

Для выполнения проекта используйте данные RLMS. Волна остается та же, что и в домашнем задании. Можно поменять используемые данные на любые другие и сформулировать другой исследовательский вопрос.

1. Опишите используемые данные, если приняли решение поменять исследовательский вопрос.
2. Сформулируйте исследовательский вопрос (повторите старый или сформулируйте новый), который можно изучить с помощью ваших данных.

Достаточно сформулировать вопрос о наличии статистической связи между переменными. Стоит осознавать, что обнаружение причинно-следственных связей — очень трудная задача, браться за неё мы не требуем, хотя и не запрещаем.

В постановке вопроса в качестве зависимой переменной должна использоваться количественная переменная. Среди предикторов должно быть минимум две количественных переменных и минимум одна факторная.

3. Дайте краткую характеристику используемым переменным. Приведите описательные статистики. Изобразите ваши наблюдения в осях первых двух главных компонент, выделив их из количественных регрессоров и зависимой переменной.
4. Оцените две регрессионных модели: простую и более общую сложную. Это могут быть как те же, модели, которые вы использовали в домашнем задании, так и новые.

Для сложной модели проинтерпретируйте коэффициенты и постройте доверительные интервалы для них. С помощью теста сделайте выбор между простой и сложной моделью.

При проведении тестов и построении доверительных интервалов используйте робастные к гетероскедастичности методы.

5. Проведите тест Рамсея для отобранной модели. Проинтерпретируйте результат тестирования.
6. Для отобранной модели проверьте наличие мультиколлинеарности.
7. Сравните модели с LASSO по прогнозной силе.

Поделите выборку на две части: обучающую и тестовую. На обучающей оцените простую модель, сложную модель и модель LASSO. Постройте прогнозы на тестовую часть выборки. Выберите наилучшую из моделей.

8. Для финальной модели и выдуманного нового наблюдения постройте прогноз и предиктивный интервал для прогноза.
9. Проверьте наличие гетероскедастичности графически и с помощью хотя бы одного формального теста.
10. Описанные выше пункты являются минимальным требованием. Можно сделать больше! Удачи :)
11. Работу следует представить в виде отчёта в pdf или html формате. В начале работы должен идти текст с графиками, в конце работы в качестве приложения должен идти код. Общий объем текста (без приложений) должен составлять не более 10 страниц.

Дедлайн сдачи - 6 декабря 2021, 20:59. До указанного времени файл в формате pdf должен быть загружен по ссылке <https://www.dropbox.com/request/YbF0bdSMZYWalPxJqZec>.

Если RLMS надоел, то можно посмотреть в сторону:

- Russian Federal State Statistic Service, <https://rosstat.gov.ru/>;
- Russian Federal State Statistic Service (Сборник Регионы России), <https://rosstat.gov.ru/folder/210/document/13204>;
- Joint Economic and Social Data Archive HSE, <http://sophist.hse.ru/>;
- World Bank Open Data, <https://data.worldbank.org/>;
- OECD Data, <https://data.oecd.org/>.