Лабораторная работа №1. Тема: Прогнозирование моделью линейной регрессии

Рассмотрим пример построения линейной регрессионной модели на основе информации об **ожидаемой продолжительности жизни мужчин, число лет (y)**, рождаемости населения на 1000 человек (x_1), смертности населения на 1000 человек (x_2), числе браков на 1000 человек (x_3), числе разводов на 1000 человек (x_4), коэффициенте младенческой смертности (x_5), соотношении денежного дохода и прожиточного минимума, % (x_6), соотношении средней оплаты труда и прожиточного минимума трудоспособного населения, % (x_7), численности населения с денежными доходами ниже прожиточного минимума в % от численности населения (x_8), числа зарегистрированных преступлений на 100000 населения (x_9)

Разбивка признаков дана по административно-территориальному делению по состоянию на 2013 год. Регионы сгруппированы по Общероссийскому классификатору экономических регионов (ОКЭР) по состоянию на 2013 год.

Группа регионов, по которым вам нужен осуществить прогнозирование, определяется на основе вариантов, которые приведены на втором листе в файле с исходными данными. Номер варианта будет выложен в Google Sheets на отдельном листе. Данные по самим ОКЭР не берем, только по конкретным регионам.

Работа выполняется в Python (предпочтительно Jupyter Notebook – можно делать в https://colab.research.google.com/ , если есть проблемы с установкой через Anaconda). Выполненную работу необходимо прислать на адрес KulaevMA@yandex.ru. Тема письма должна быть оформлена следующим образом: [МГТУ][ПАД2023][ЛР1][ФАМИЛИЯ]. В письме должны находиться три файла: исходные данные по вашему варианту, ipynb файл с откомментированным кодом, реализующий все этапы лабораторной, текстовый отчет.

Текстовый отчет выполняется в формате «по ГОСТу», в том числе с основными элементами как Цель, Формулировка, Основная часть, Заключение и тд. Основная часть может быть разбиты на логичные подэтапы. Скриншоты получаемых результатов и отсылки на код должны присутствовать в тексте. Каждая часть отчета должна быть подкреплена сущностными выводами (**интерпретацией результатов!**), например, исходя из значений коэффициентов заметно, что наибольший положительный вклад величиной 10 в продолжительность жизни вносит доля лиц, злоупотребляющих алкоголем.

Основные этапы выполнения работы:

1. Нормирование (масштабирование) исходных данных.
2. Расчет весов линейной регрессии по аналитической формуле – предлагается использовать numpy.
3. Построение и интепретация корреляционной матрицы. Определение степени мультиколлинеарности на основе числа обусловленности. Для их расчета не нужно писать самописные методы, предлагается использовать scipy.
4. Анализ регрессионных остатков.
5. Определение весов линейной регрессии градиентным методом – предлагается использовать numpy. Проанализировать изменение ошибки от итерации к итерации.
6. Сравнение результатов по аналитическому и градиентному методу.
7. С помощью библиотеки sklearn сделать fit-predict модели линейной регрессии. Сравнить результаты с ранее полученными.
8. С помощью библиотеки statmodels получить «эконометрический» результат обучения модели линейной регрессии. Проинтерпретировать все его составляющие (в т.ч. те, которые изучались только теоретически), сравнить с предыдущими результатами. Пример отчета - https://www.statsmodels.org/stable/index.html P.S. Регрессионным статистическим гипотезам будет уделено особое внимание на защите.
9. Сравнить качество получаемых моделей на основе коэффициента детерминации и MSE.
10. Сделать итоговый вывод касательно причин различия в результатах при выполнении работ разными методами, а также по получаемым моделям в целом. Провести сравнительный анализ.