Домашнее задание № 6

Регрессионный анализ.

Крайний срок сдачи: 23 декабря 2020 г. (до конца дня).

Каждое задание оценивается в 2 балла.

1

N1 Рассмотрим базу данных "LifeCycleSavings"(`https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/LifeCycleSavings.html`), содержащую информацию о среднем коэффициенте персональных сбережениях жителей 50 стран. Этот коэффициент для конкретного жителя вычисляется как отношение его совокупных личных сбережений к располагаемому доходу. Согласно гипотезе Модильяни, среднее по стране значение этого коэффициента зависит от

- процента населения моложе 15 лет (LifeCycleSavings$pop15);
- процента населения старше 75 лет (LifeCycleSavings$pop75);
- располагаемого дохода на душу населения (LifeCycleSavings$dpi);
- процентной скорости изменения располагаемого дохода на душу населения (LifeCycleSavings$ddpi).

Представленные данные являются усреднёнными показателями за 1960–1970 гг.

(i) Для переменных "sr"(как y-переменной) и "pop15" (как x-переменной) постройте ядерную оценку регрессии при различных вариантах выбора ядра (гауссовское ядро и ядро Епанечникова) и различных методах выбора параметра bandwidth (критерий Акаике, обобщённый метод кросс-проверки). Найдите наилучший метод в смысле наименьшей среднеквадратичной ошибки.

(ii) Повторите вычисления для остальных трёх объясняющих переменных вместо "pop15". Выберите 2 переменные, которые по Вашему мнению наилучшим образом объясняют коэффициент персональных сбережений (в дальнейшем эти переменные будем называть V1 и V2). Объясните свой выбор.

(iii) На основе V1 и V2 постройте многомерную регрессию методом LOESS и линейную регрессию. Разделите случайным образом все страны на 2 группы: в одну группу отнесите примерно 80 % стран, в другую - 20 %. Оцените параметры модели LOESS и линейной регрессии по большей группе и проверьте качество моделей по меньшей. Выясните, какая из построенных моделей является более точной.

(iv) Найдите выбросы в переменных V1 и V2, используя их диаграммы размаха. Удалите соответствующие страны из таблицы и заново постройте многомерную регрессию методом LOESS и линейную регрессию (как в предыдущем пункте). Проверьте, улучшилось ли качество подгонки для оставшихся в таблице стран.

N2 Рассмотрим модель

$$Y_i = \sin(\pi X_i/2) + \varepsilon_i, \qquad i = 1..n,$$

где X_i последовательность i.i.d. случайных величин, равномерно распределённых на [0,1], и ε_i - i.i.d., имеющие нормальное распределение со средним 0 и дисперсией 0.01.

Данное задание направлено на имплементацию аналога метода LOESS, при котором усреднение ведётся по k ближайшим соседям. Более точно, алгоритм состоит из следующих шагов.

1. Определим количество соседей наблюдения номер i равным одному и тому же числу, $k_i = k$ (например, $k = 5$).

2. Для каждой точки X_i расположим точки $X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$ в порядке их отдалённости от X_i :

$$|X_{(i,1)} - X_i| \le |X_{(i,2)} - X_i| \le ... \le |X_{(i,n-1)} - X_i|$$

и определим оценку

$$\hat{r}_n(X_i) = \frac{1}{k_i}\sum_{j=1}^{k_i} Y_{(i,j)},$$

где $Y_{(i,j)}$ является наблюдаемым значением Y в точке $X_{(i,j)}$. Постройте график, на котором для каждой точки X_i будут отображаться наблюдаемые значения Y_i и оценённые значения $\hat{r}_n(X_i)$.

3. Для каждого $i = 1..n$, вычислите ошибки

$$e_i = \hat{r}_n(X_i) - Y_i$$

и определите

$$\delta_i = B\left(e_i\right),$$

где B - весовая функция, равная

$$B(x) = \begin{cases} \left(1 - x^2\right)^2, & |x| < 1, \\ 0, & |x| \ge 1. \end{cases}$$

4. Переопределите $k_i = round(k_i/\delta_i)$ (то есть для оценивания в точках, для которых e_i большое, мы используем усреднение по большему количеству наблюдений).

5. Повторите шаги 2-4 несколько раз (например, 10 раз).

Выясните, улучшается ли качество оценки от многократного повторения данного алгоритма.

2

T1 Пусть дан набор точек $(x_i, y_i), i = 1..n$. Для описания регрессионной зависимости между y_i и x_i будем использовать оценку Надарая-Ватсона

$$\hat{r}(x) = \frac{\sum_{i=1}^n y_i K\big((x - x_i)/h\big)}{\sum_{i=1}^n K\big((x - x_i)/h\big)}$$

с треугольным ядром

$$K(x) = (1 - |x|) \cdot \mathbb{I}\{|x| \le 1\}$$

и параметром $h > 0$. Для случая $n = 6$ и $x_i = i$, $\forall i = 1..6$, вычислите сглаживающую матрицу H и эффективное количество степеней свободы (след матрицы H), если

(i) $h = 1/2$;

(ii) $h = 3/2$.

Комментарий. Напомним, что слаживающая матрица H - это такая матрица, что

$$\hat{\vec{y}} = H\vec{y},$$

где $\vec{y} = (y_1, ..., y_n)^\top, \hat{\vec{y}} = (\hat{r}(x_1), ..., \hat{r}(x_n))^\top$.

T2 Рассмотрим модель линейной регрессии

$$\begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & ... & x_{1m} \\ ... & ... & ... \\ x_{n1} & ... & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_1 \\ ... \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_n \end{pmatrix} =: X\vec{\beta} + \vec{\varepsilon}$$

с $n \ge m$. Напомним, что ключевую роль при оценивании вектора $\vec{\beta}$ играет тот факт, что матрица $Q = X^\top X$ является обратимой.

(i) Докажите, что если $x_{ij} = u_i^{j-1}, i = 1..n, j = 1..m$, где $u_1, .., u_n$ - различные значения (полиномиальная регрессия), то столбцы матрицы X линейно независимы.

(ii) Докажите, что если столбцы матрицы X линейно независимы, то матрица Q положительно определена, то есть $\vec{v}^\top Q \vec{v} > 0$ для любого ненулевого вектора $\vec{v} \in \mathbb{R}^m$.

Комментарий. Обратите внимание, что в (ii) нужно доказать строгое неравенство.

T3 Рассмотрим простейшую линейную регрессию

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1..n,$$

где ε_i - i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ и $x_1, ..., x_n$ - детерминированные точки, хотя бы две из которых различны. Обозначим через $\hat{\alpha}, \hat{\beta}$ оценки параметров α, β, полученные методом наименьших квадратов

$$\big(\hat{\alpha}, \hat{\beta}\big) := \arg\min_{\alpha,\beta} \mathcal{R}(\alpha, \beta), \quad \text{где} \quad \mathcal{R}(\alpha, \beta) = \sum_{i=1}^{n} \big(y_i - \alpha - \beta x_i\big)^2$$

Докажите, что эти оценки имеют нормальное распределение и являются несмещёнными оценками параметров α и β.

T4* В обозначениях предыдущей задачи докажите, что

$$\frac{1}{n-2} \mathcal{R}(\hat{\alpha}, \hat{\beta})$$

является несмещённой оценкой параметра σ^2.