Санкт-Петербургский политехнический университет имени Петра Великого

Институт прикладной математики и механики

Высшая школа прикладной математики и физики

Задача восстановления зависимостей
Отчёт по лабораторной работе №2

Выполнил:

Студент: Ли Жуйци
Группа: 5040102/10201

Принял:

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

2022 г.

Содержание

Список иллюстраций

1. Постановка задача

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

Модель данных будем искать в классе линейных функций:

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

С неотрицательной первой производной: $\beta_2 > 0$

2. Теория

2.1. Информационное множество

Интервальное множество решений , которое необходимо построить и оценить в задании 1, называется информационным множеством. В качестве точечных оценок информационного множества будут использованы следующие величины:

• Середина наибольшей диагонали

• Центр тяжести (среднее суммы всех вершин)

• Оценка , полученная решением исходной задачи в точечной постановке (с серединами интервалов) методом наименьших квадратов

2.2. Коридор совместных зависимостей

Коридором совместных зависимостей называется множество, образованное всеми решениями с параметрами из информационного множества.

2.3. Предсказание значений

Предсказание осуществляется посредством построения сечения коридора совместных зависимостей в указанных точках. Соотношение прогнозных и исходных интервалов в исходных точках измерений является одним из показателей качества построенной модели.

3. Исследование

3.1. Выбор рассматриваемой области

На рисунке 1 показан график исходных данных.

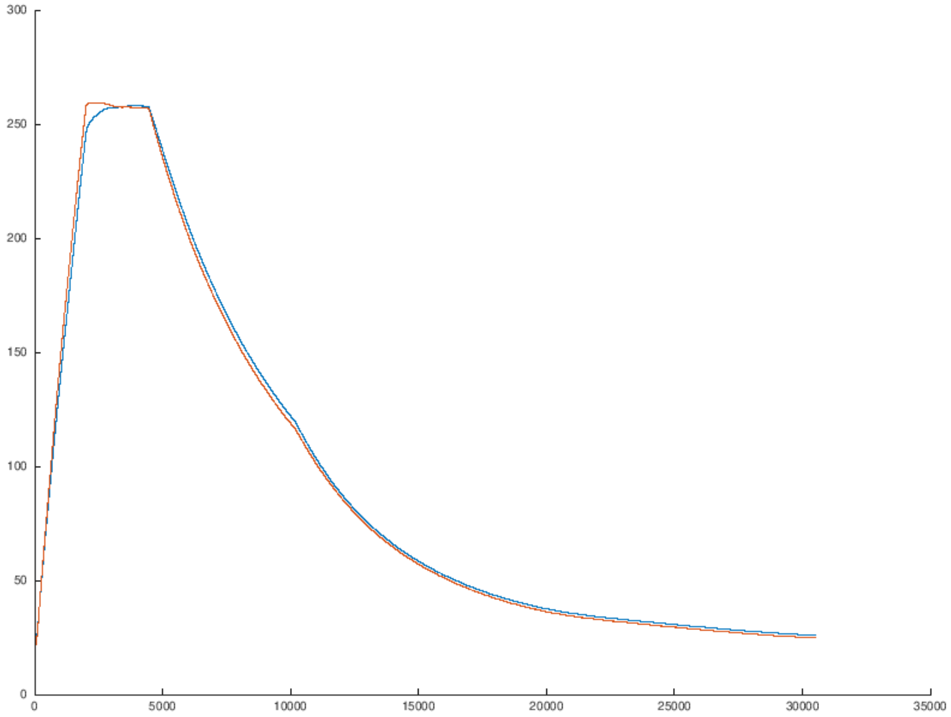


Рис. 1. Исходные данные

Выберем хорошо представимый линейной моделью участок: $x \in [0, 500]$ и $x \in [1000, 2000]$

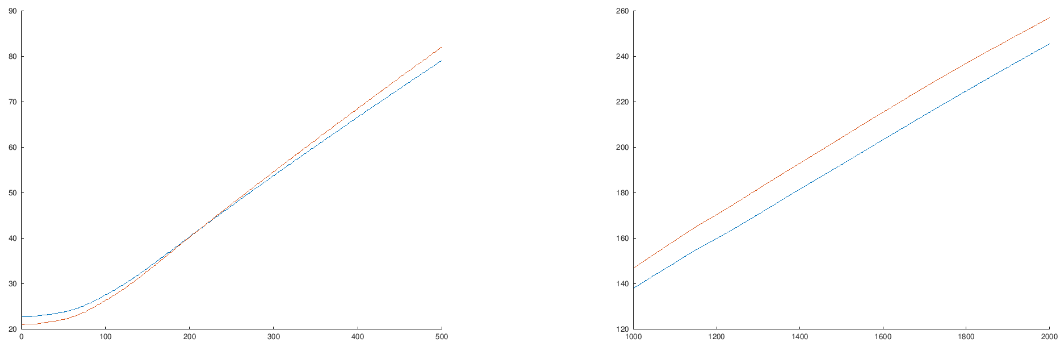


Рис. 2. Уточнённый рассматриваемый участок

Оставим только нижнюю линию, и выберем на ней 10 точек:

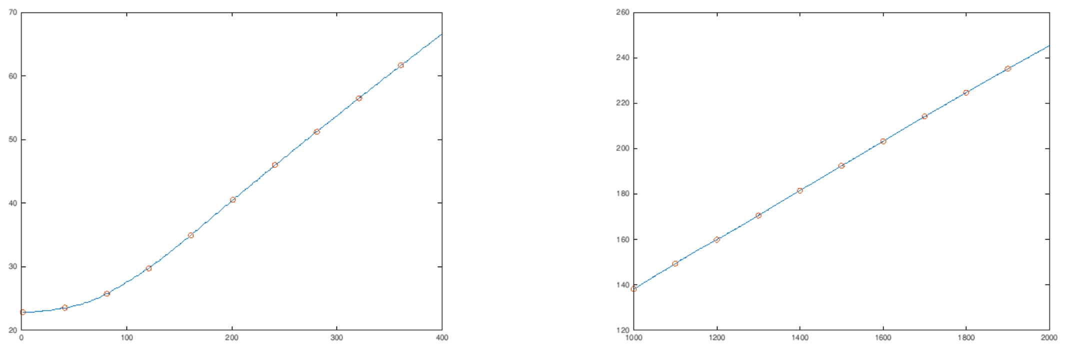


Рис. 3. Уточнённый рассматриваемый участок

Посмотрим на выбранные значения:

	1	2	3	4	5	6	7	8	9	10
x	1	41	81	121	161	201	241	281	321	361
y	22.8	23.5	25.7	29.7	34.9	40.5	46	51.2	56.5	61.7

Таблица 1. $x \in [0, 500]$

	1	2	3	4	5	6	7	8	9	10
x	1000	1100	1200	1300	1400	1500	1600	1700	1800	1900
y	138.1	149.3	159.8	170.5	181.5	192.4	203.2	214.2	224.7	235.2

Таблица 2. $x \in [1000, 2000]$

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

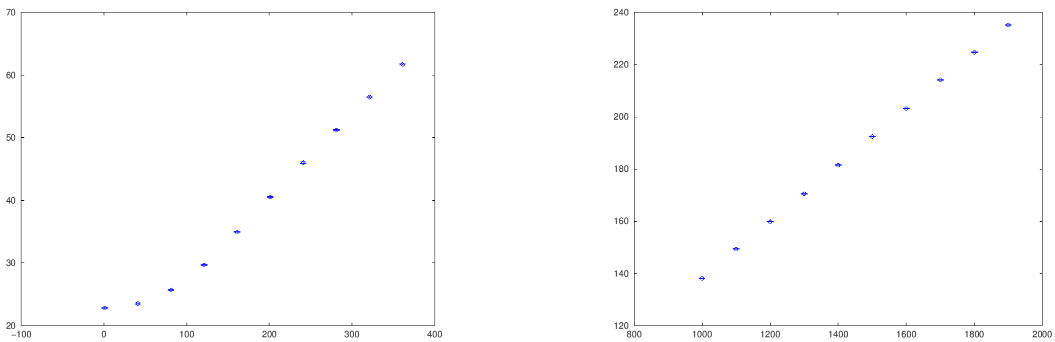


Рис. 4. Входные данные с интервальной неопределённостью $x \in [0, 500]$ и $x \in [1000, 2000]$

3.2. Параметры модели

Точечная оценка параметров регрессии. Сначала проведем точечную оценку параметров регрессии. Пусть модель задается в классе линейных функций

$$y = \beta_1 + \beta_1 x, \tag{2}$$

где x — номер измерения в выборке; y — угол поворота вала двигателя.

В результате на участке $x \in [0, 500]$ получены значения $\beta_1 = 18.3226$ и $\beta_1 = 0.1156$, на участке $x \in [1000, 2000]$ получены значения $\beta_1 = 30.2285$ и $\beta_1 = 0.1080$.Таким образом, по результатам построения линейной модели методом МНК имеем следующий вид:

$$y = 18.3226 + 0.1156x, x \in [0, 500] \tag{3}$$

$$y = 30.2285 + 0.1080x, x \in [1000, 2000] \tag{4}$$

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:

Для согласования с данными поставим задачу оптимизации и решим методами линейного программирования. В соотвествии с подходом к варьированию величины неопределенности поставим задачу в виде

$$\begin{cases} mid\ y_i - w_i \cdot rad\ y_i \le X\beta \le mid\ y_i + w_i \cdot rad\ y_i, \quad i = 1, m, \\ \sum\limits_{i=1}^{m} w_i \longrightarrow \min \\ w_i \ge 0, \quad i = 1, 2, \ \ldots, \ m, \\ w, \beta = ? \end{cases}$$

Здесь X — матрица $m \times 2$, в первом столбце которой элементы, равные 1, во втором — значения x_i. В качестве значений середины и радиуса возьмем $mid\ y_i = y_i$ и $rad\ y_i = 1$.

Значение весов в задаче оптимизации:

$$w = [71.5, 27.75, 1.0, 11.75, 10.5, 5.25, 1.0, 1.0, 2.5, 3.75]$$
$$\beta = [18.3226, 0.1156] x \in [0, 500]$$
$$w = [1.0, 2.80, 1.000, 1.60, 1.0, 1.0, 1.0, 2.6, 1.0, 3.800]$$
$$\beta = [18.3226, 0.1156] x \in [1000, 2000]$$

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

$$rad\ y_i := \varepsilon = \max_i \varepsilon_i w_i. \tag{5}$$

Информационное множество параметров I. 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:

$$\hat{\beta}_{maxdig} = \frac{1}{2}(b_1 - b_2). \tag{6}$$

где b_1 и b_2 - наиболее удалённые друг от друга вершины многогранника.

- Центр тяжести информационного множества:

$$\hat{\beta}_{gravity} = \frac{1}{n}\sum_{i=1}^{n} b_i. \tag{7}$$

где b_i – вершина многогранника, n – количество.

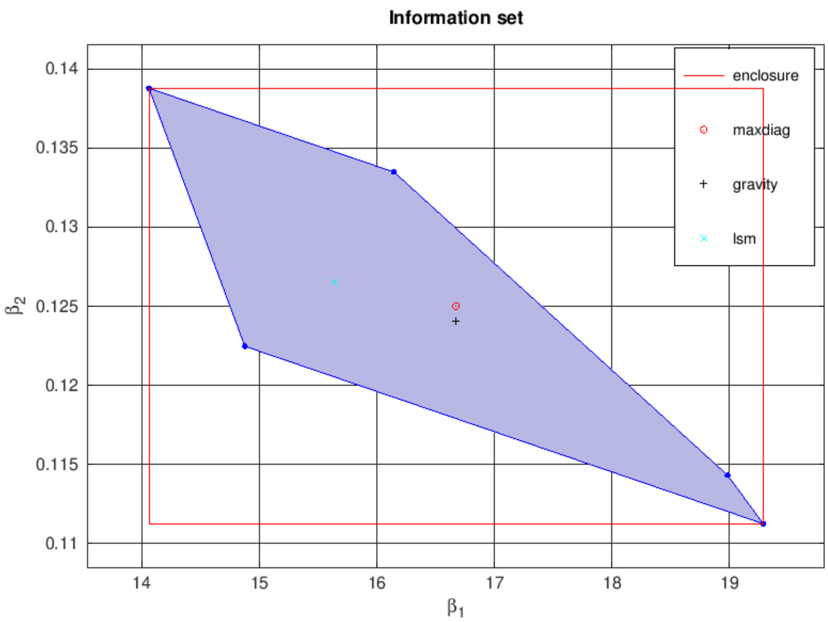


Рис. 5. Информационное множество I $x \in [0, 500]$

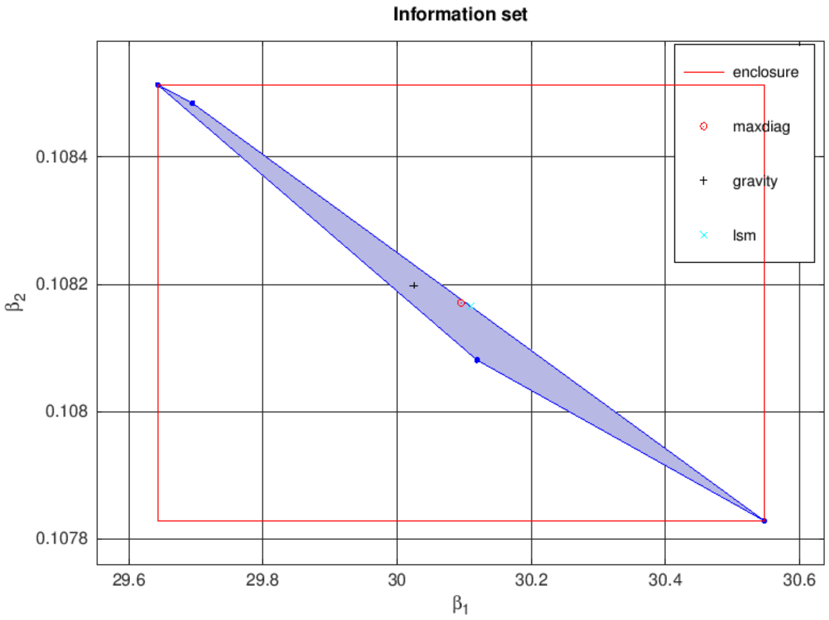
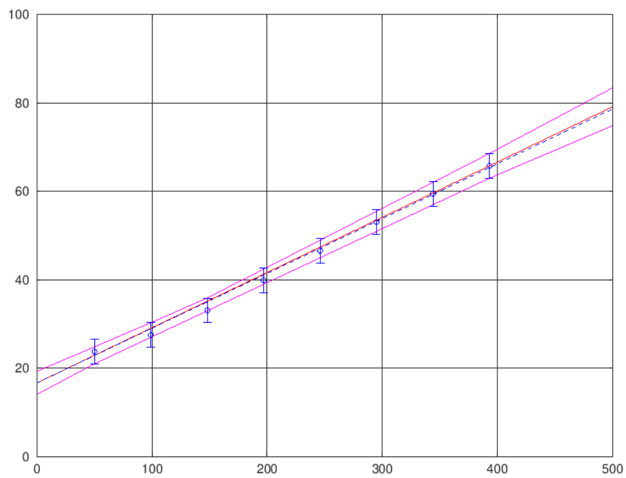


Рис. 6. Информационное множество I $x \in [1000, 2000]$

Коридор совместности γ. Коридор совместности γ. На рис. 6 изображены диаграмма рассеяния данных и коридор совместности для полученных параметров модели регрессии для заданной модели погрешности данных.соответствующим центру тяжести множества, показанного на рис. 5.



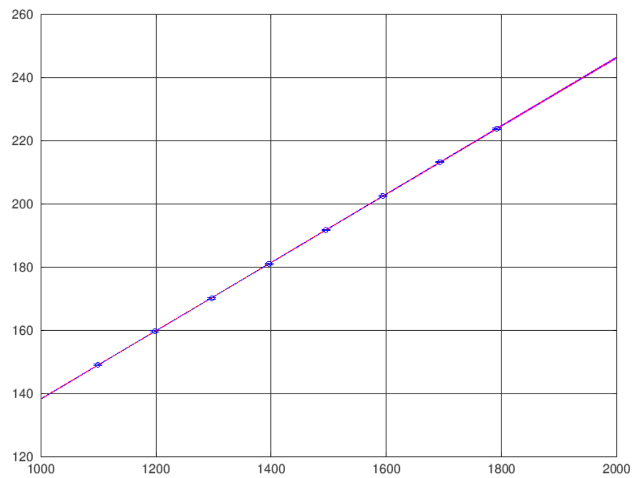


Рис. 7. Коридор совместных зависимостей, весь диапазон в участке $x \in [0, 500]$ и $x \in [1000, 2000]$

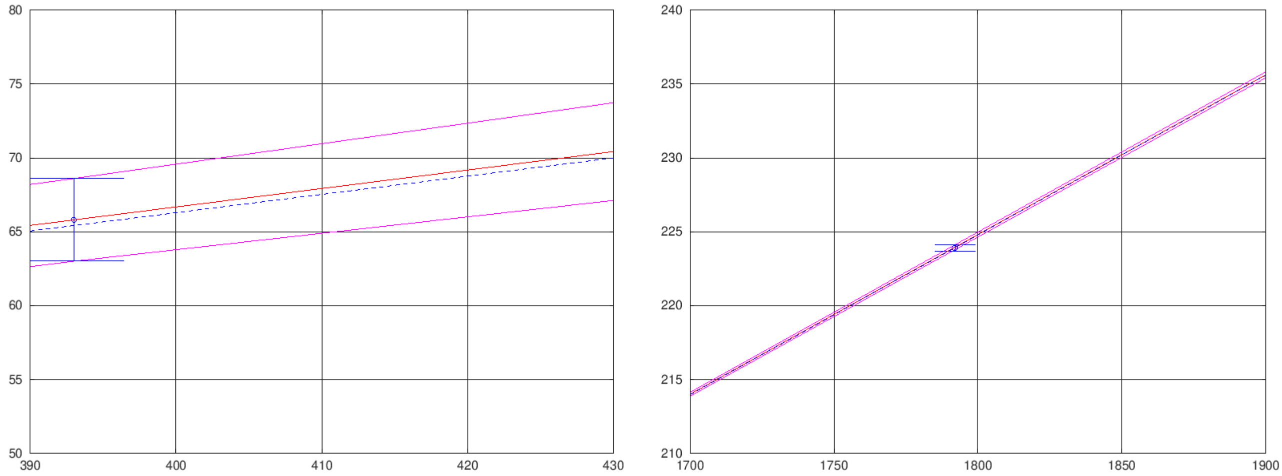


Рис. 8. Коридор совместных событий в окрестности первого наблюдения $x \in [0, 500]$ и $x \in [1000, 2000]$

Прогноз значений выходной переменной. Важнейшим назначением регрессионной модели является предсказание значений выходной переменной для заданных значений входной.

С помощью информационного множества I для построенной модели.

$$y(x) = [14.061, 19.289] + [0.111, 0.139]x, x \in [0, 500]. \tag{8}$$

На основании этой модели получим прогнозируемые значения выходной переменной. Пусть

$$x_p = [30, 100, 200, 300, 400]. \tag{9}$$

Тогда

x_p	y_p	rad y_p
30	[18.22,22.63]	0.465
100	[27.12,30.41]	0.228
200	[39.37,42.84]	0.148
300	[51.61,56.19]	0.200
400	[63.78,69.57]	0.865

С помощью информационного множества I для построенной модели.

$$y_1(x) = [29.643, 30.547] + [0.108, 0.109]x_1, x_1 \in [1000, 2000]. \tag{10}$$

На основании этой модели получим прогнозируемые значения выходной переменной. Пусть

$$x_{p1} = [1000, 1200, 1400, 1600, 1800]. \tag{11}$$

Тогда

x_{p1}	y_{p1}	rad y_p
30	[138.2,138.4]	0.108
100	[159.8,159.9]	0.062
200	[181.4,181.6]	0.070
300	[203.0,203.3]	0.111
400	[224.6,225.0]	0.165

где y_p - интервальный прогноз значений у в точках x_p

rad y_p - радиус прогнозных интервалов. Неопределённость прогноза растёт по мере удаления от области, в которой производились исходные измерения.

Граничные точки множества совместности на участке $x \in [0, 500]$. В данном случае граничными оказались точки с номерами 1, 2, 3, 8. Убедимся в этом посмотрев детально на каждую из точек подробнее

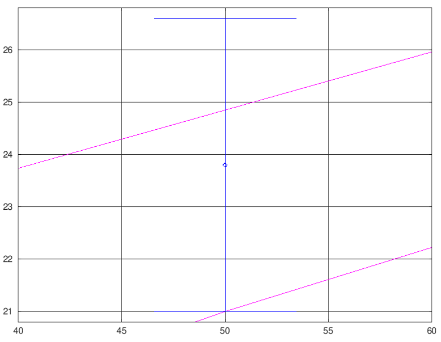


Рис. 9. точка 1

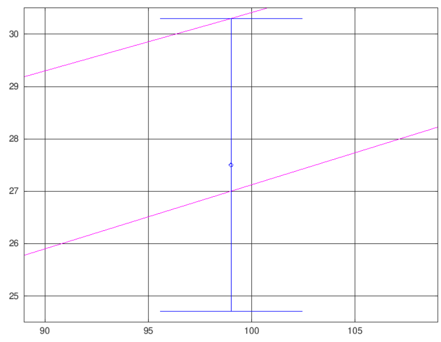


Рис. 10. точка 2

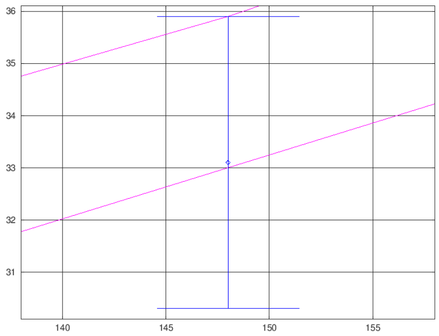


Рис. 11. точка 3

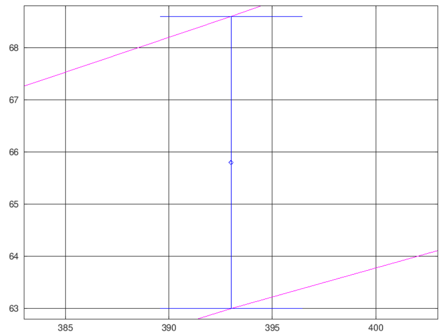


Рис. 12. точка 4

Как мы видим, точки 2 и 3 касаются верхней границы множества. Точка 1 – нижней. точка 8 касаются верхней и нижней границы множества. Убедимся, в том, что остальные точки (4,5,6,7) не являются граничными.

Тем самым набор точек [1, 2, 3, 8] может полностью определить модель.

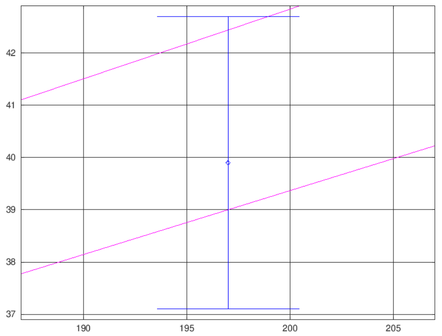


Рис. 13. точка 4

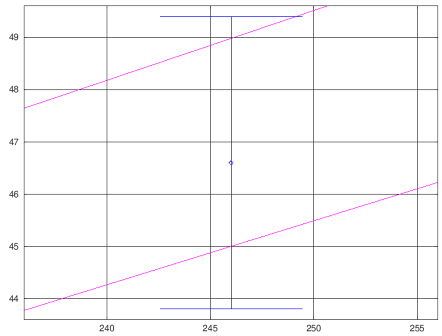


Рис. 14. точка 5

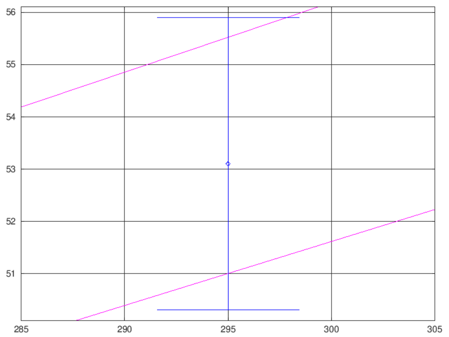


Рис. 15. точка 4

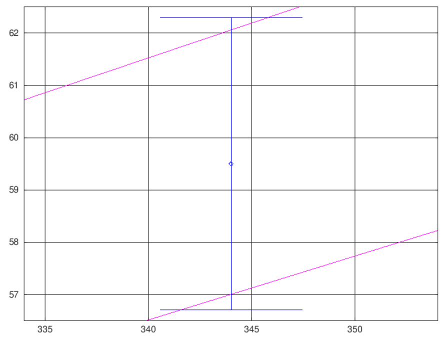


Рис. 16. точка 5

4. Обсуждение

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.

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 совместно- сти и обнаружены граничные точки коридора совместности. По полученной модели были вычислены прогнозы за пределами области измерений.

5. 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-example (20.02.2022).

3. https://github.com/Li-Rui-QI/interval-uncertainty.git