

Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт

Кафедра «Прикладная математика»

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

Выполнил студент:
Габдрахманов Булат Маратович
группа: 5040102/20201

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

Санкт-Петербург
2023 г.

Содержание

1	Постановка задачи	2
2	Теория	2
2.1	Точечная линейная регрессия	2
2.2	Информационное множество	2
3	Реализация	3
4	Результаты	3
5	Заключение	11

Список иллюстраций

1	Первая выборка, Y_1	4
2	Точечная линейная регрессия для Y_1	4
3	Информационное множество для Y_1	5
4	Коридор совместных значений для Y_1	6
5	Вторая выборка, Y_2	7
6	Точечная линейная регрессия для Y_2	7
7	Информационное множество для Y_2	8
8	Коридор совместных значений для Y_2	8
9	Третья выборка, Y_3	9
10	Точечная линейная регрессия для Y_3	9
11	Информационное множество для Y_3	10
12	Коридор совместных значений для Y_3	10

1 Постановка задачи

Для заданных выборок необходимо найти точную линейную регрессию, информационные множества и коридоры решений и сравнить полученные результаты.

2 Теория

2.1 Точечная линейная регрессия

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

$$y = \beta_0 + \beta_1 x \quad (1)$$

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

$$\begin{aligned} \sum_{i=1}^m w_i &\rightarrow \min \\ \text{mid}y_i - w_i \cdot \text{rad}y_i &\leq X\beta \leq \text{mid}y_i + w_i \cdot \text{rad}y_i \\ w_i &\geq 0, i = 1, \dots, m \\ w, \beta &-? \end{aligned} \quad (2)$$

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

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

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

Информационное множество может быть построено, как пересечение полос, заданных

$$\underline{y}_i \leq \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} \leq \overline{y}_i \quad (4)$$

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

3 Реализация

Весь код написан на языке Python (версии 3.9). [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. $y_i = [\min_{t \in S_i} S_i, \max_{t \in S_i} S_i]$, $i = [-0.5, -0.25, +0.25, +0.25]$, $y_i \in Y_1$.

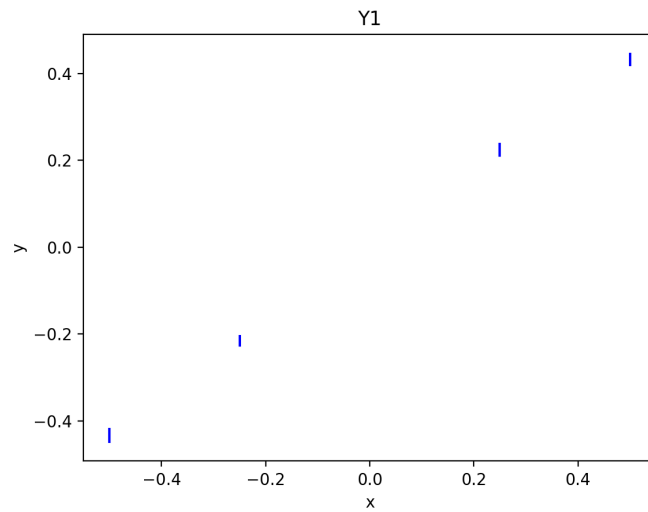


Рис. 1: Первая выборка, Y_1

Построим линейную регрессию, решив задачу 2 для выборки Y_1 .

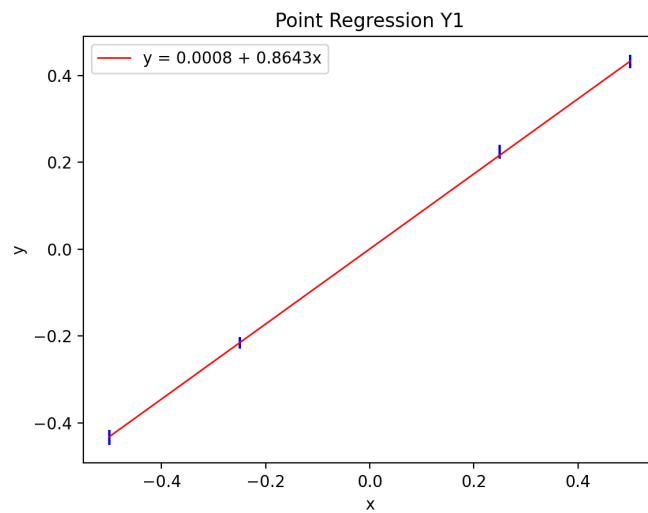


Рис. 2: Точечная линейная регрессия для Y_1

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.

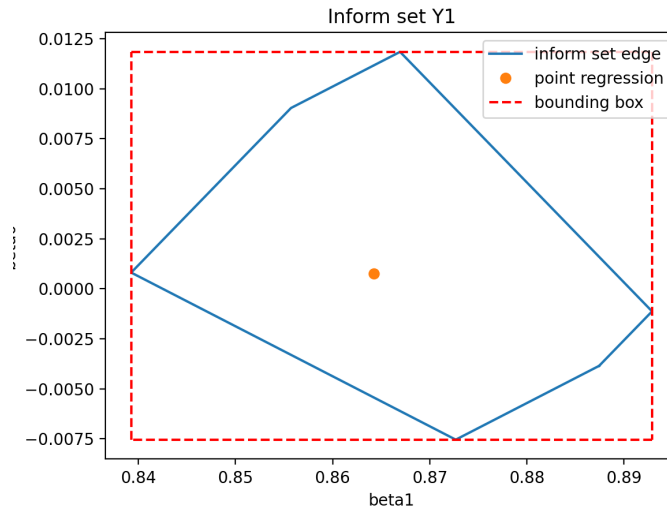


Рис. 3: Информационное множество для Y_1

На рис. 3 можно заметить, 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 .

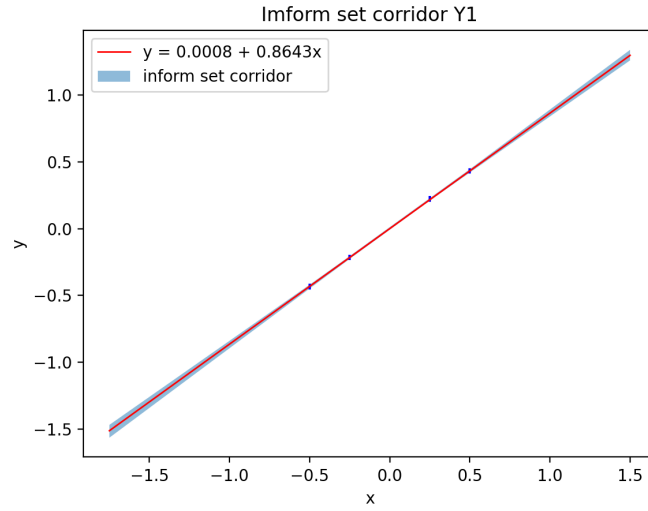


Рис. 4: Коридор совместных значений для Y_1

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. $y_i = [\text{median}(S_i) - \varepsilon, \text{median}(S_i) + \varepsilon]$, $\varepsilon = \frac{1}{2^{14}}$ $i = [-0.5, -0.25, +0.25, +0.25]$, $y_i \in Y_2$. Y_2 имеет вид.

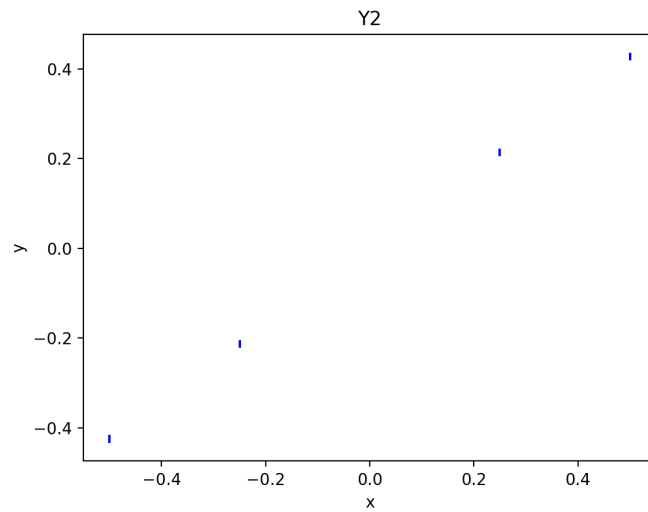


Рис. 5: Вторая выборка, Y_2

Построим точечную линейную регрессию для Y_2 .

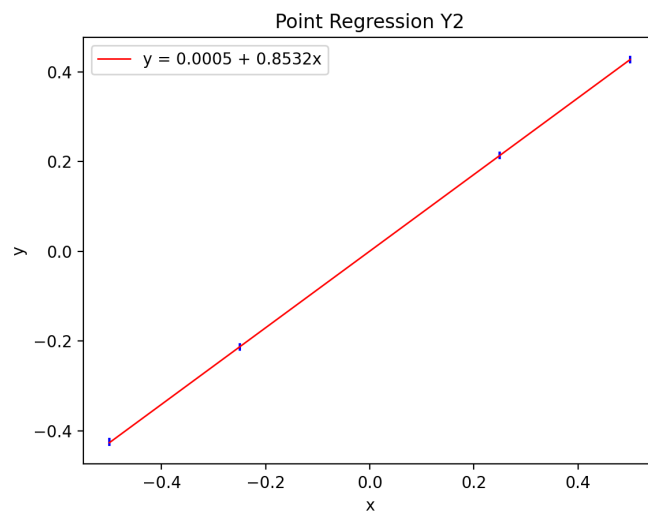


Рис. 6: Точечная линейная регрессия для Y_2

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 .

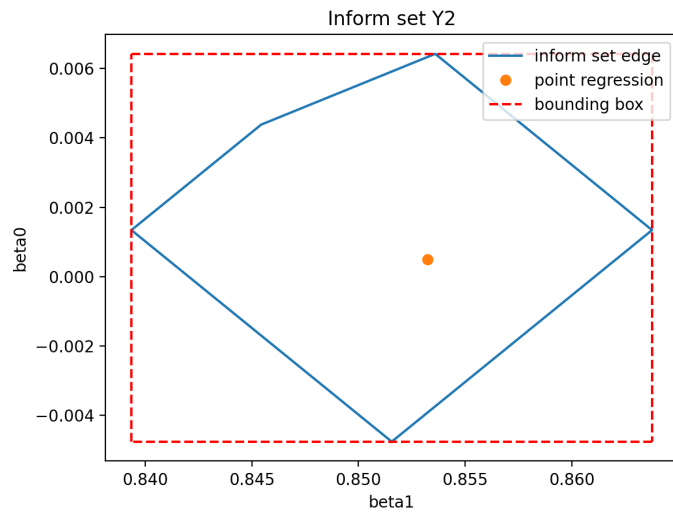


Рис. 7: Информационное множество для Y_2

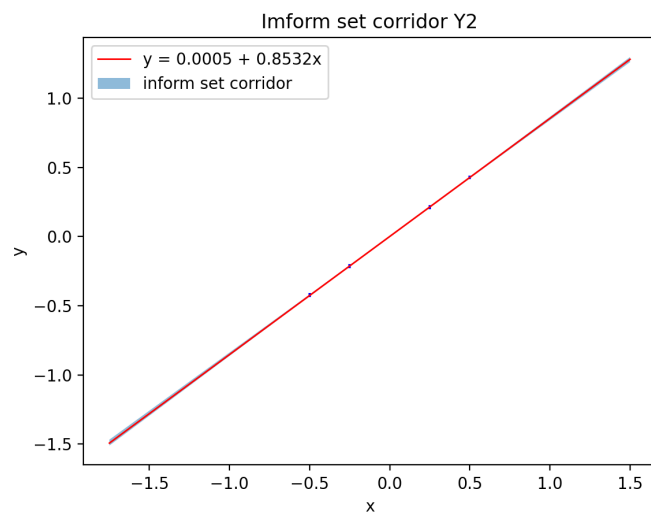


Рис. 8: Коридор совместных значений для Y_2

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.

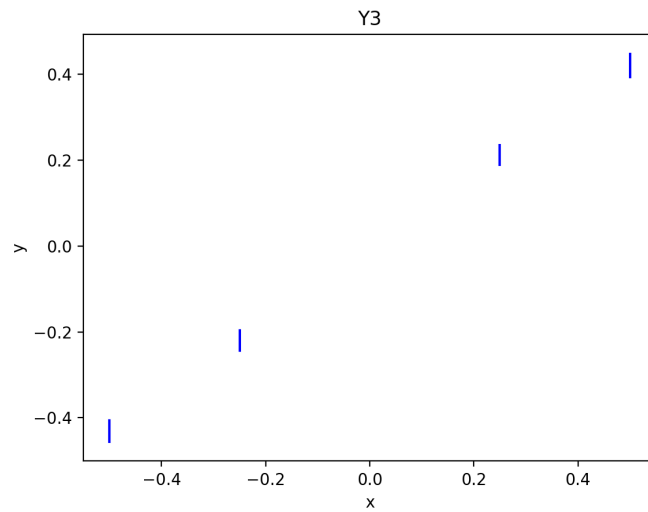


Рис. 9: Третья выборка, Y_3

Построим точечную регрессию.

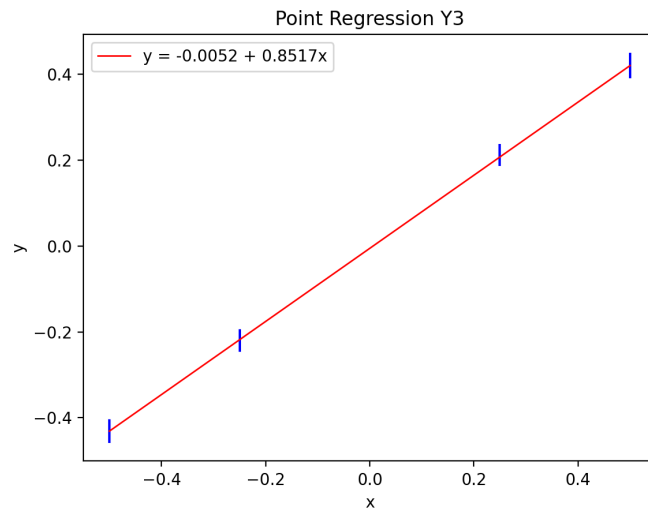


Рис. 10: Точечная линейная регрессия для Y_3

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.

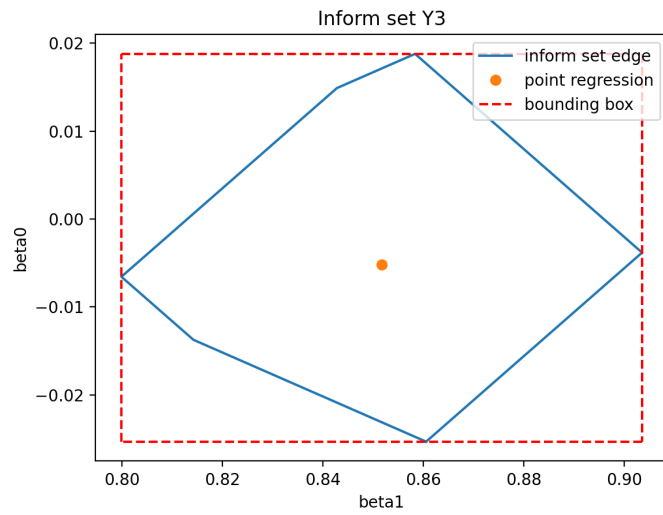


Рис. 11: Информационное множество для Y_3

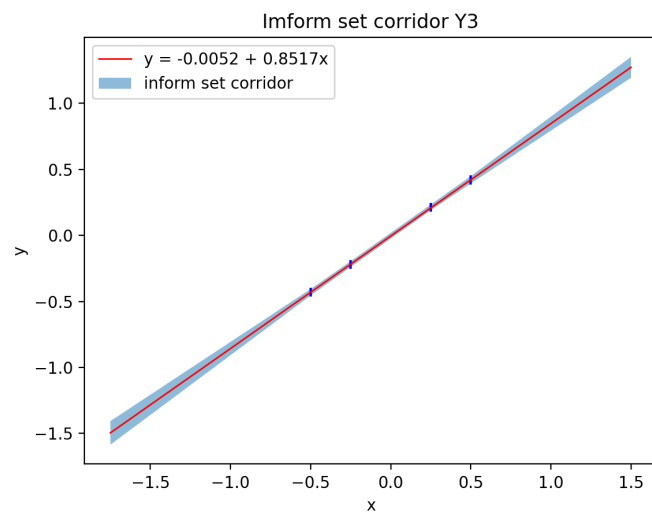


Рис. 12: Коридор совместных значений для Y_3

5 Заключение

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

0	β_0	β_1
Y_1	0.00076	0.86426
Y_2	0.0005	0.85324
Y_3	-0.0052	0.85169

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 .

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