Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический иститут

Кафедра «Прикладная математика»

Отчёт по лабораторной работе №2 по дисциплине «Анализ данных с интервальной неопределённостью»

Выполнил студент:
Куксенко Кирилл Сергеевич
группа: 5040102/20201

Проверил:
к.ф.-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2023 г.

Содержание

Список иллюстраций

1 Постановка задачи

2 Теория

2.1 Точечная линейная регрессия

Рассматривается задача восстановления зависимости для выборки $(X, (Y))$, $X = \{x_i\}_{i=1}^n$, $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$, x_i - точеный, $\mathbf{y}_i$ - интервальный. Пусть искомая модель задана в классе линейных функций

$$y = \beta_0 + \beta_1 x \tag{1}$$

Поставим задачу оптимизацию 2 для нахождения точечных оценок параметров β_0, β_1.

$$\begin{gathered} \sum_{i=1}^{m} w_i \to \min \\ \mathrm{mid}\mathbf{y}_i - w_i \cdot \mathrm{rad}\mathbf{y}_i \le X\beta \le \mathrm{mid}\mathbf{y}_i + w_i \cdot \mathrm{rad}\mathbf{y}_i \\ w_i \ge 0, i = 1, ..., m \\ w, \beta - ? \end{gathered} \tag{2}$$

Задачу 2 можно решить методами линейного программирования.

2.2 Информационное множество

Информационным множеством задачи восстановления зависимости будем называть множество значений всех параметров зависимости, совместных с данными в каком-то смысле.

Коридором совместных зависимостей задачи восстановления зависимости называется многозначное множество отображений Υ, сопоставляющее каждому значению аргумента x множество

$$\Upsilon(x) = \bigcup_{\beta \in \Omega} f(x, \beta) \tag{3}$$

, где Ω - информационное множество, x - вектор переменных, β - вектор оцениваемых параметров.

Информационное множество может быть построено, как пересечение полос, заданных

$$\underline{\mathbf{y}_i} \le \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im} \le \overline{\mathbf{y}_i} \tag{4}$$

, где $i = \overline{1,n}\mathbf{y}_i \in \mathbf{Y}, x_i \in X$, X - точечная выборка переменных, $\mathbf{Y}$ - интервальная выборка откликов.

3 Реализация

Весь код написан на языке Python (версии 3.7.3). Ссылка на GitHub с исходным кодом.

4 Результаты

Данные были взяты из файлов *data/dataset1/+0_5V/+0_5V_0.txt*. Обынтерваливание было произведено следующим образом.

$$\mathbf{x}_i = [(x_i - \delta_i) - \varepsilon, (x_i - \delta_i) + \varepsilon], \varepsilon = \frac{100}{2^{14}} \tag{5}$$

где x_i - точечное значение, δ_i - точечная погрешность. Набор δ_i получен из соответствующих файлов в *data/dataset1/ZeroLine.txt*

Построим линейную регрессию и найдём информационное множество для двух выборок с разной степенью совместности.

Рассмотрим первую выборку X_1.

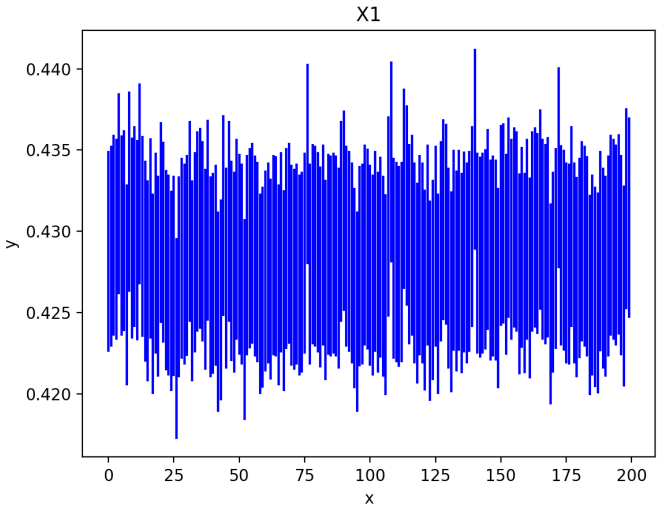

Рис. 1: Первая выборка, X_1

Индекс Жаккара первой выборки равен $JK(X_1) = 0.5115$ (в этой работе $JK(X) \in [0, 1]$).

Построим линейную регрессию, решив задачу 2 для выборки X_1.

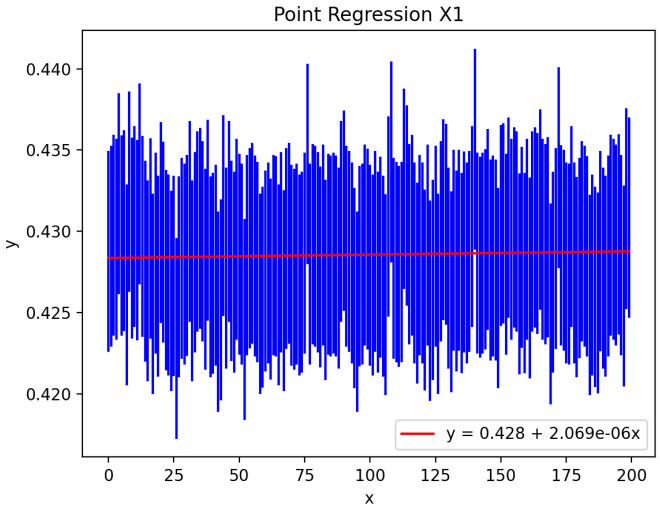

Рис. 2: Точечная линейная регрессия для X_1

Получим следующие оценки для параметров: $\beta_0 = 0.428, \beta_1 = 2.069e^{-6}$. Тогда полученная модель имеет вид $y = 0.428 + 2.069e^{-6}x$.

Найдём для данной выборки информационное множество.

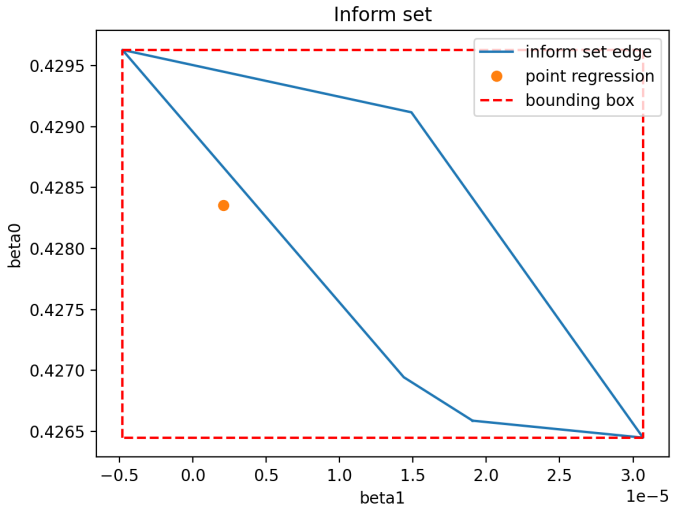

Рис. 3: Информационное множество для X_1

На рис. 3 можно заметит, что найденные параметры β_0, β_1 решением задачи 2 лежат вне информационного множества.

Построим коридор совместных значений для выборки X_1 и информационного множества 3 и оценим значения выходной переменной y вне пределов значений входной переменной x.

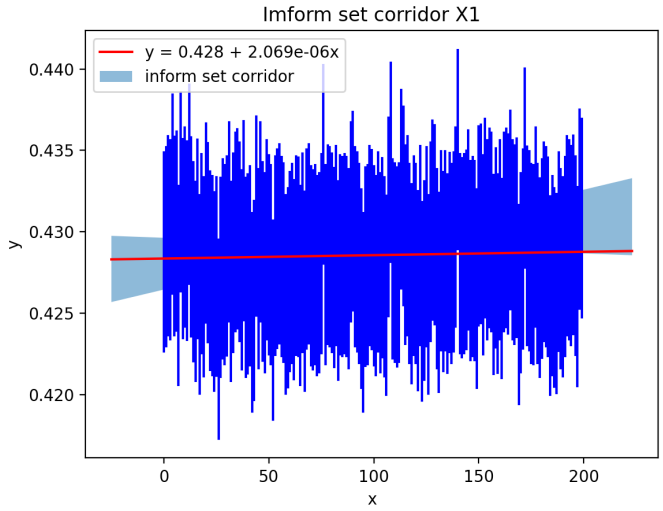

Рис. 4: Коридор совместных значений для X_1

На рис. 4 видно, что построенная точечная регрессия лежит вне коридора совместных значений, что согласуется с рис. 3.

Проведём аналогичные построения для выборки X_2, построенную аналогичным образом с X_1 за исключением отсутствия учёта δ_i. X_2 имеет вид.

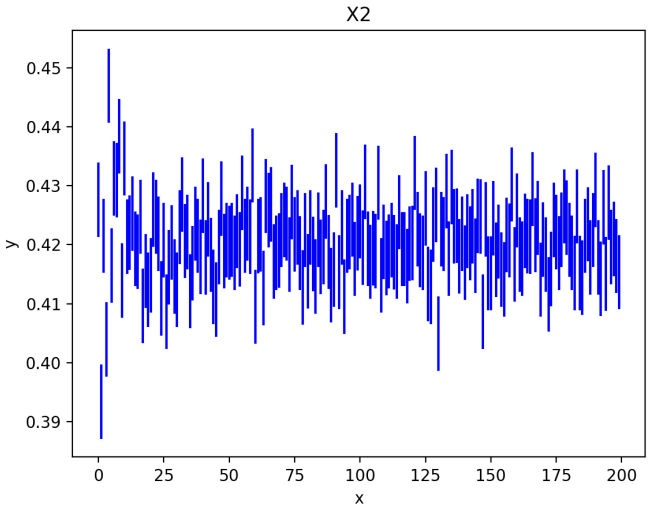

Рис. 5: Вторая выборка, X_2

Индекс Жаккра X_2 равен $JK(X_2) = 0.1855$.
Построим точечную линейную регрессию для X_2.

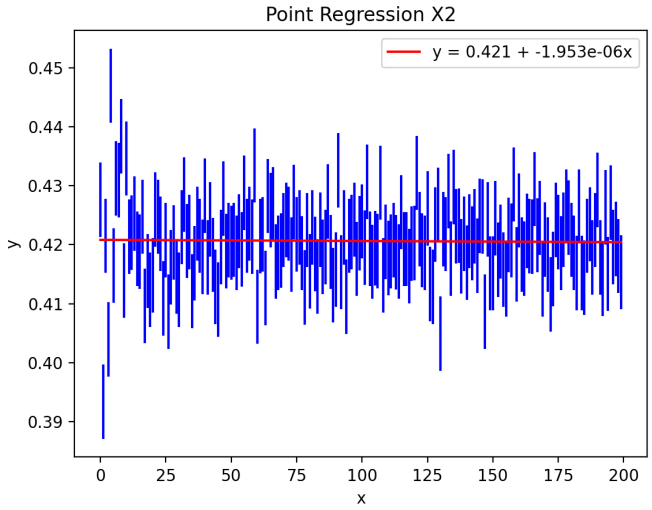

Рис. 6: Точечная линейная регрессия для X_2

Для X_2 получили следующие оценки параметров: $\beta_0 = 0.421, \beta_1 = -1.953e^{-6}$.

Построим информационное множество и коридор совместных значений для X_2.

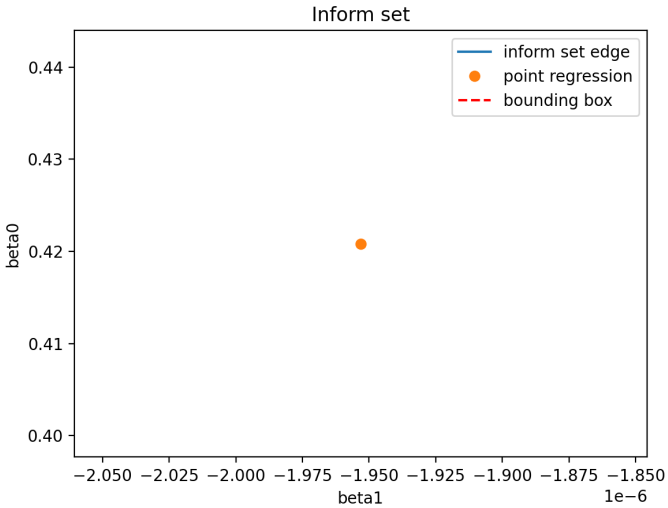

Рис. 7: Пустое информационное множество для X_2

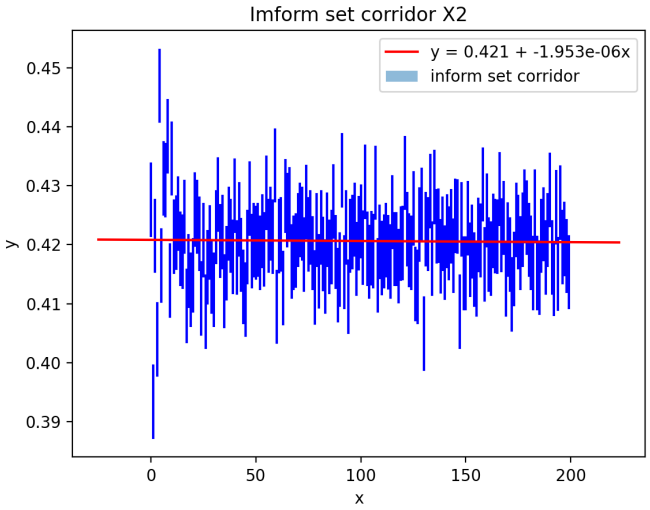

Рис. 8: Коридор совместных значений для X_2

В итоге для X_2 получили пустое информационное множество, и значит для X_2 и модели 1 не существует коридора совместных значений.

5 Обсуждение

Из полученных результатов можно заметить следующее. Может оказаться, что, в случае малой совместности или в случае отсутствия совместности, точечная регрессия не попадает в информационное множество, что видно на рис. 3, 4, 7. Также видно, что точечная регрессия может не пересекать все интервалы исходной выборки рис. 2. Стоит отметить, что с уменьшением степени совместности, размер информационного множества и ширина коридора совместности уменьшаются, и в определённый момент информационное множество может оказатся пустым. (рис. 3, 4, 7, 8), что вполне ожидаемо При этом для обоих выборок оценки параметров, полученных с помощью точечной линейной регрессии, мало отличаются. Также заметно, что ширина коридора сильно увеличивается за пределами значений входной переменной 4.