Санкт-Петербургский политехнический университет Петра Великого
Высшая школа прикладной математики и вычислительной физики
Кафедра прикладной математики

Курсовая работа
по дисциплине «Стохастические модели и анализ данных»
на тему

Восстановление зависимостей

Выполнил студент гр. 5040102/00201
Чепулис М.А.

Преподаватель
Баженов А.Н.

Санкт-Петербург
2021 год

Оглавление

Постановка задачи 3
Параметры модели 5
Коридор совместных зависимостей 7
Прогноз за пределы интервала: 8
Граничные точки множества совместности 8
Заключение 9
Приложение: 11
Использованная литература 11

Постановка задачи

Необходимо выбрать массив данных и восстановить линейную зависимость с учётом интервальной неопределённости данных.

Модель данных будем искать в классе линейных функций:

$$y = \beta_1 + \beta_2 x$$

С неотрицательной первой производной: $\beta_2 > 0$

Ниже показан график исходных данных

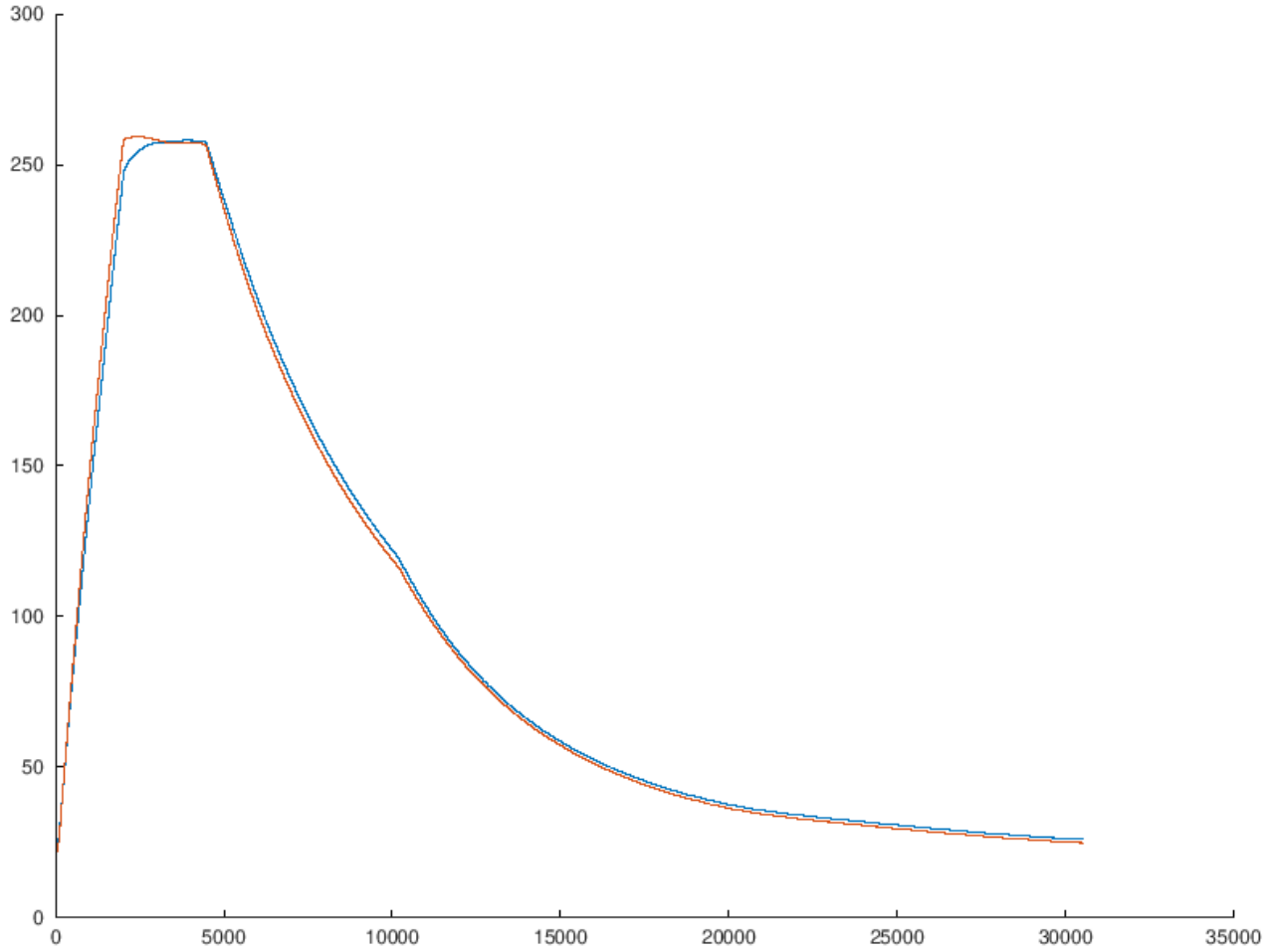


Рисунок 1 Исходные данные

Решение

Выбор рассматриваемой области

Выберем хорошо представимый линейной моделью участок: $x \in [500, 1000]$

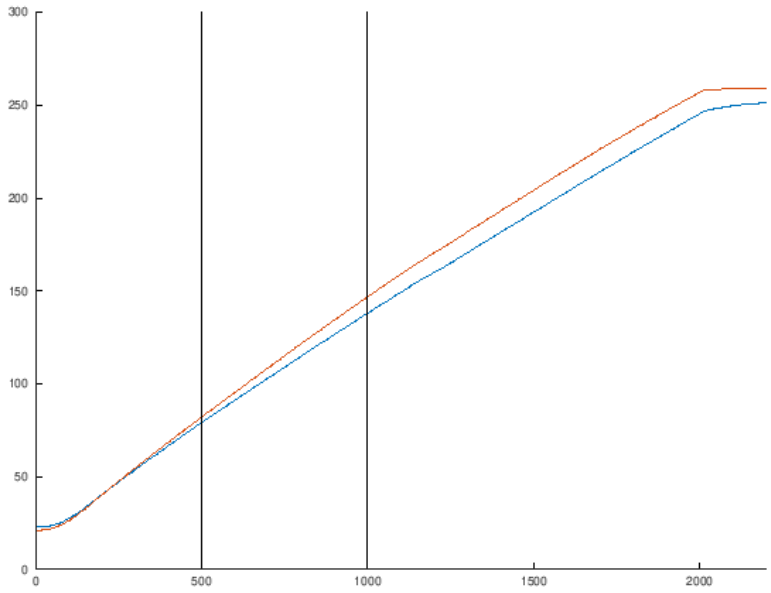


Рисунок 2 Уточнённый рассматриваемый участок

Оставим только нижнюю линию, и выберем на ней 5 точек:

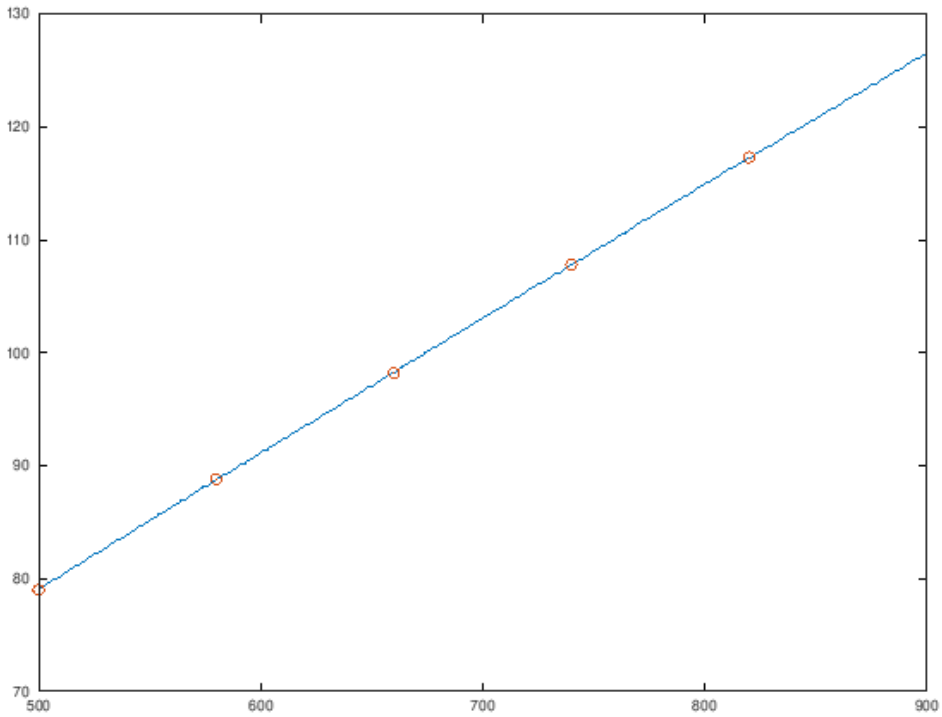


Рисунок 3 Выбранные точки из исходных данных

Посмотрим на выбранные значения:

x = 500 580 660 740 820

y = 79.0 88.8 98.2 107.8 117.3

В качестве начальной погрешности зададим $\varepsilon = 0.1$, одинаковую для всех наблюдений. Этот выбор связан с последним значащим разрядом в данных.

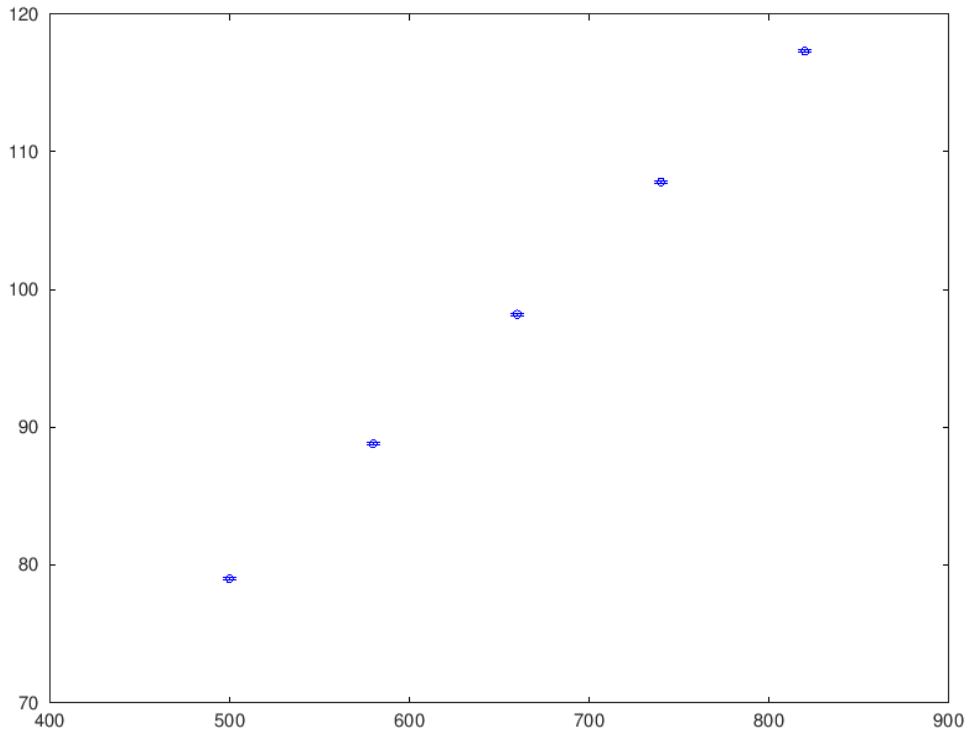


Рисунок 4 Входные данные с интервальной неопределённостью

Параметры модели

Сперва построим линейную модель методом МНК как на точечных значениях:

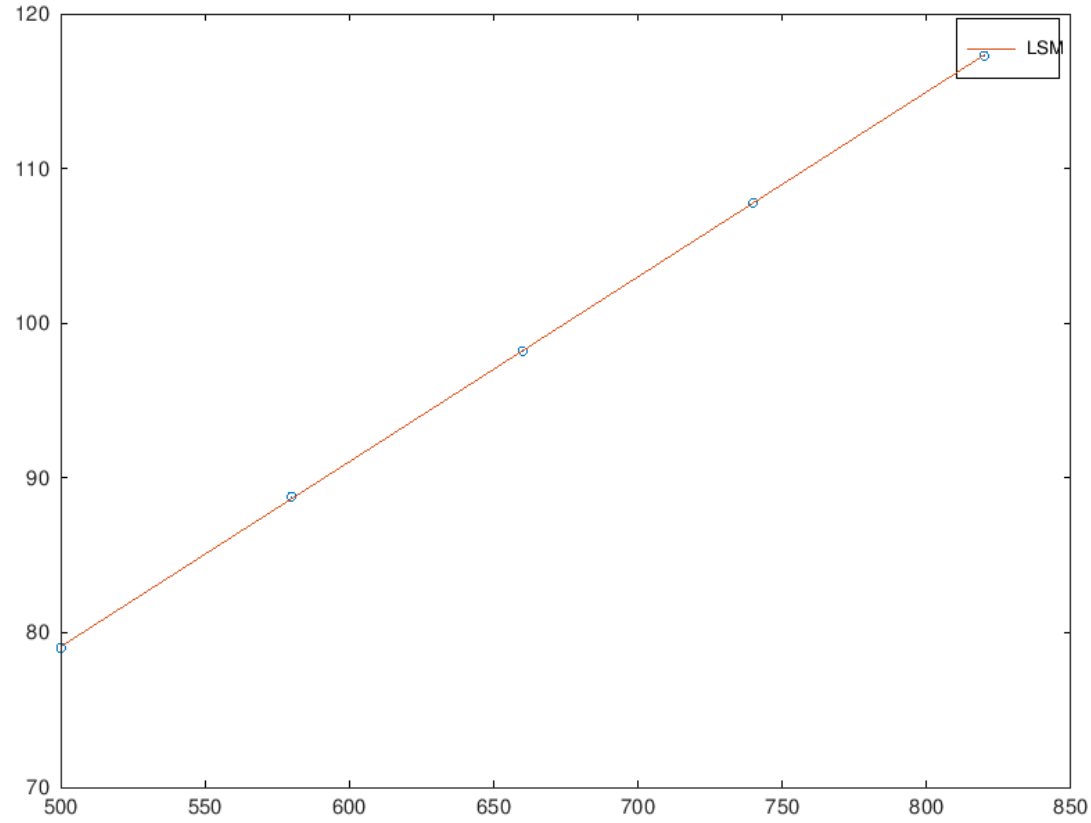


Рисунок 5 МНК линейная регрессия

$$\beta_1 = 19.35, \quad \beta_2 = 0{,}12$$

При переходе к интервальному случаю, при попытке определить информационное множество мы обнаруживаем, что оно пусто. Предположим, что погрешность была недооценена. Для согласования с данными поставим задачу оптимизации и решим её методом линейного программирования [1]:

$$mid\ \boldsymbol{y}_i - w_i \cdot rad\ \boldsymbol{y}_i \le X\beta \le\ mid\ \boldsymbol{y}_i + w_i \cdot rad\ \boldsymbol{y}_i, \qquad i = 1, m,$$

$$\sum_{i=1}^{m} w_i \to min,$$

$$w_i \ge 0, \qquad i = 1, m,$$

$$w, \beta = ?$$

где X – матрица $m \times 2$, в первом столбце которой элементы равные 1, во втором – значения x_i.

В качестве значений $mid\ \boldsymbol{y}_i = y_i, rad\ \boldsymbol{y}_i = \varepsilon_i$

Значение весов в задаче оптимизации:

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

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

Как мы видим, требуются небольшие корректировки погрешности, потому не будем считать второе наблюдение выбросом.

Увеличим погрешность всех измерений:

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

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].

Сразу обозначим на графике несколько точечных оценок:

- Центр наибольшей диагонали информационного множества:

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

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

- Центр тяжести информационного множества:

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

  где b_i – вершина многогранника, n – их количество.

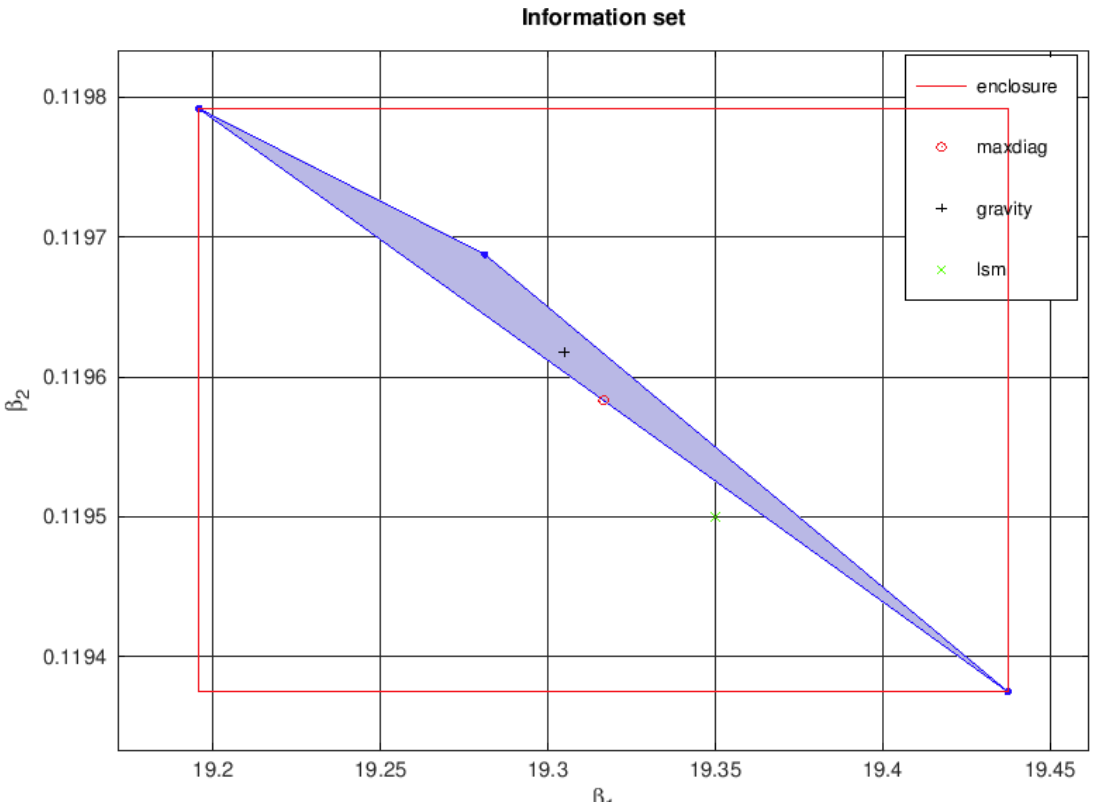


Рисунок 6 информационное множество линейной модели

Заметим, что значения, полученные при помощи МНК оказались за границами информационного множества.

Коридор совместных зависимостей

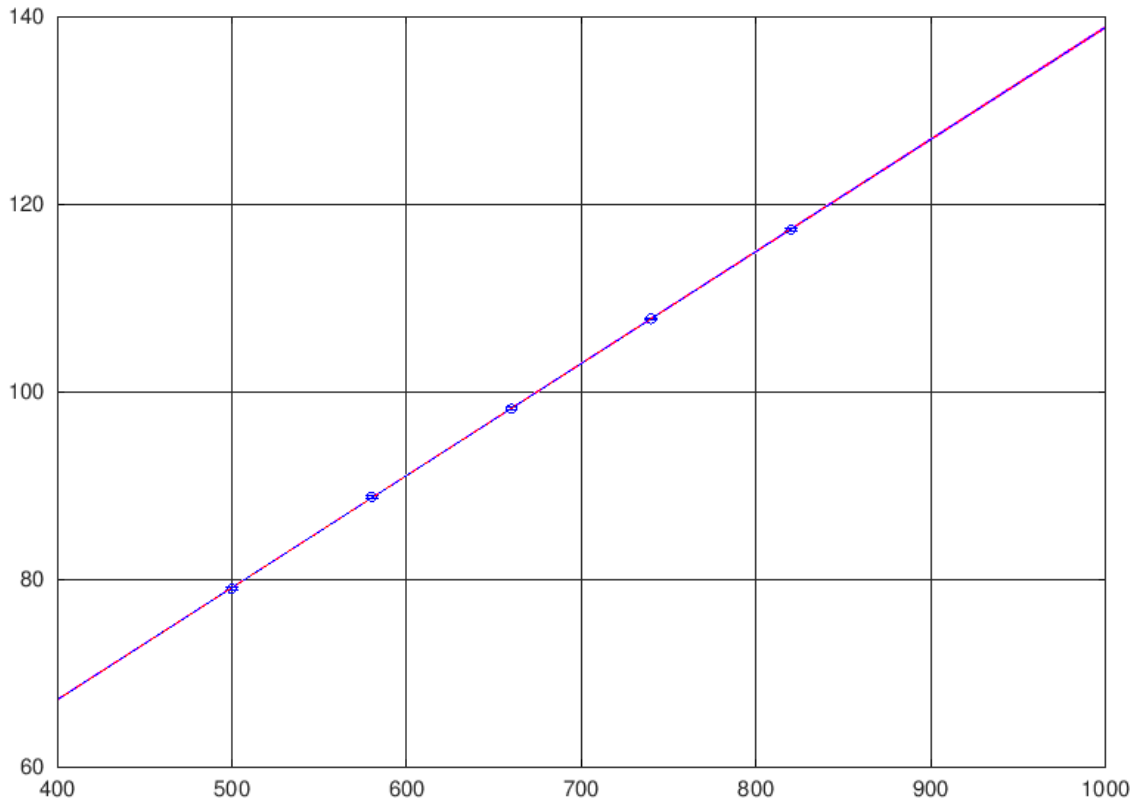


Рисунок 7 Коридор совместных зависимостей, весь диапазон

Однако коридор совместных событий слился в одну прямую. Рассмотрим подробнее, что происходит возле первой точки:

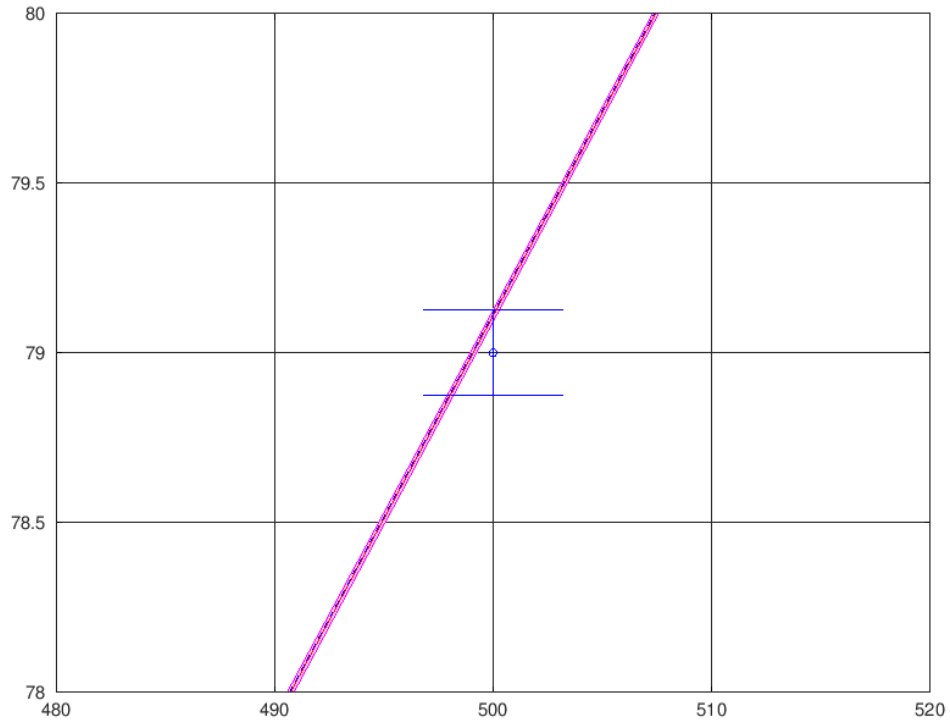


Рисунок 8 Коридор совместных событий в окрестности первого наблюдения

Прогноз за пределы интервала:

С помощью построенной выше модели

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

Можно получить прогнозные значения выходной переменной:

Возьмём 3 точки:

$$x_p = [450, 600, 950]$$

Тогда $y_p = \hat{\boldsymbol{y}}(x_p)$

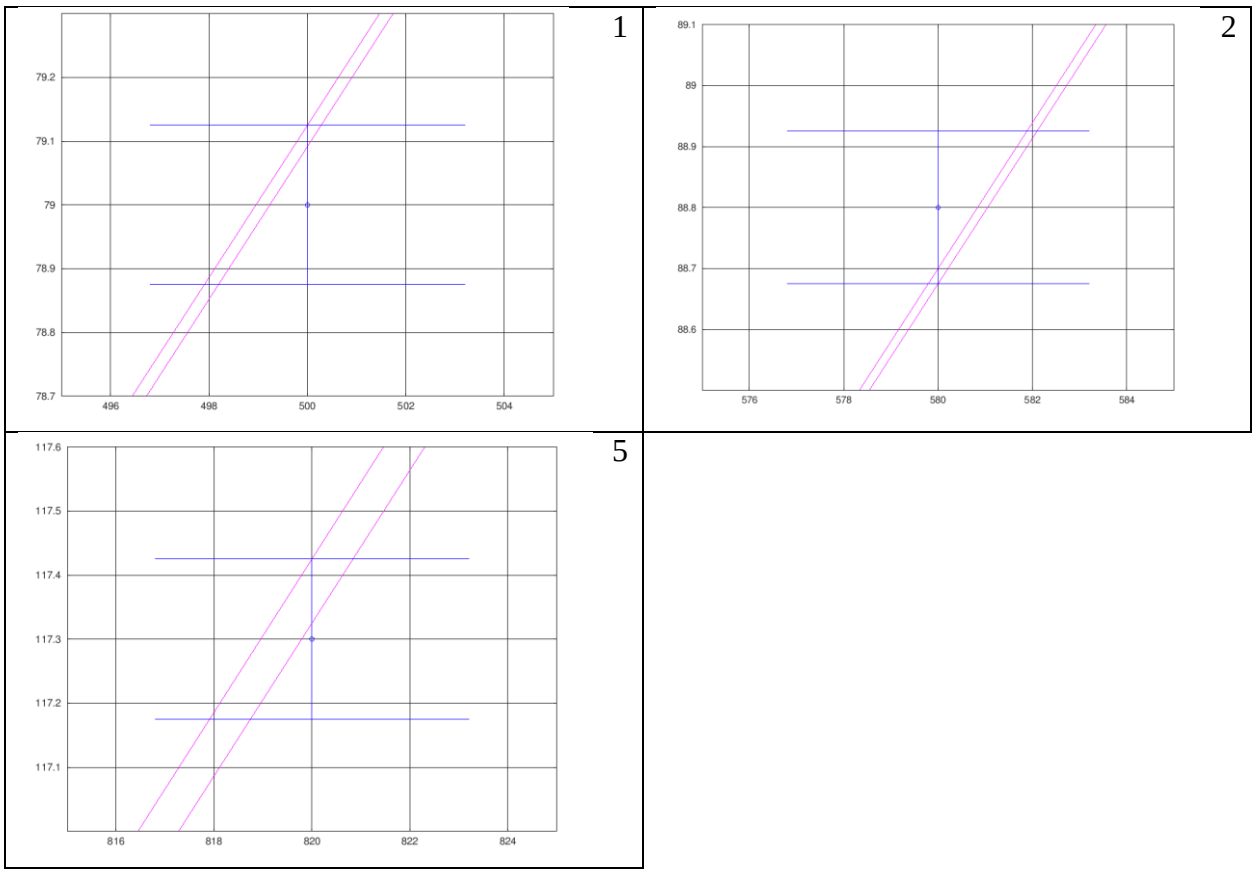
x_p	y_p	$rad\ y_p$
450	[73.102, 73.156]	0.02
600	[91.063, 91.094]	0.01
950	[132.844, 132.998]	0.07

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

Граничные точки множества совместности

В данном случае граничными оказались точки с номерами 1, 2, 5.

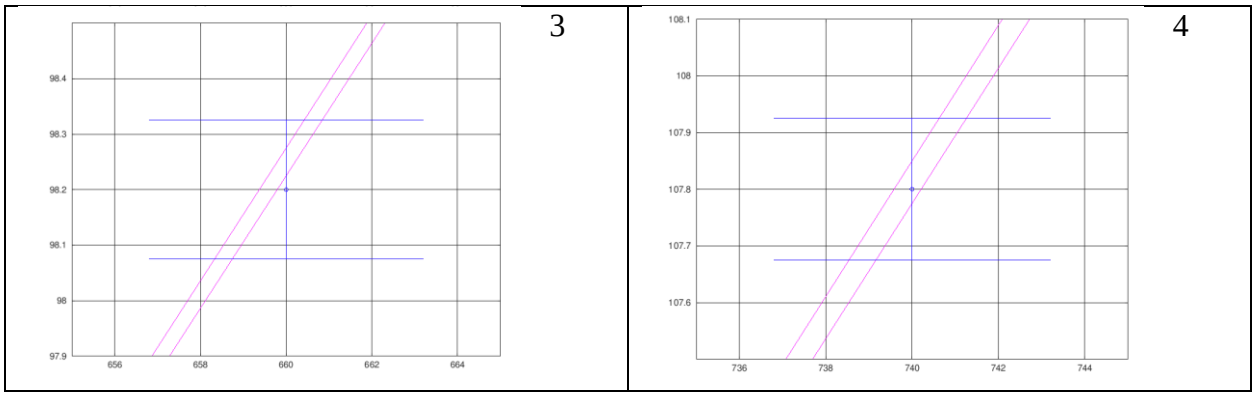
Убедимся в этом посмотрев детально на каждую из точек подробнее



Как мы видим, точки 1 и 5 касаются верхней границы множества.

Точка 2 – нижней.

Убедимся, в том, что остальные точки (3 и 4) не являются граничными.



Тем самым набор точек [1, 2, 5] может полностью определить модель.

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.

Приложение:

Ссылка на проект с кодом реализации:
https://github.com/MChepulis/Lab-Stochastic-models-and-data-analysis

Использованная литература

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 приложение». Ижевск. 2021. с.200.
2. С.И.Жилин. Примеры анализа интервальных данных в Octave https://github.com/szhilin/octave-interval-examples