов

   Подготовьте данные. Для простоты отфильтруйте нулевые значения целевой переменной, так как модели с цензурированными выборками в эту работу не входят. Преобразуйте категориальные переменные, если считаете это необходимым. Сделайте иные преобразования данных, если считаете их необходимыми (например, бинаризация признаков или логарифмирование). Обоснуйте свои преобразования.

2. 1.5 балла

   Отберите признаки, которые вы включите в модель. Кратко обоснуйте, почему вы берёте каждый признак, и какой знак вы ожидаете получить при соответствующей переменной в линейной модели. Нет ограничений по количеству и набору переменных, важно, чтобы он был логичен и обоснован. Однко желательно принять во внимание нелинейные признаки и признаки взаимодействия.

   Подготовьте описательные статистики данных. Обязательные: среднее, медиана, дисперсия, минимум, максимум. Дополнительные – любые по желанию. Например, отвергается ли гипотеза о нормальности теста Шапиро-Уилка и что-то ещё.

   Визуализируйте ваши признаки. Обязательно: гистограммы и ящики с усами. Иное – по желанию. Проинтерпретируйте графики. Есть ли в данных выбросы или иные аномалии? Если да, то как вы планируете с ними работать.

3. 1 балл

   Проверьте, есть ли в ваших данных мультиколлинеарность. Используйте для этого VIF и CN. Если мультиколлинеарность присутствует, разберитесь с ней как настоящие ковбои.

4. 0.5 балла

   Оцените линейную модель и проинтерпретируйте результаты. Как можно объяснить знаки, которые не совпали с вашими исходными предположениями? Получились ли переменные значимы по отдельности и в целом? В данном пункте не следует ожидать высокого R^2 из-за особенностей данных. Игнорируйте этот факт. Такие мелочи не должны вас волновать, ведь вы занимаетесь наукой.

   Протестируйте остатки на нормальность любым из тестов.

   Бонус 1 балл.

   Если остатки вашей модели получились не нормальными, вероятно предпосылка о нормальности случайных ошибок тоже нарушается. Воспользуйтесь процедурой бутстрепа и получите эмпирические доверительные интервалы. Алгоритм описан в этом видео: https://www.youtube.com/watch?v=UBSExb568B8 Обратите внимание, что случайная выборка генерируется с возвращением.

5. 0.5 баллов

   Постройте следующие прогнозы: точечный, индивидуальный и для среднего. В качестве значений независимых переменных возьмите медианные значения наблюдений.

6. 0.5 баллов

   Предположите, какие из регрессоров могут порождать гетероскедастичность. Обоснуйте идейно зависимость между этими переменными и дисперсией ошибок.

7. 1 балл

   Попытайтесь выявить гетероскедастичность по предположенным ранее переменным двумя способами: графически с помощью остатков регрессионной модели и вручную проведите тест Голдфельда-Квандта. Если не найдёте имплементации в любом из языков (маловероятно), реализуйте тест самостоятельно. Совпали ли результаты?

8. 1.5 балла

   Вне зависимости от того, обнаружена ли гетероскедастичность в предыдущих пунктах, вручную (в матрицах) оцените модель взвешенного МНК. Разумеется, предположите, что ковариационная матрица ошибок диагональна. Засчитывается как использование пакетных реализаций, так и решение вручную в матрицах, но в обоих случаях надо предоставить коэффициенты, стандартные ошибки и p-value.

   Изменились ли результаты значимости по сравнению с обычной линейной моделью?

9. 1 балл

   Вычислите робастные ошибки в форме Уайта (HC0), лучше вместо пакетных реализаций построить вручную. Поясните метод расчёта (можно на словах, главное - идея). Вычислите p-value для всех переменных. Изменились ли результаты значимости по сравнению с обычной линейной моделью?

   Бонус 1 балл

   Повторите для HC3

10. 1.5 балл

    Преобразуйте ваши независимые переменные с помощью PCA. Какую долю дисперсии объясняют две первые главные компоненты? Постройте линейную регрессию зависимой переменной на две первые главные компоненты и проинтерпретируйте результат. Повысилась ли объясняющая способность модели относительно обычной линейной? Значимы ли переменные?

11. 1 балл Весёлая задача на десятку

    С помощью метода максимального правдоподобия получите оценки линейной модели. Для этого в явном виде в матричной форме выпишите функцию правдоподобия.Далее найдите оценки коэффициентов (дисперсии ошибок – по желанию) двумя способами:

    (a) Решите задачу аналитически (совпадёт с МНК, просто перепишите предыдущий результат)

    (b) Решите задачу с помощью методов численного дифференцирования. В случаях когда градиент трудно выписать в явной форме, его можно вычислить приближённо, через приращения. Пример кода для функции нескольких переменных приложен в письме с домашним заданием, обобщите его на векторный случай. Напишите функцию, вычисляющую приближённый вектор-градиент правдоподобия в точке. Так как эта задача выпуклая, отлично справится обычный градиентный спуск без дополнительных наворотов. Если кто-то забыл или не знал, что это такое:
    https://github.com/esokolov/ml-course-hse/blob/master/2019-fall/lecture-notes/lecture02-linregr.pdf

    Похожи ли результаты?