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 $(X, \mathbf{(}Y\mathbf{)})$, $X = \{x_i\}_{i=1}^n, \mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$, 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 S_X были взяты из файлов *data/dataset1/X/X_0.txt*, где $X \in \{-0\_5, -0\_25, +0\_25, +0\_5\}$. Набор δ_i получен из соответствующих файлов в *data/dataset1/ZeroLine.txt*.

Набор значений X точечный и одинаков для всех выборок. $X = [-0.5, -0.25, 0.25, 0.5]$. Набор значений отклика Y интервальный и разный для каждой выборки.

Построим линейную регрессию и найдём информационное множество для нескольких выборок.

Рассмотрим первую выборку Y_1. Y_1 следующим образом. $y_i = [\min_{t \in S_i} S_i, \max_{t \in S_i} S_i]$, $i = [-0.5, -0.25, +0.25, +0.25], y_i \in Y_1$.

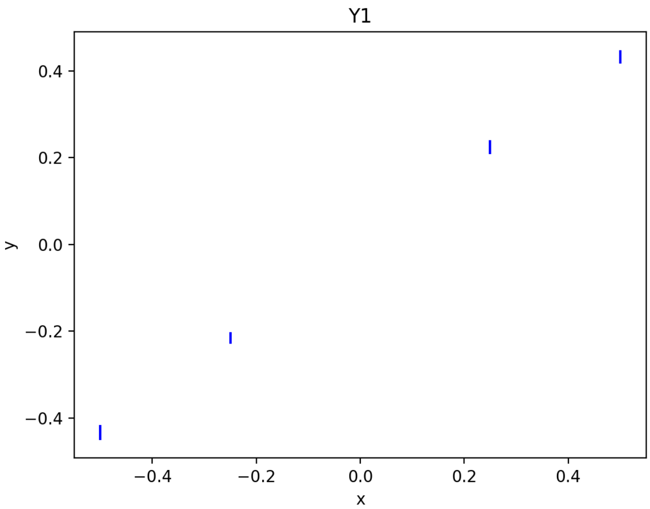

Рис. 1: Первая выборка, Y_1

Построим линейную регрессию, решив задачу 2 для выборки Y_1.

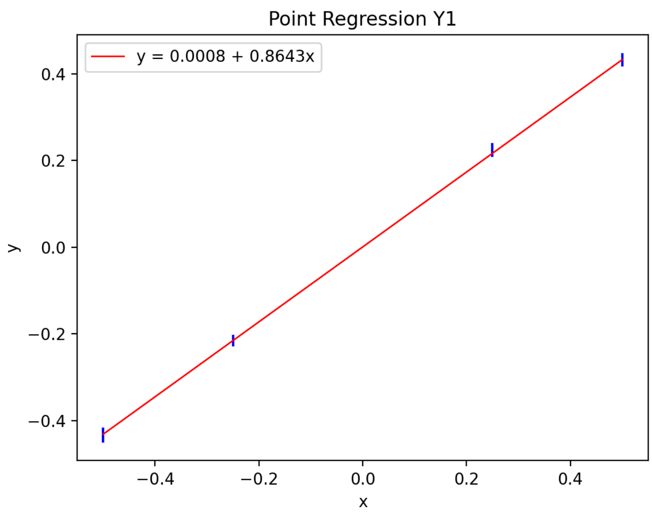

Рис. 2: Точечная линейная регрессия для Y_1

Получим следующие оценки для параметров: $\beta_0 = 0.00076, \beta_1 = 0.86426$. Тогда полученная модель имеет вид $y = 0.00076 + 0.86426x$.

Найдём для данной выборки информационное множество.

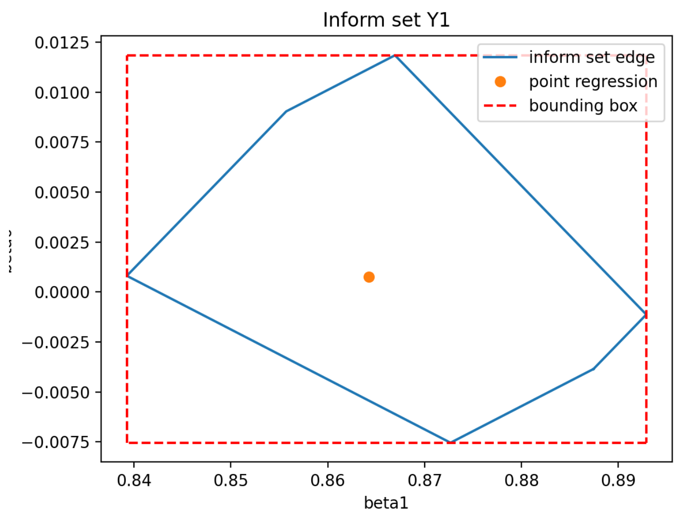

Рис. 3: Информационное множество для Y_1

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

Построим коридор совместных значений для выборки Y_1 и информационного множества 3 и оценим значения выходной переменной y вне пределов значений входной переменной x.

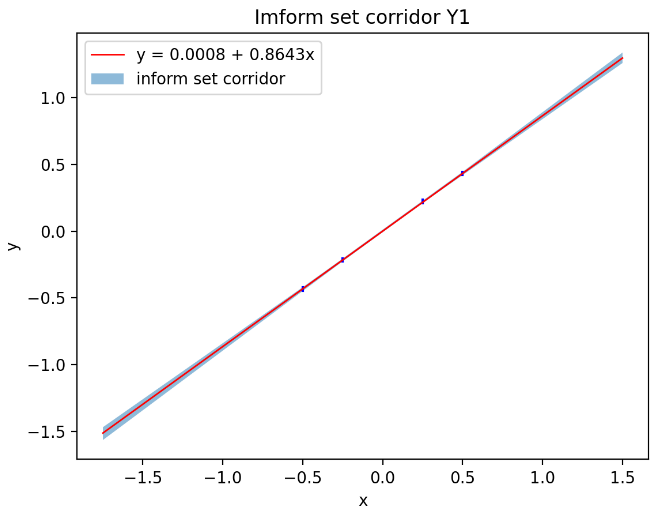

Рис. 4: Коридор совместных значений для Y_1

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

Проведём аналогичные построения для выборки Y_2, построенную следующим образом. $y_i = [median(S_i) - \varepsilon, median(S_i) + \varepsilon]$, $\varepsilon = \frac{1}{2^{14}}$ $i = [-0.5, -0.25, +0.25, +0.25], y_i \in Y_2$. Y_2 имеет вид.

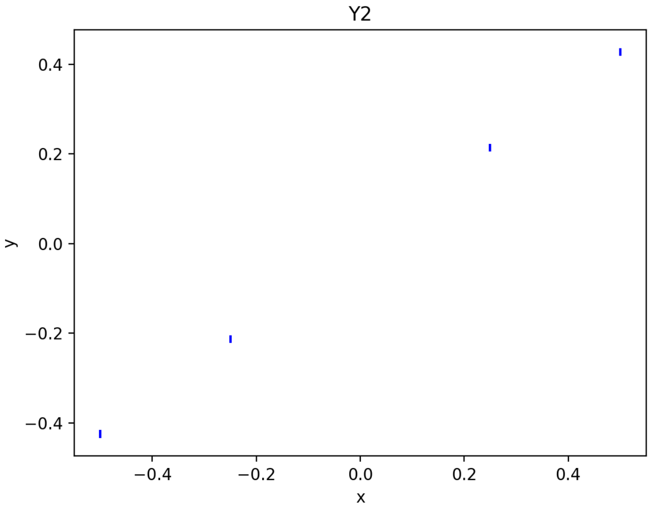

Рис. 5: Вторая выборка, Y_2

Построим точечную линейную регрессию для Y_2.

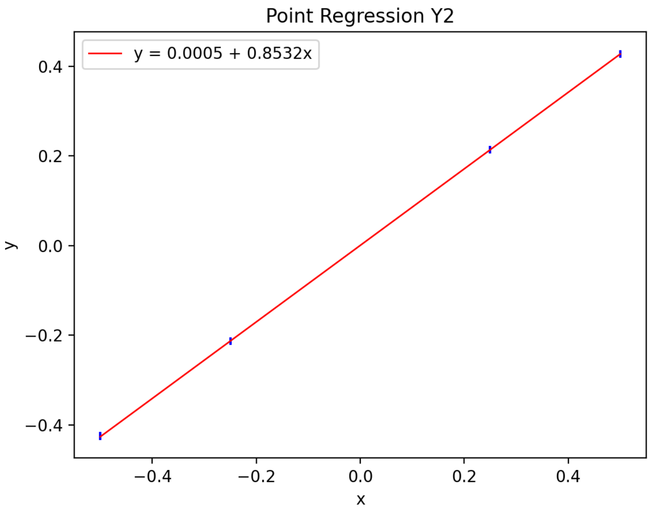

Рис. 6: Точечная линейная регрессия для Y_2

Для Y_2 получили следующие оценки параметров: $\beta_0 = 0.0005, \beta_1 = 0.85324$. Построим информационное множество и коридор совместных значений для Y_2.

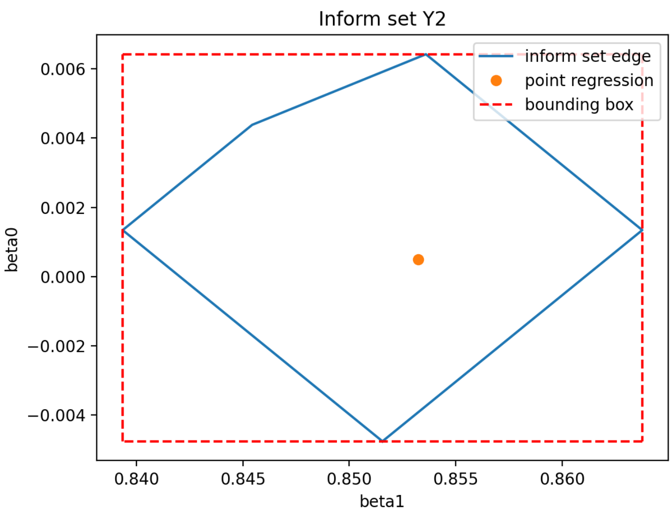

Рис. 7: Информационное множество для Y_2

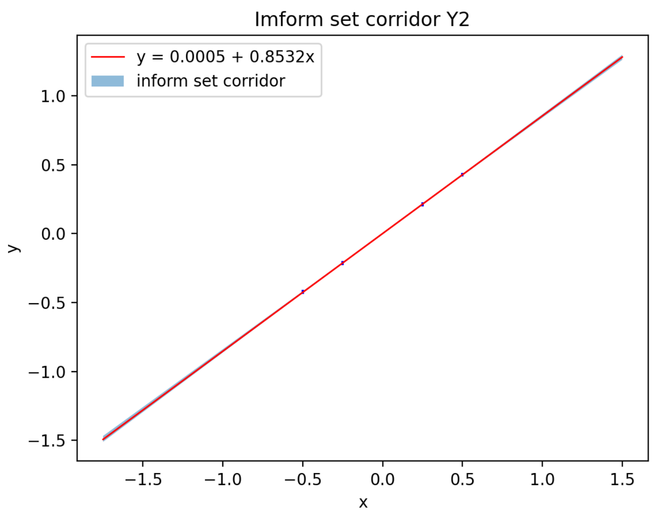

Рис. 8: Коридор совместных значений для Y_2

В итоге для Y_2 получили, что точечная регрессия также попала в информационное множество.

Теперь проведём аналогичные построения для Y_3, построенную аналогично Y_1, за исключением отсутствия учёта δ_i. Y_3 имеет вид.

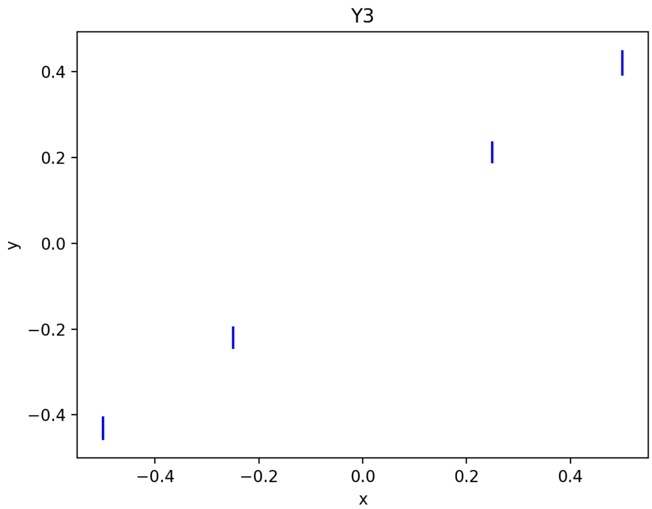

Рис. 9: Третья выборка, Y_3

Построим точечную регрессию.

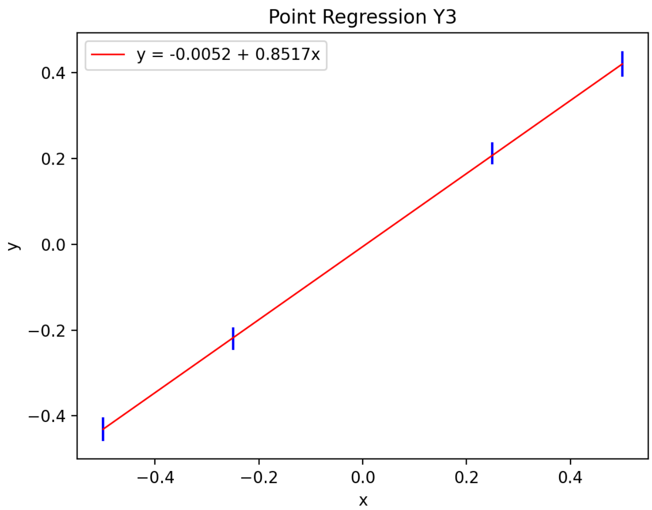

Рис. 10: Точечная линейная регрессия для Y_3

Для Y_3 точечная линейная регрессия дала следующие оценки: $\beta_0 = -0.0052, \beta_1 = 0.85169$. Информационное множество и коридор совместных значений имеют следующий вид.

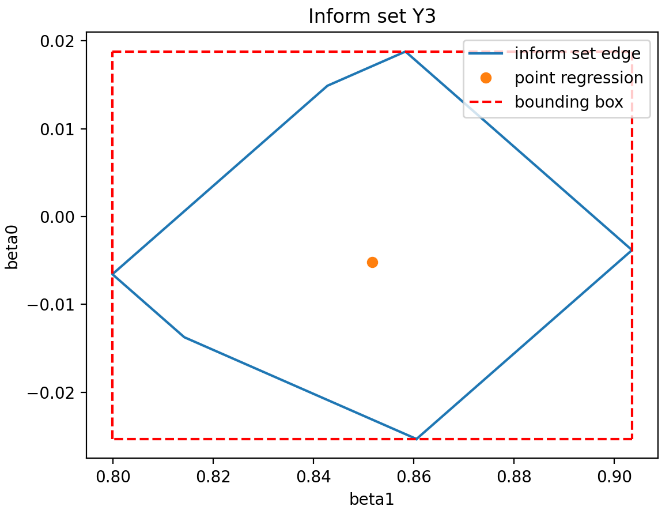

Рис. 11: Информационное множество для Y_3

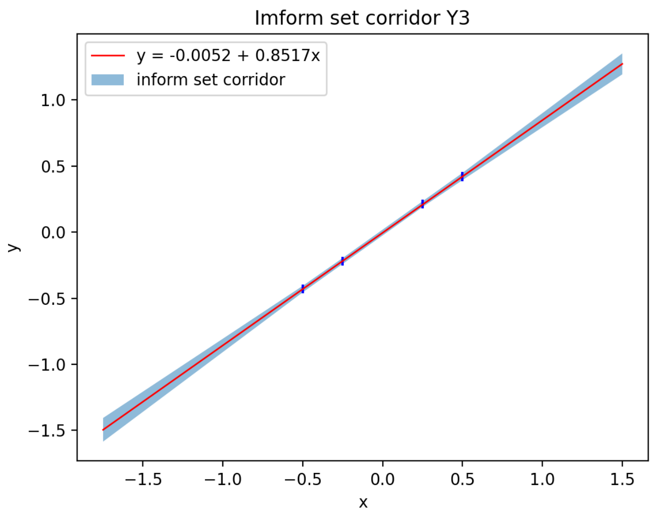

Рис. 12: Коридор совместных значений для Y_3

5 Обсуждение

Из полученных результатов можно заметить следующее. Наиболее маленькое информационное множество было получено для выборки Y_2 (рис. 3, 7, 11), что неудивительно, так как Y_2 имеет наименьшую интервальную неопределённость. Соответственно для Y_2 получили и наиболее узкий коридор совместных значений (рис. 4, 8, 12).

Видно, что для выборок Y_1, Y_2 точечная линейная регрессия дала более точный результат, близкий к ожидаемому $\beta_0 = 0.0, \beta_1 = 1.0$. Для Y_3 получили более неточную оценку, так оценка параметра β_0 для Y_3 отличается на порядок от соответствующей оценки для Y_1, Y_2.

Также стоит отметить, что во всех случаях точечная линейная регрессия попала в информационное множество.