Прикладная статистика и анализ данных.

Задание 2.

- Дедлайн **25 февраля 23:59**. После дедлайна работы не принимаются кроме случаев наличия уважительной причины.
- Выполненную работу нужно отправить на почту `mipt.stats@yandex.ru`, указав тему письма "`[asda] Фамилия Имя - задание 2`". Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- По задачам 2-4 прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `2.N.ipynb` и `2.N.pdf`, где `N` — ваш номер из таблицы с оценками. Задачи 2 и 4 выполняются на языке R. Задачу 3 можно выполнить в Питоне в отдельном ноутбуке.
- Задачу 1 необходимо оформить в tex'е и прислать `pdf` или же прислать фотку в правильной ориентации рукописного решения, где все четко видно.
- Решения, размещенные на каких-либо интернет-ресурсах не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Не забывайте делать пояснения и выводы.

1. **(2 балла)** Во взвешенном методе наименьших квадратов каждому наблюдению задается некоторый известный вес w_i. Задача имеет вид $\sum\limits_{i=1}^{n} w_i \left(Y_i - x_i^T\theta\right)^2 \to \min\limits_{\theta}$. Найдите решение задачи в матричном виде.
2. **(4 балла)** Скачайте данные о стоимости квартир в Москве:

   https://raw.githubusercontent.com/bdemeshev/em301/master/datasets/flats_moscow.txt

   Описание данных доступно по ссылке

   https://github.com/bdemeshev/em301/blob/master/datasets/flats_moscow_description.txt

   Обучите линейную регрессионную модель для предсказания цены квартиры от всех других параметров. Проверьте гипотезы о незначимости признаков и постройте доверительные интервалы для коэффициентов модели. Определите признаки желаемой для себя квартиры в Москве и постройте предсказательный интервал ее цены.
3. **(4 балла)** Проведите эксперимент по определению реального уровня значимости критерия для проверки гипотезы о незначимости коэффициента в гауссовской линейной модели, если на самом деле в данных присутствует гетероскедастичность. Для этого смоделируйте некоторым образом двумерные данные x и посчитайте по ним ожидаемый отклик $y(x) = \theta_0 + \theta_1 x^{(1)} + \theta_2 x^{(2)}$, где коэффициенты выберите по своему усмотрению, причем $\theta_2 = 0$. Зашумите набор значений $y(x_i)$ некоторым шумом, дисперсия которого зависит от x или от номера наблюдения. По таким данным обучите линейную модель и проверьте гипотезу $\mathsf{H}_0\colon \theta_2 = 0$. Повторите эксперимент несколько раз и посчитайте долю случаев, в которых гипотеза отвергается. Распределение шума должно быть одинаковым в каждом эксперименте.
4. См. ноутбук.