Санкт-Петербургский политехнический университет им. Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

КУРСОВАЯ РАБОТА

Восстановление зависимостей

по дисциплине «Стохастические модели и анализ данных»

Выполнила

студент гр. 5040102/00201 А.Г. Жаворонкова

Преподаватель

к.ф.-м.н., доцент ВШПМиВФ ФМИ А.Н. Баженов

Санкт-Петербург
2022 год

Содержание

Список иллюстраций

Список таблиц

1 Постановка задачи

Необходимо выбрать массив данных и восстановить линейную зависимость с учётом интервальной неопределённости данных. Модель данных будем искать в классе линейных функций

$$y = \beta_1 + \beta_2 x, \quad \beta_2 > 0. \tag{1.1}$$

На рисунке 1 показан график исходных данных.

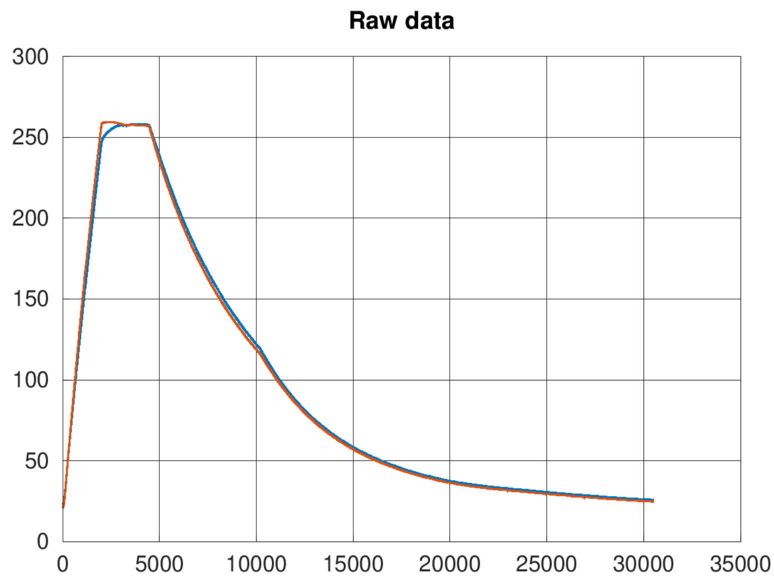

Рис. 1: Исходные данные

2 Исследование

2.1 Выбор рассматриваемой области

Выберем хорошо представимый линейной моделью участок $x \in [500, 1000]$, график этого участка изображён на рисунке 2.

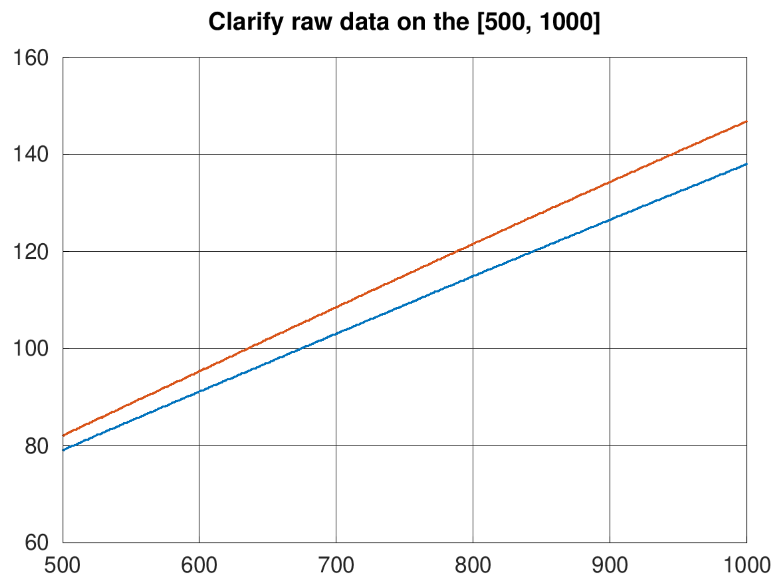

Рис. 2: Уточнённый рассматриваемый участок

Оставим только синюю линию и выберем на ней 5 точек (рисунок 3).

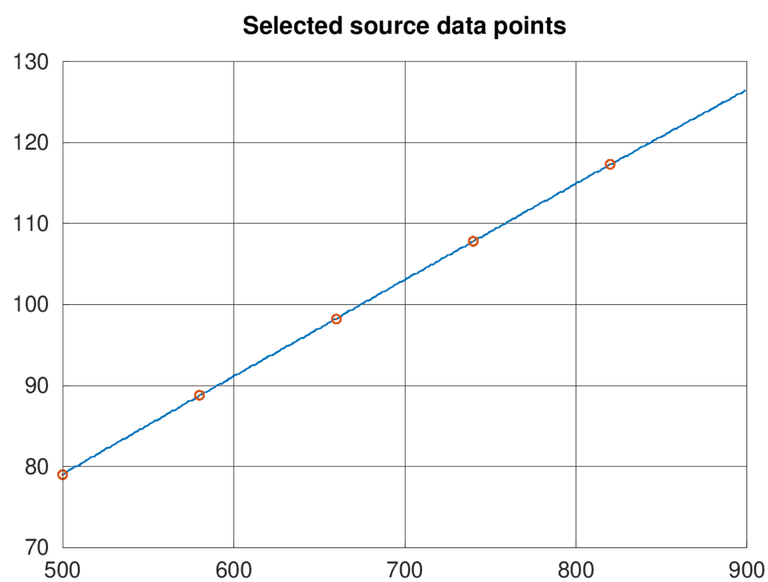

Рис. 3: Выбранные точки из исходных данных

Посмотрим на выбранные значения:

	1	**2**	**3**	**4**	**5**
$\boldsymbol{x}$	500	580	660	740	820
$\boldsymbol{y}$	79.0	88.8	98.2	107.8	117.3

Таблица 1: Значения исследуемых точек

В качестве начальной погрешности зададим $\varepsilon = 0.1$. Погрешность будет одинаковая для всех наблюдений. Этот выбор связан с последним значащим разрядом в данных (рисунок 4).

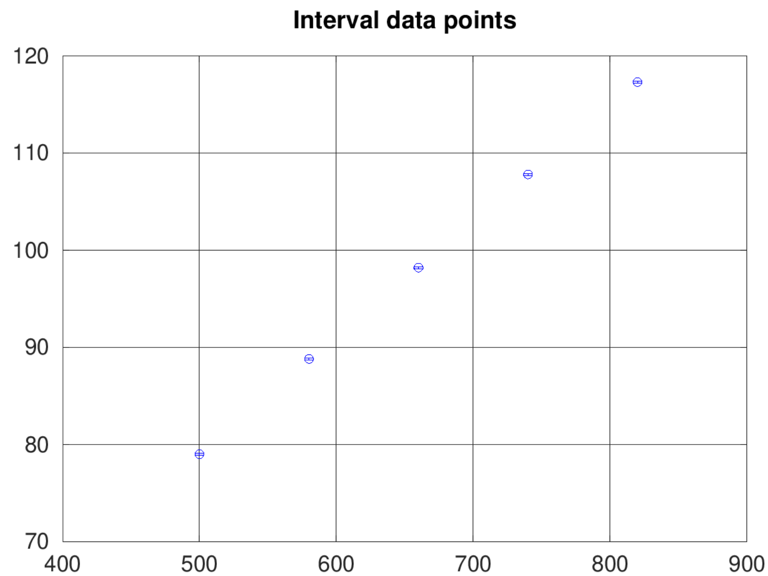

Рис. 4: Входные данные с интервальной неопределённостью

2.2 Параметры модели

Для начала построим линейную модель методом МНК как на точечных значениях. Построенная модель изображена на рисунке 5.

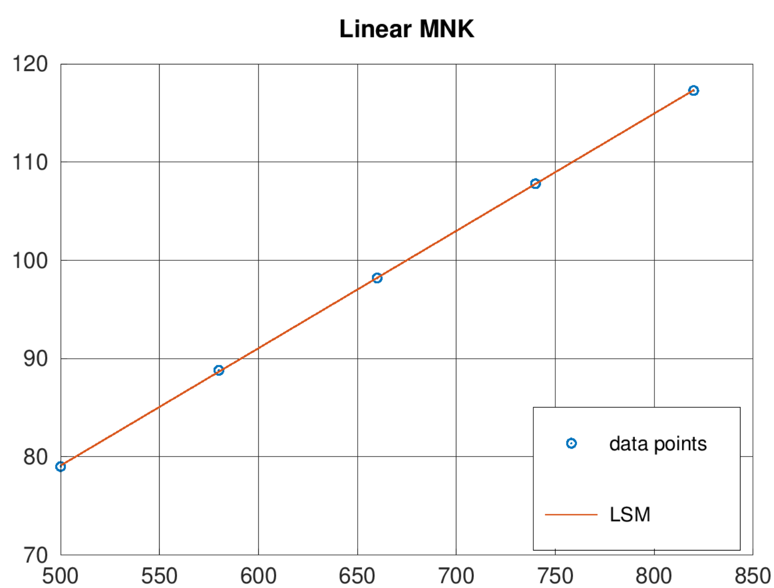

Рис. 5: МНК линейная регрессия

В результате получены значения $\beta_1 = 19.35$ и $\beta_2 = 0.12$. Таким образом, по результатам построения линейной модели методом МНК имеем следующий вид:

$$y = 19.35 + 0.12x.$$

Перейдём к интервальному случаю. При попытке определить информационное множество мы обнаруживаем, что оно пусто. Предположим, что мы недооценили погрешность. Для согласования с данными поставим задачу оптимизации и решим её методом линейного программирования:

$$\begin{aligned} \text{mid } \boldsymbol{y}_i - w_i \cdot \text{rad } \boldsymbol{y}_i \leq & X\beta \leq \text{mid } \boldsymbol{y}_i + w_i \cdot \text{rad } \boldsymbol{y}_i, \quad i = \overline{1, m}, \\ & \sum_{i=1}^{m} w_i \to \min, \\ & w_i \geq 0, \quad i = \overline{1, m}, \\ & w, \beta - ?, \end{aligned} \tag{2.1}$$

где X – матрица $m \times 2$, в первом столбце которой элементы равны 1, во втором – значения x_i. В качестве значений $\text{mid } \boldsymbol{y}_i = y_i,\ \text{rad } \boldsymbol{y}_i = \varepsilon_i$.
По результатам решения задачи оптимизации получаем следующие значения:

$$w = [1.0,\ 1.25,\ 1.0,\ 1.0,\ 1.0]$$

$$\beta = [19.25,\ 0.12]$$

Как видим, требуются небольшие корректировки погрешности, потому не будем считать второе наблюдение выбросом. Затем увеличим погрешность всех измерений:

$$\text{rad } \boldsymbol{y}_i = \max_i w_i \cdot \varepsilon.$$

Построим новое информационное множество параметров модели. Информационное множество задачи построения линейной зависимости по интервальным данным задаётся системой линейных неравенств. Данное множество представляет собой выпуклый многогранник.
Нам понадобятся две точечные оценки:

- Центр наибольшей диагонали информационного множества:

$$\hat{\beta}_{\text{maxdiag}} = \frac{1}{2}(b_1 - b_2),$$

  где b_1 и b_2 – наиболее удалённые друг от друга вершины многогранника;

- Центр тяжести информационного множества:

$$\hat{\beta}_{\text{gravity}} = \frac{1}{n}\sum_{i=1}^{n} b_i,$$

  где b_i – вершины многогранника, а n – их количество.

Построим график информационного множества нашей задачи и нанесём на него точечные оценки (рисунок 6.

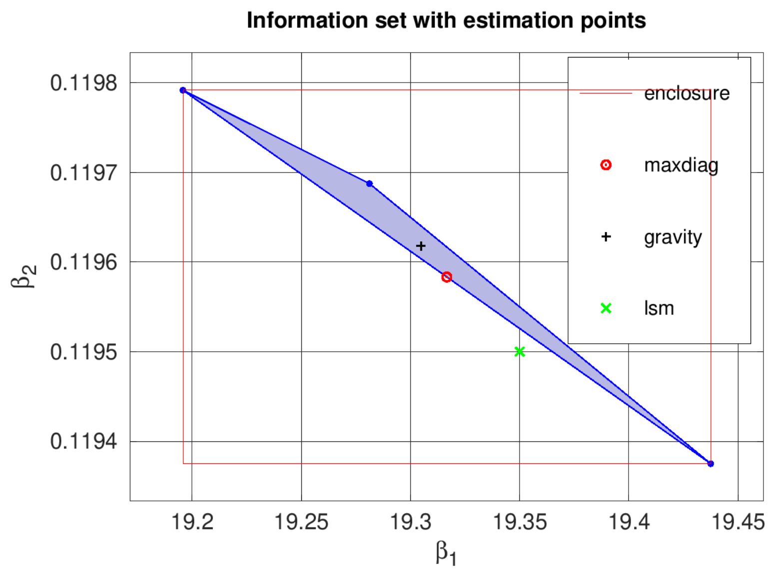

Рис. 6: Информационное множество линейной модели с точечными оценками

Заметим, что значения, полученные при помощи МНК оказались за границами информационного множества.
По результатам построения мы получили следующие внешние интервальные оценки параметров модели

$$\beta_1 = [19.1958,\ 19.4375], \quad \beta_2 = [0.1194,\ 0.1198].$$

2.3 Коридор совместных зависимостей

Построим коридор совместных зависимостей всего рассматриваемого диапазона (рисунок 7). Видим, что он сливается в одну прямую.

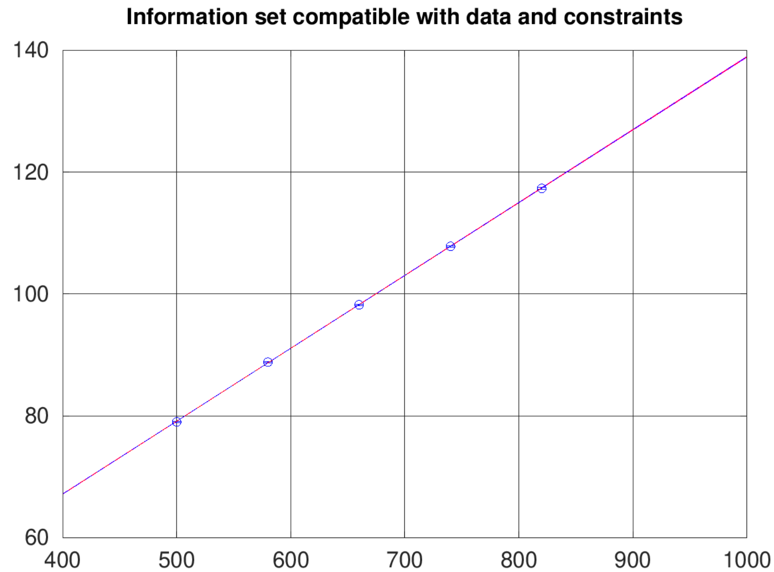

Рис. 7: Коридор совместных зависимостей, весь диапазон

Рассмотрим более подробно, что происходит возле какой-нибудь одной точки, например, первой, на рисунке 8.

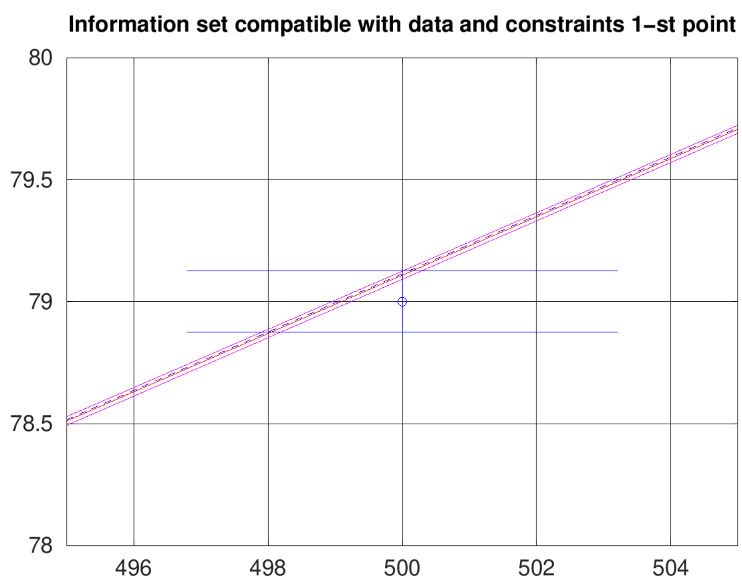

Рис. 8: Коридор совместных зависимостей, первая точка

2.4 Прогноз за пределы интервала

С помощью построенной выше модели и найденных внешних интервальных оценок параметра имеем следующую модель:

$$\hat{\boldsymbol{y}}(x) = [19.1958,\ 19.4375] + [0.1194,\ 0.1198]x.$$

На основании этой модели получим прогнозируемые значения выходной переменной. Пусть

$$x_p = [250,\ 450,\ 600,\ 950,\ 1800],$$

тогда $y_p = \hat{\boldsymbol{y}}(x_p)$. Посмотрим на получившиеся значения в таблице ниже:

x_p	y_p	rad y_p
250	[43.15, 43.31]	0.08
450	[73.10, 73.16]	0.03
600	[91.06, 91.09]	0.02
950	[132.84, 133.00]	0.08
1800	[234.31, 234.82]	0.25

Таблица 2: Прогнозы за пределы интервала

Неопределённость прогноза растёт по мере удаления от области, в которой производились исходные измерения. Это обусловлено видом коридора зависимости, расширяющимся за пределами области измерений.

2.5 Граничные точки множества совместности

Для нашей задачи граничными оказались точки под номерами 1, 2, 5. Убедимся в этом, посмотрев детально каждую из точек подробнее. Из рисунков ниже можем сделать вывод, что точки 1, 2, 5 действительно являются граничными. А именно точки 1 и 5 касаются верхней границы множества, а точка 2 - нижней.

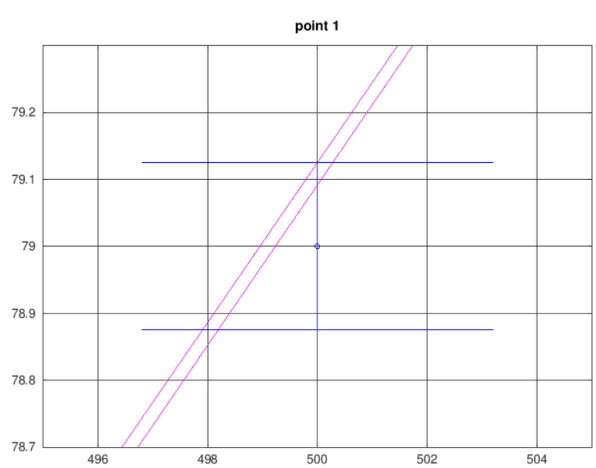

Рис. 9: Точка 1.

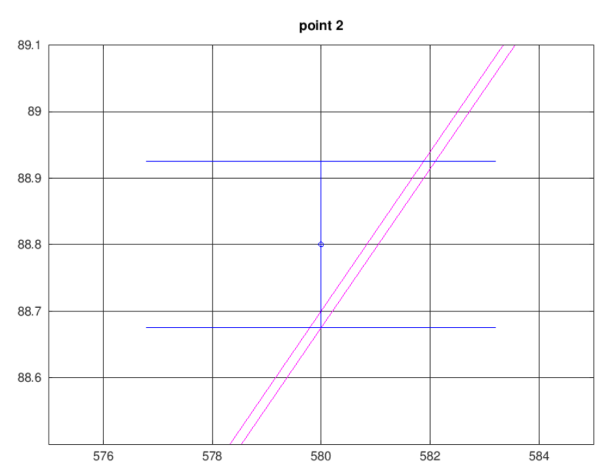

Рис. 10: Точка 2.

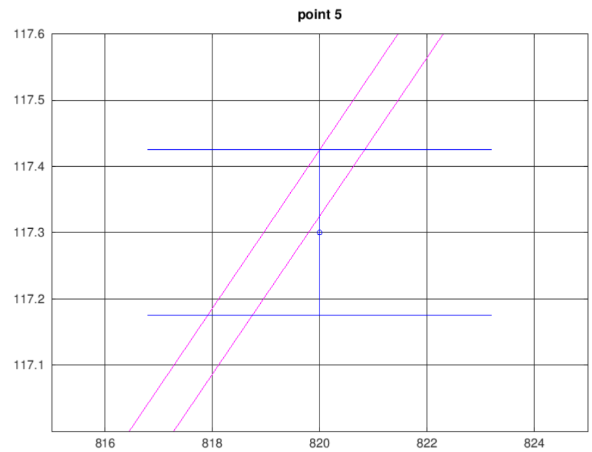

Рис. 11: Точка 5.

Исходя из рисунков ниже, убеждаемся, что точки 3 и 4 не являются граничными.

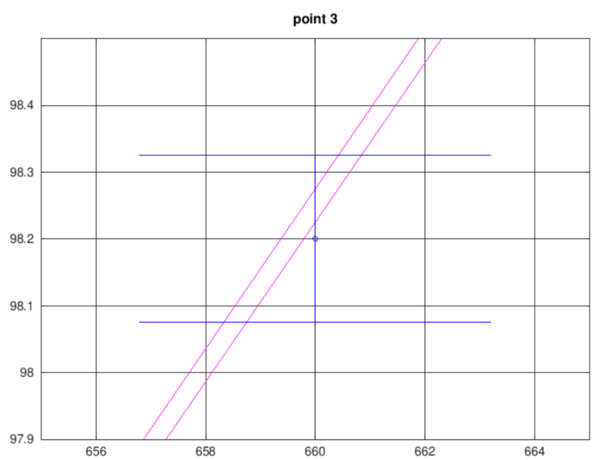

Рис. 12: Точка 3.

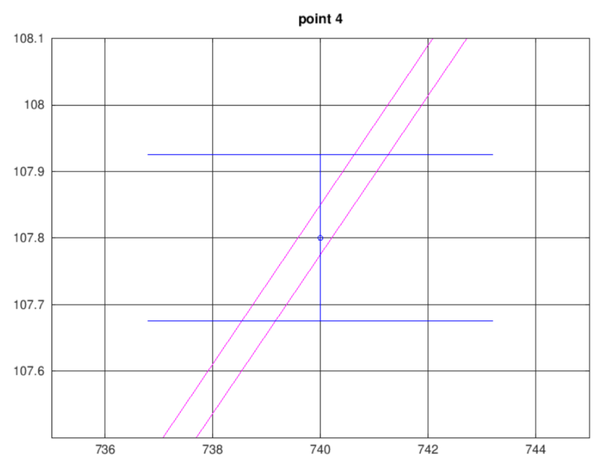

Рис. 13: Точка 4.

Таким образом, можно сделать вывод, что набор точек 1, 2 и 5 может полностью определить модель.

3 Заключение

В ходе работы была построена линейная модель данных. Сначала были рассмотрены точечные наблюдения, а затем – наблюдения с интервальной неопределённостью.
Для заданных наблюдений была выбрана погрешность, но выборка оказалась несовместной. Таким образом, мы сделали вывод, что в выборке отсутствуют выбросы и причина несовместности – недооценённая погрешность.
Чтобы улучшить оценку погрешности, была сформирована и решена задача линейного программирования, после корректировки которой выборка стала совместной. Мы получили информационное множество для параметров линейной модели, построили коридор совместности и обнаружены граничные точки коридора совместности. По полученной модели были вычислены прогнозы за пределами области измерений.

Список литературы

[1] А.Н. Баженов, С.И.Жилин, С.И. Кумков, С.П.Шарый. Обработка и анализ данных с интервальной неопределённостью. РХД. Серия «Интервальный анализ и его приложени». Ижевск. 2021.с.200 (20.02.2022).

[2] Жилин С.И. Примеры анализа интервальных данных в Octave [Электронный ресурс] / Режим доступа: https://github.com/szhilin/octave-interval-examples (20.02.2022).

A Приложение

Ссылка на проект с кодом исследования и отчётом:
https://github.com/Zhavoronkova-Alina/Stochastic-models-and-data-analysis