Стохастические модели и анализ данных
Работа по восстановлению зависимости

Никита Лансков

29 декабря 2021 г.

Содержание

1 Постановка задачи

Требуется выбрать массив данных с интервальной неопределенностью и восстановить линейную зависимость.

Модель данных будем искать в классе линейных функций

$$y = \beta_1 + \beta_2 x \tag{1}$$

При условии: $\beta_2 > 0$

2 Параметры модели

2.1 Предобработка данных

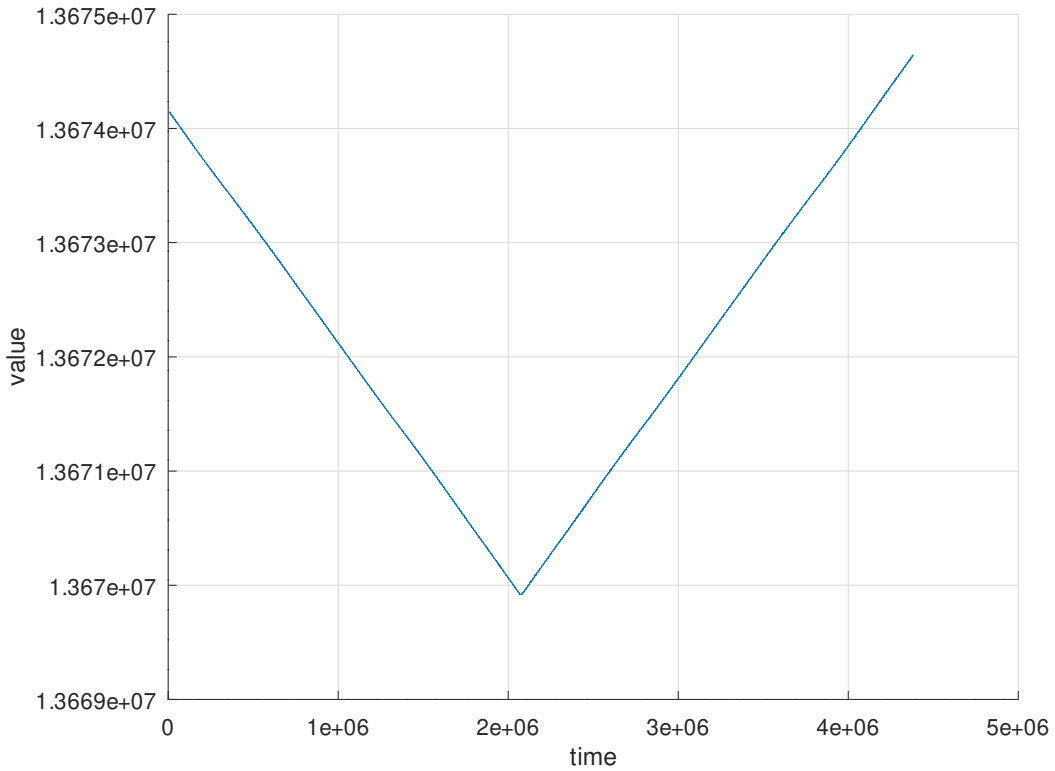


Рис. 1: Все данные

Выберем область, которую будем рассматривать, например: $t \in [2.5e^6, 4.3e^6]$

Даллее объединим точки, значения в которых отличаются менее чем на 20.

Ну и в качестве значений нашей небольшой подвыборки возьмем первые 10 точек на нечетных позициях.

в итоге получим данные следующего вида.

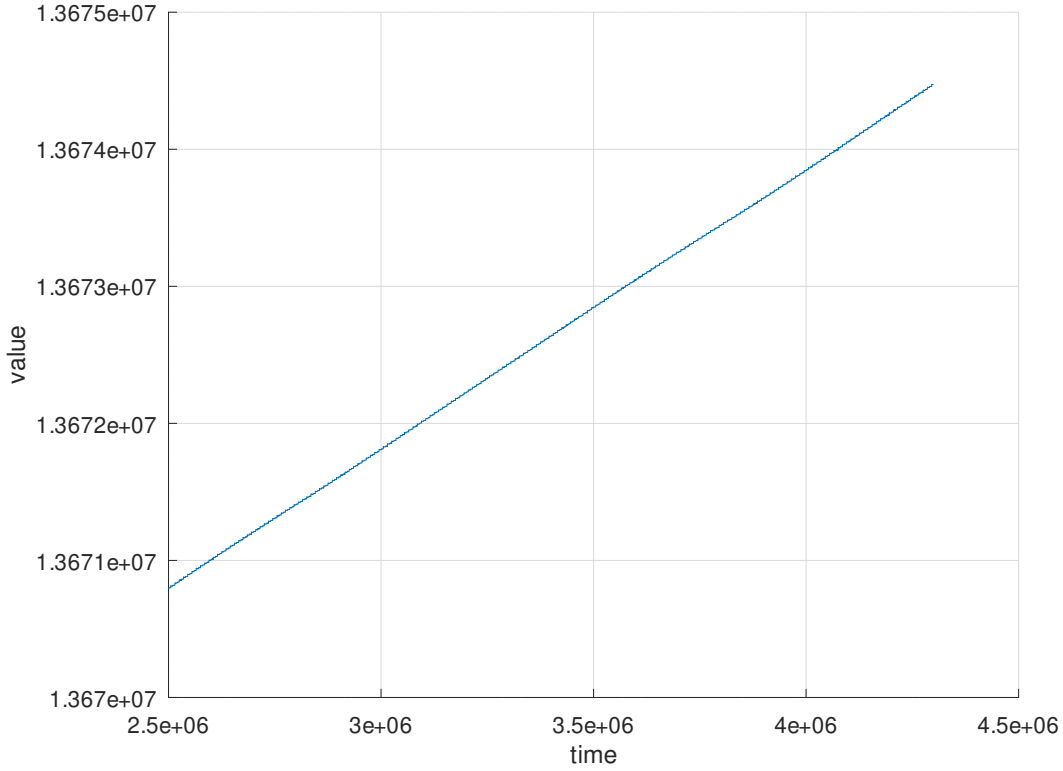


Рис. 2: Выбранный диапазон значени

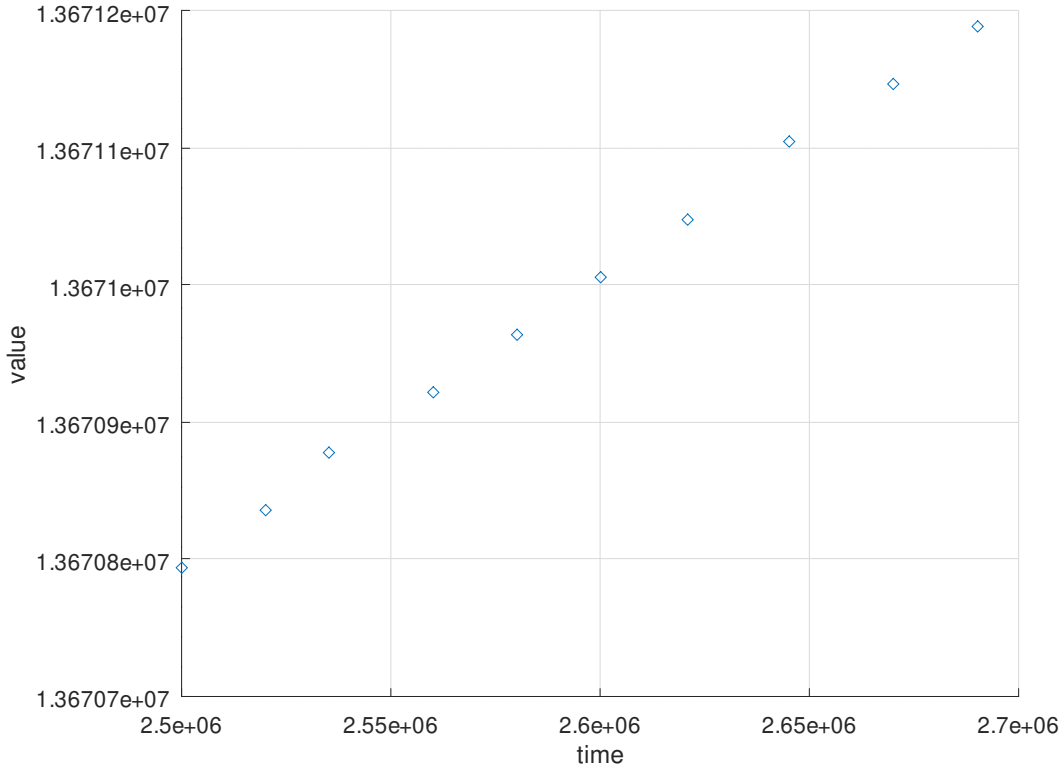


Рис. 3: Итоговая выборка из 10 значений

2.2 Линейная модель МНК для точечных значений

В качестве начальной погрешности возьмем $\varepsilon = 1$, в соответствии с последним значащим разрядом в данных.

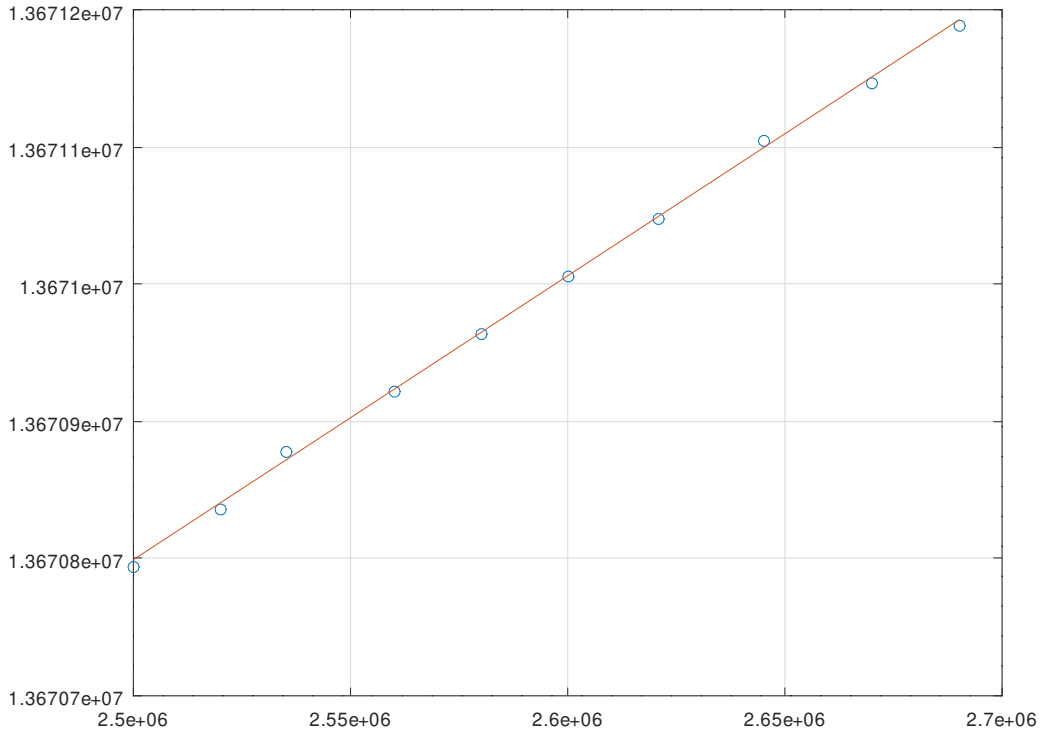


Рис. 4: МНК

Для точечного случая по методу наименьших квадратов получаем следующие значения параметров:

$$\beta_1 = 1.3666e^7, \beta_2 = 2.0722e^{-3}$$

2.3 Модель для интервального случая

При переходе к интревальному случаю, обнаруживаем, что информационное множество оказывается пустым. Попробуем это исправить, решив задачу оптимизации для уточнения погрешности [1].

$$\begin{cases} mid\ y_i - w_i\ \cdot rad\ y_i \leqslant X\beta \leqslant mid\ y_i + w_i\ \cdot rad\ y_i, i = \overline{1,m} \\ \sum_{i=1}^{m} w_i \to \min \\ w_i \geqslant 0, i = \overline{1,m} \\ w, \beta = ? \end{cases} \tag{2}$$

Где X - матрица $m \times 2$, в первом столбце единичные значения, а во втором значения x_i.

$$mid\ y_i = y_1,\ rad\ y_i = \varepsilon_i$$

Полученные значения в задаче оптимизации:

$$w = [3.77, 3.12, 7.71, 1.00, 1.00, 1.78, 1.00, 7.87, 1.39, 1.00]$$

$$\beta = [1.3666e^{7}, 2.0710e^{-3}]$$

Увеличим погрешность всех измерений:

$$rad\ y_i = \max_i\ w_i \cdot \varepsilon$$

Построим новое информационное множество параметров модели. Поскольку информационное множество задачи построения линейной зависимости по интервальным данным задаётся системой линейных неравенств, то оно представляет собой выпуклый многогранник. [2] Дополнительно обозначим на графитке центр наибольшей диагонали информационного множества и его центр тяжести.

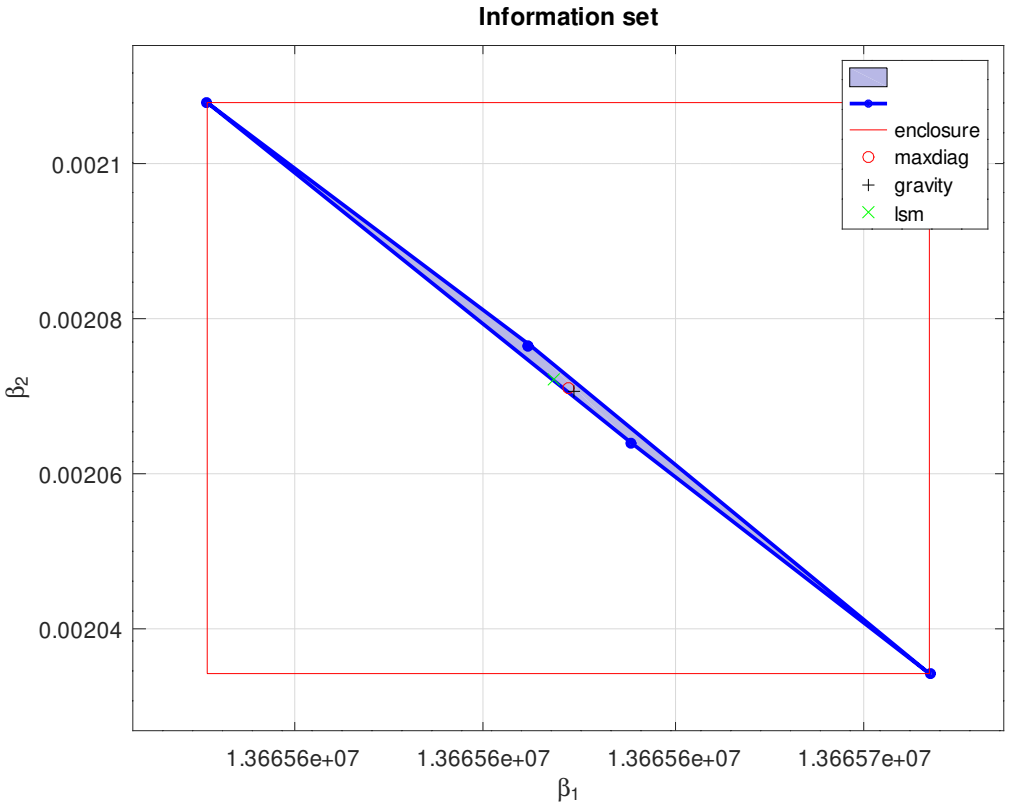


Рис. 5: Информационное множество

3 Коридор совместных зависимостей

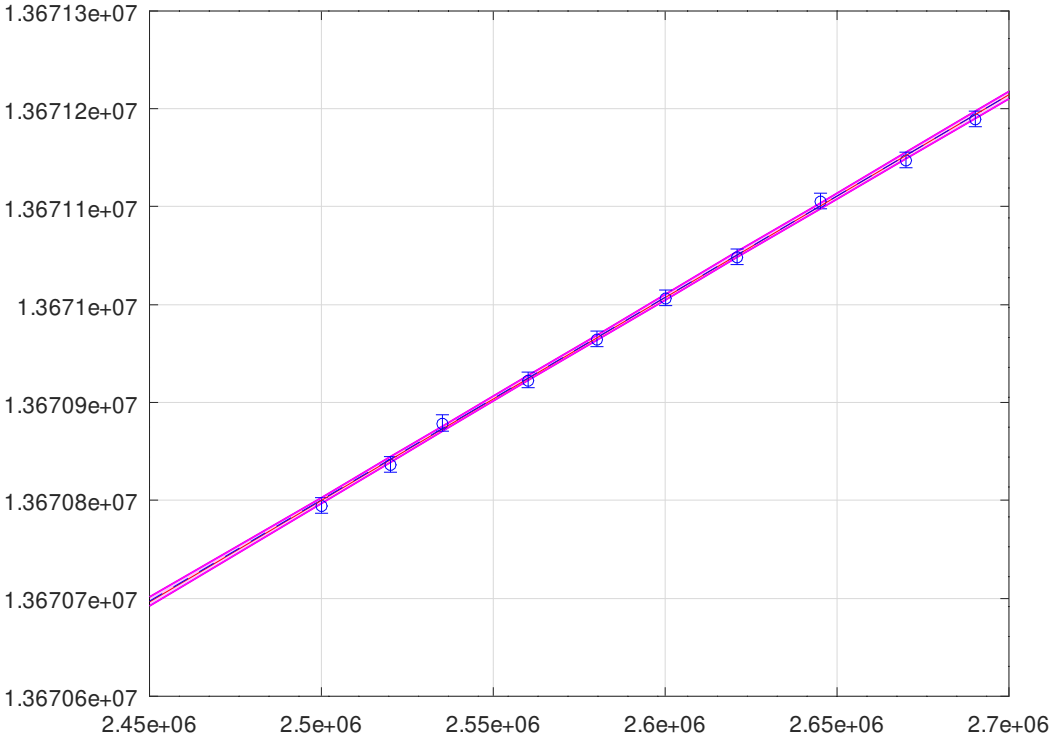


Рис. 6: Коридор совместных зависимостей

4 Прогноз за пределы интервала

При помощи построенной модели:

$$\hat{y}(x) = [1.3666e^7, 1.3666e^7] + [2.0342e^{-3}, 2.1079e^{-3}]x$$

Спрогнозируем значения для $x_p = [2.75e^6; 2.78e^6; 2.81e^6; 2.86e^6; 2.95e^6]$

x_p	y_p	$rad\ y_p$
$2.75e^6$	$[1.36713e^7, 1.36713e^7]$	6.2748
$2.78e^6$	$[1.36714e^7, 1.36714e^7]$	7.3793
$2.81e^6$	$[1.36714e^7, 1.36715e^7]$	8.4839
$2.86e^6$	$[1.36715e^7, 1.36716e^7]$	10.3249
$2.95e^6$	$[1.36717e^7, 1.36717e^7]$	13.6386

Таблица 1: Прогноз значений

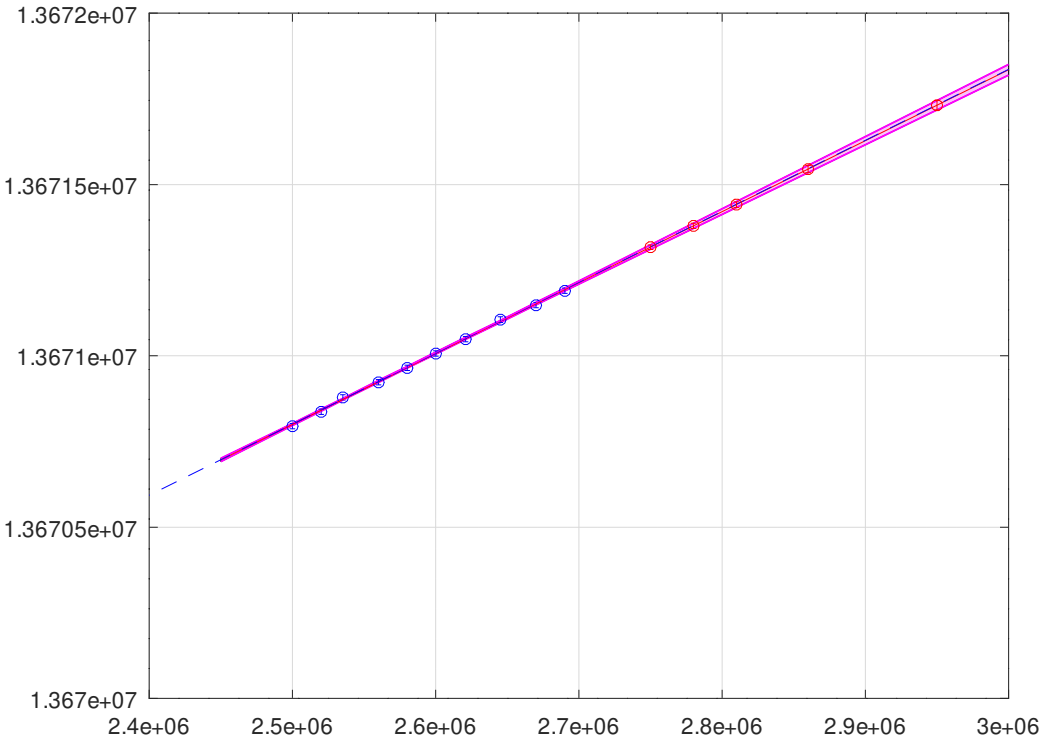
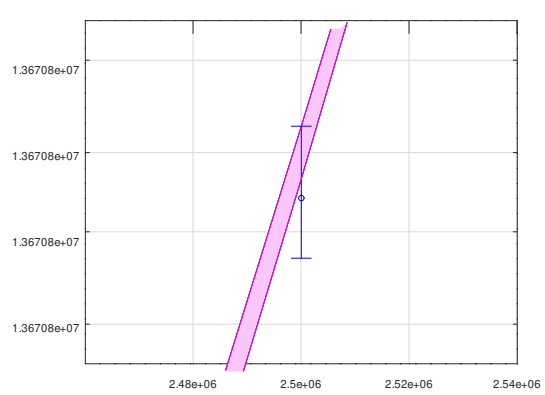


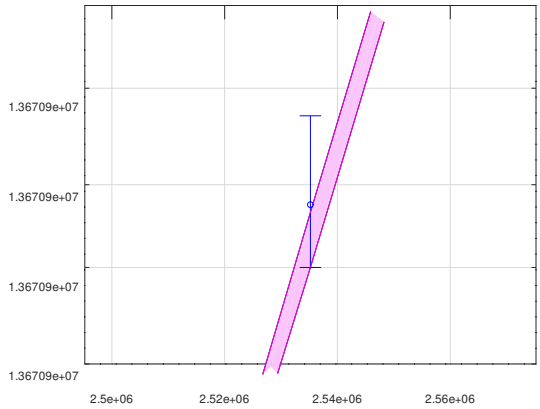
Рис. 7: Прогноз за пределы интрервала

5 Граничные точки множества совместности

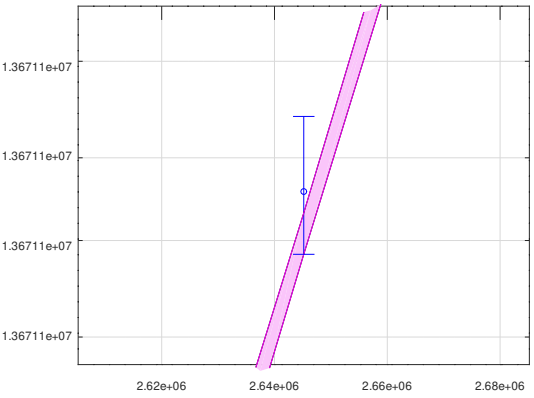
Граничными оказались точки с номерами 1, 3, 8, 9, 10



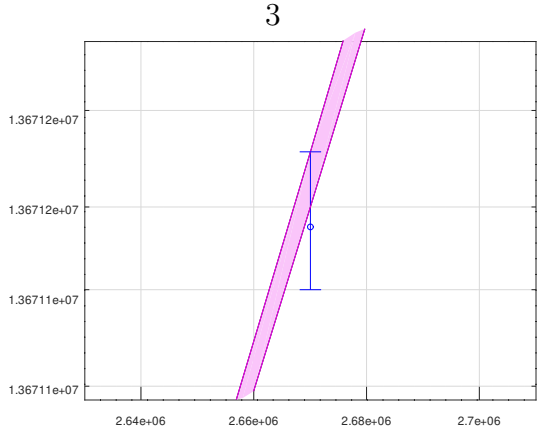
1



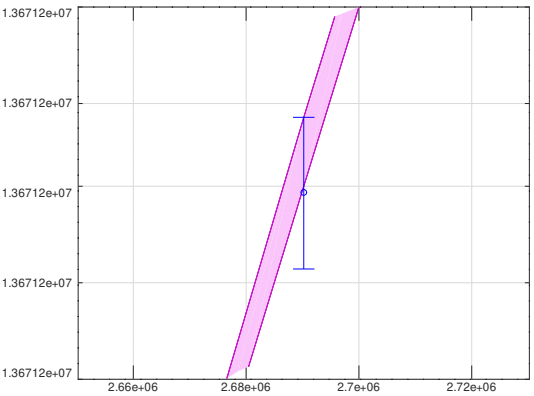
3



8



9



10

Рис. 8: Граничные точки

Остальные точки не являются граничными:

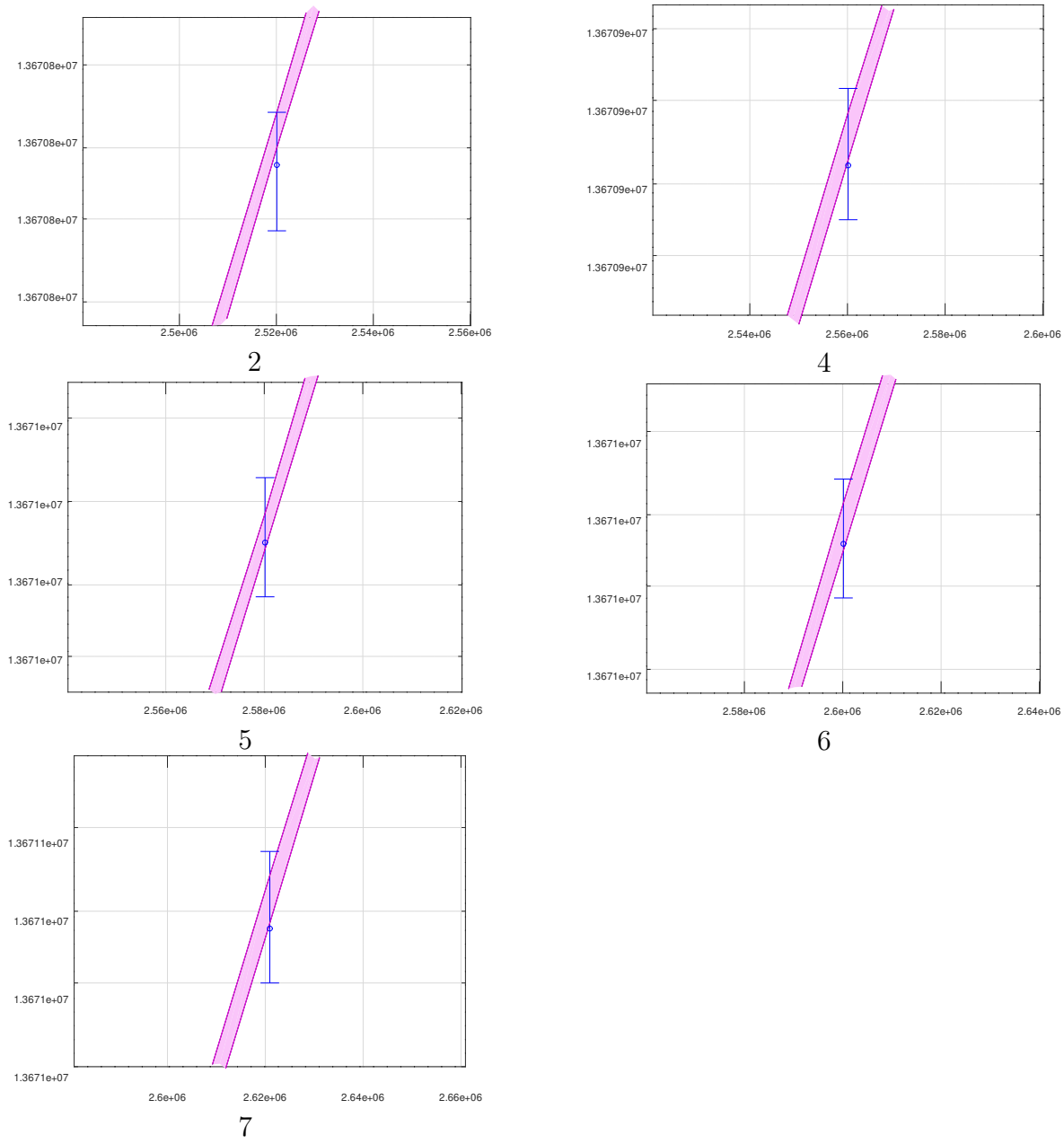


Рис. 9: Не граничные точки

Точка под номером два подозрительно похожа на граничную, но на практике у нее слишком большое отклонение от границы, поэтому она была отфильтрована из граничных.

6 Заключение

В ходе работы была построена линейная модель данных. Наблюдения рассматривались сначала как просто точечные, далее – как значения с интервальной неопределённостью.

Была задана погрешность наблюдений, однако выборка оказалась несовместной. Было принято решение, что в выборке отсутствуют выбросы и причина несовместности – недооценённая погрешность.

Для улучшения оценки погрешности была сформирована и решена задача линейного программирования. После корректировки выборка стала совместной.

Было получено информационное множество для параметров линейной модели, построен коридор совместности и обнаружены граничные точки коридора совместности. По полученной модели были вычислены прогнозы за пределами области измерений.

Все материалы по данной работе доступны по ссылке: [3]

Список иллюстраций

Список таблиц

Список литературы

[1] С.И.Кумков С.П.Шарый А.Н.Баженов, С.И.Жилин. Обработка и анализ данных с интервальной неопределённосью. "РХД. Серия «Интервальный анализ и его приложение». Ижевск. 2021.

[2] С.И. Жилин. Примеры анализа интервальных данных в octave. https://github.com/szhilin/octave-interval-examples.

[3] Материалы работы. https://github.com/LanskovNV/poly-master-3/tree/main/interval_analysis/.