1 Данные

В этом домашнем задании вам предстоит работать с данными о лесных пожарах. Вам предстоит оценить зависимость сгоревшей площади от различных погодных условий. Описание данных можно найти по ссылке:

http://www.dsi.uminho.pt/pcortez/forestfires/forestfires-names.txt

Сами данные лежат здесь:

http://www.dsi.uminho.pt/ pcortez/forestfires/forestfires.csv

Сдаваемая работа должна включать в себя файл Word или PDF(LaTex) с ответами на заявленные вопросы и скрипт на R. Сдача только скрипта и/или плагиат кода караются нулём баллов. Скрины формул из LaTex, вставленные в Word, караются нулём баллов.

Результаты оценки всех моделей в соответствующих пунктах необходимо представить в виде таблицы. Можно и нужно использовать для этого выдачу из R. Для преобразования таблиц в теховский код вам пригодятся пакеты texreg и xtable. В таблицах должны присутствовать названия переменных, оценки коэффициентов, стандартные ошибки и p-value.

2 Задания

1. 0 баллов

   Подготовьте данные. Для простоты отфильтруйте нулевые значения целевой переменной, так как модели с цензурированными выборками вы ещё не проходили.

2. 1 балл 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.

3. 0.5 балла

   Оцените линейную модель и проинтерпретируйте результаты. Как можно объяснить знаки, которые не совпали с вашими предположениями?

4. 1 балл

   Постройте прогноз для среднего и индивидуальный прогноз. В качестве значений независимых переменных возьмите медианные значения наблюдений.

5. 0.5 баллов

   Предположите какая из независимых переменных может порождать гетероскедастичность. Обоснуйте идейно связь переменной и дисперсии ошибок.

6. 1 балл

   Попытайтесь выявить гетероскедастичность по предположенной ранее переменной двумя способами: графически с помощью остатков регрессионной модели и вручную проведите тест Голдфельда-Квандта. Совпали ли результаты?

7. 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. Требуется вычислить оценки коэффициентов, их стандартные ошибки и p-value. Представьте результат

в виде таблицы. Также в тексте работы необходимо выписать формулы, по которым посчитаны оценки коэффициентов и их стандартные ошибки. Изменились ли результаты значимости по сравнению с обычной линейной моделью.

8. 1.5 балла

   Вычислите робастные ошибки в форме Уайта (без кросс-валидации). Поясните метод расчёта. Вычислите p-value для всех переменных. Результат оформите в виде таблицы Изменились ли результаты значимости по сравнению с обычной линейной моделью?

9. 1 балл

   Проверьте, есть ли в ваших данных мультиколлинеарность. Используйте для этогого VIF и CN.

10. 1 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.

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 двумя способами:

    (a) Решите задачу аналитически

    (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.

    Похожи ли получились результаты?