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
Демьянов Д.С.

Преподаватель
Баженов А.Н.

Санкт-Петербург
2021 год

Оглавление

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 отрицательной первой производной: $\beta_2 < 0$

Так как в нашем случае y = f(x) определена не однозначно, сначала будем искать решение в классе функций $x = \beta_1' + \beta_2' y$.

Ниже показан график исходных данных:

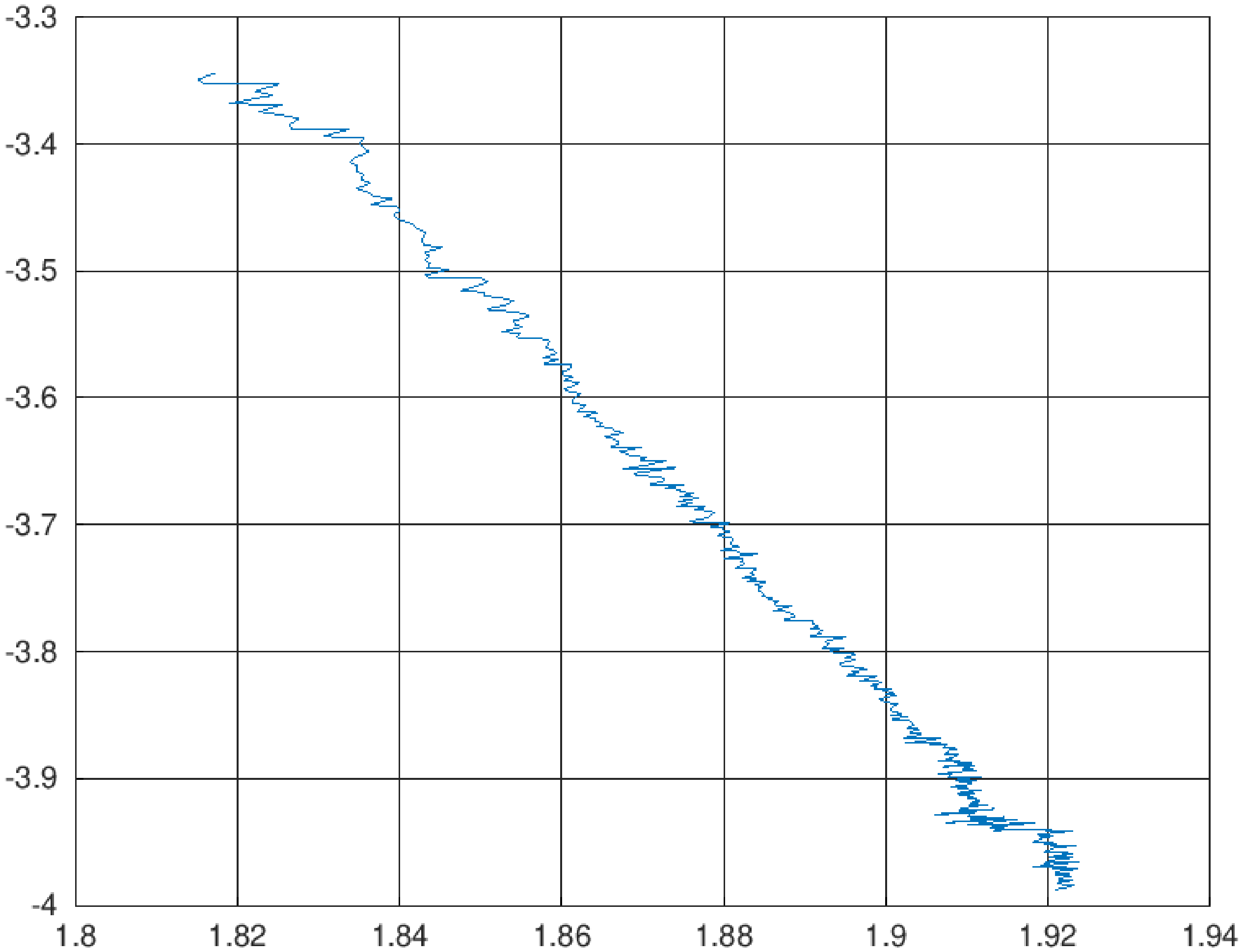

Рисунок 1 Исходные данные

Особенность задачи, что неопределенность имеют входные данные.

Решение

Подготовка данных

Выберем на данной области 5 точек:

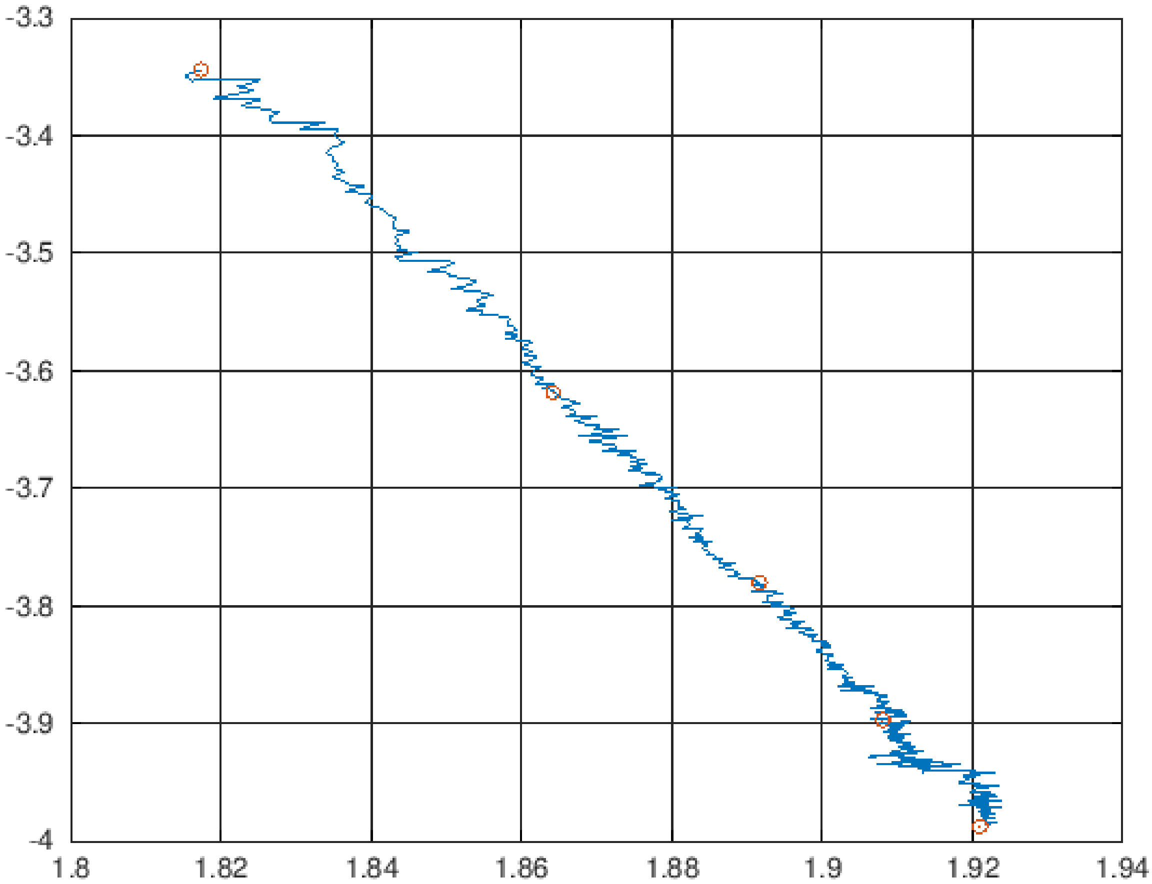

Рисунок 2 Выбранные точки из исходных данных

Посмотрим на выбранные значения:

$x = [1.81732\ , 1.86417, 1.89169, 1.90809, 1.92104]$
$y = [-3.34412, -3.61846, -3.78029, -3.89696, -3.98753]$

В качестве начальной погрешности зададим $\varepsilon = 0.05$, одинаковую для всех наблюдений.

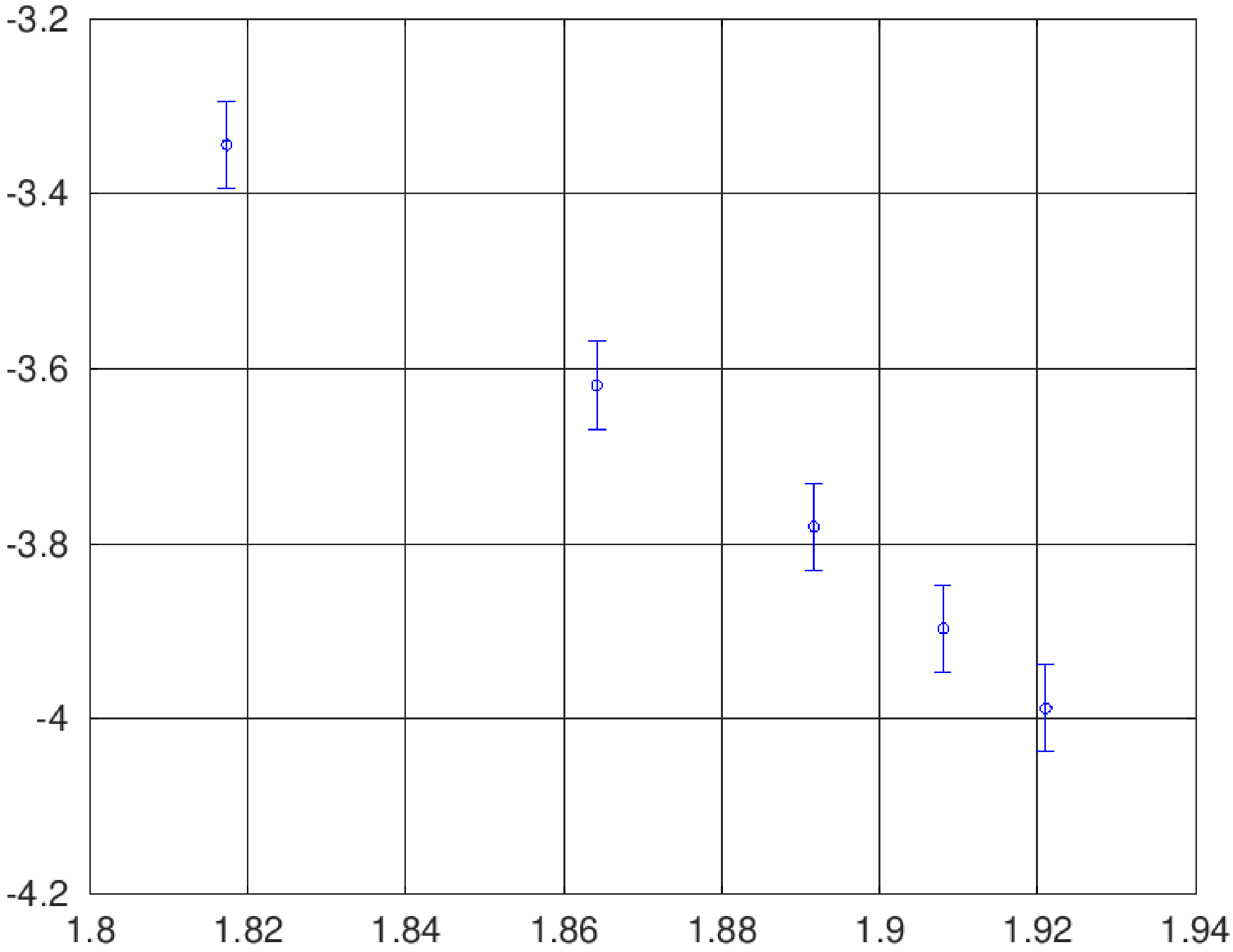

Рисунок 3 Входные данные с интервальной неопределённостью

Параметры модели

Сперва построим линейную модель методом МНК как на точечных значениях:

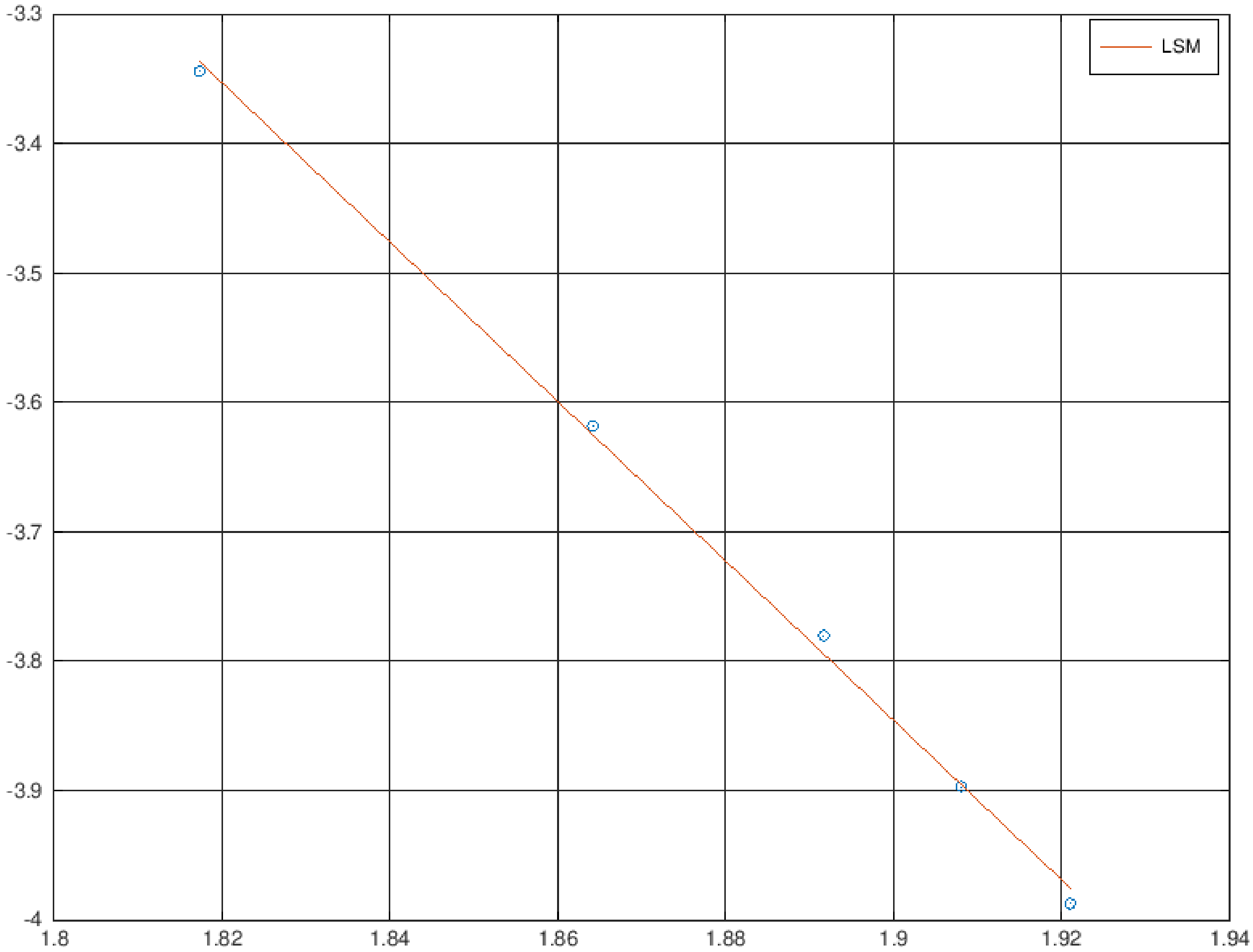

Рисунок 4 МНК линейная регрессия

$$\beta_1' = 7.8565, \quad \beta_2' = -6.1591$$

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]:

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

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

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

$$w, \beta' = ?$$

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

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

Значение весов в задаче оптимизации:

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

$$\beta = [6.1272, -5.2392]$$

Увеличим погрешность всех измерений:

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

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:

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

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.

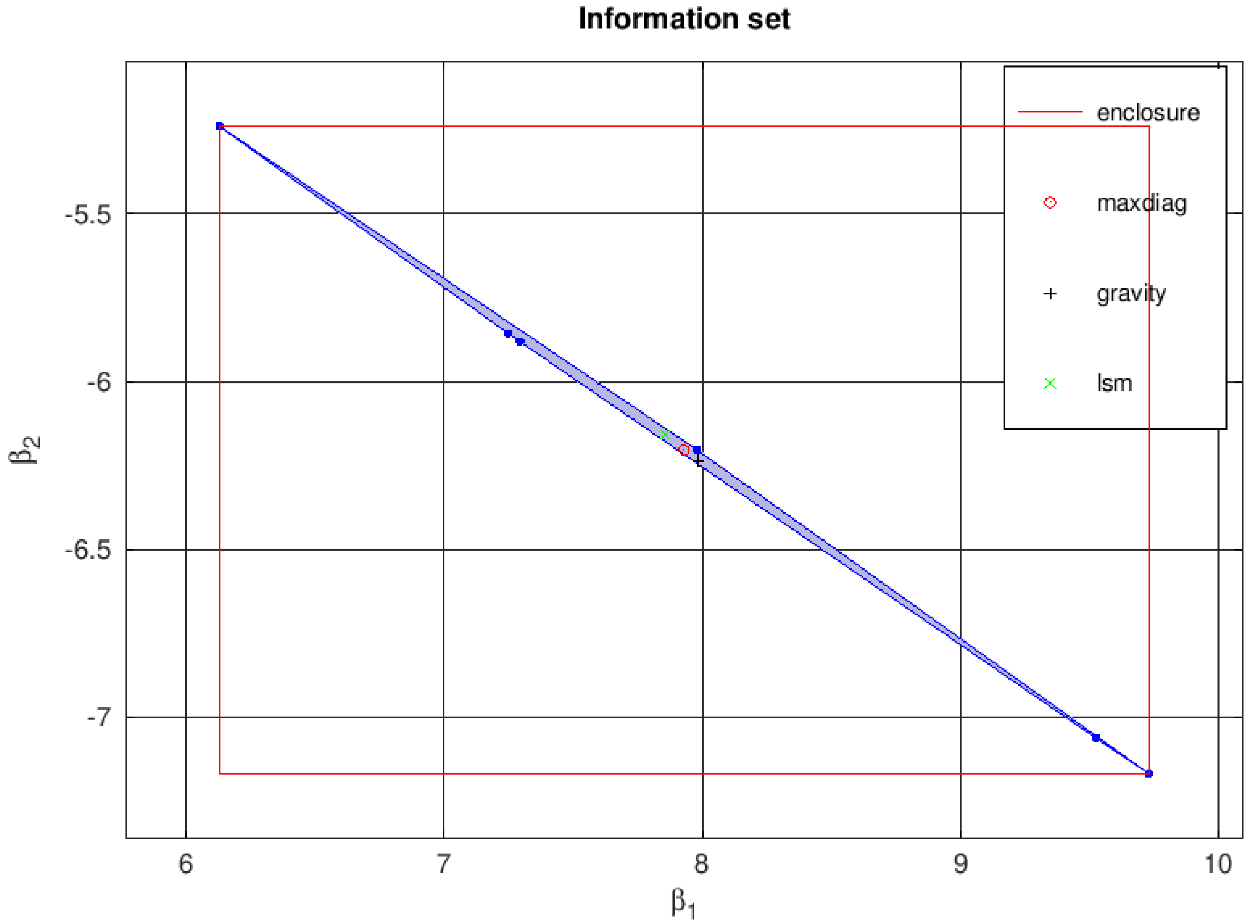

Рисунок 5 информационное множество линейной модели

Заметим, что значения, полученные при помощи МНК оказались внутри границ информационного множества.

Коридор совместных зависимостей

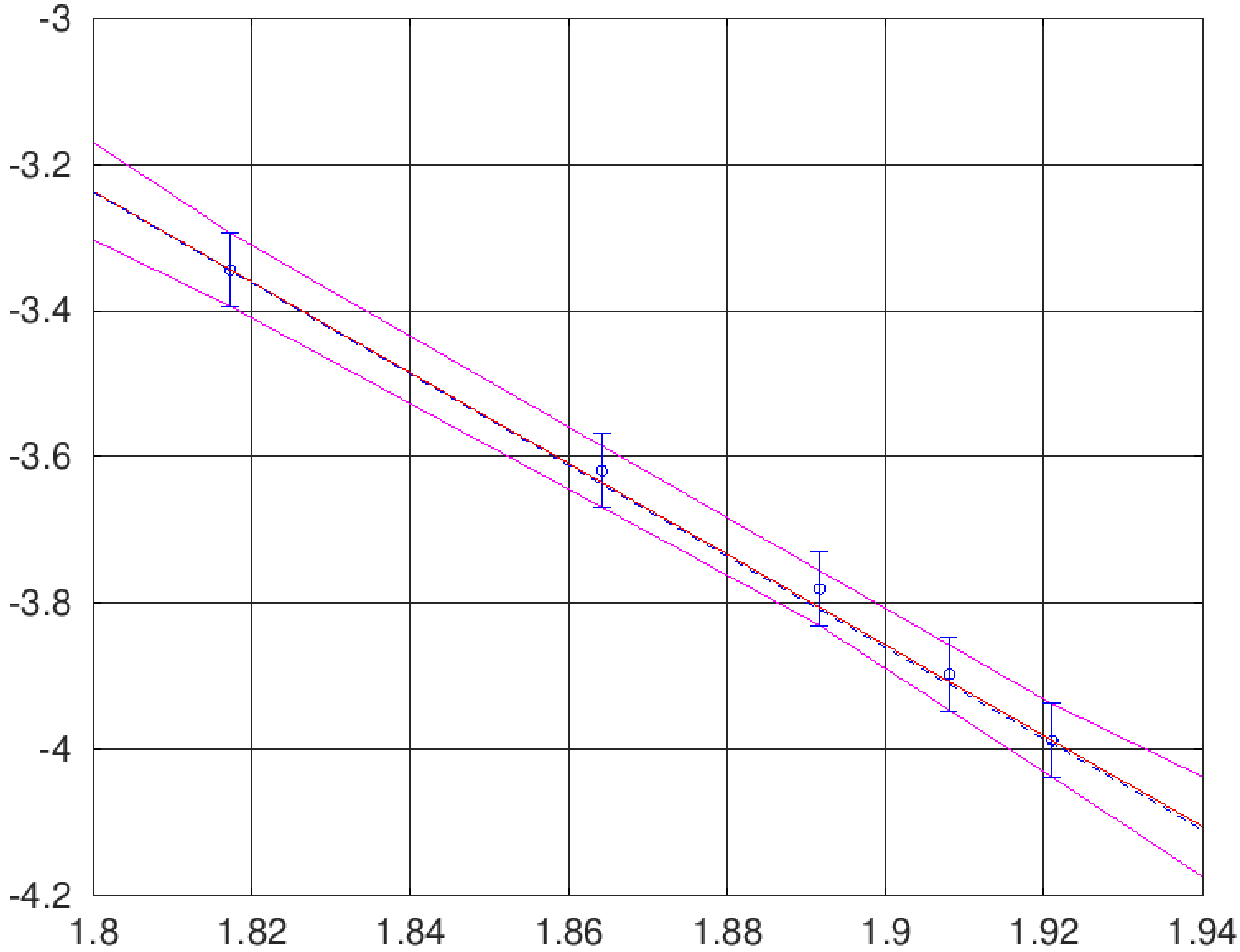

Рисунок 6 Коридор совместных зависимостей, весь диапазон

Прогноз за пределы интервала:

С помощью построенной выше модели мы получили интервальную оценку β_1', β_2' и зависимость x = g(y)

$$\widehat{\boldsymbol{x}}(y) = [6.1272, 9.7315] + [-7.1675, -5.2392]y$$

Перейдем теперь обратно к $y = \beta_1 + \beta_2 x$.

$$\beta_1 = -\frac{\beta_1'}{\beta_2'}$$

$$\beta_2 = \frac{1}{\beta_2'}$$

Получается: $\widehat{\beta_1} = [0.8549, 1.8754]$, $\widehat{\beta_2} = [-0.1909, -0.1395]$

$$\widehat{\boldsymbol{y}}(x) = [0.8549, 1.8754] + [-0.1909, -0.1395]x$$

Можно получить прогнозные значения выходной переменной:

Возьмём 3 точки:

$$x_p = [1.85;\ 1.9;\ 2.0;\ 3.0;\ 5.0]$$

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

x_p	y_p	$rad\ y_p$
1.85	[-3.5855, -3.4968]	0.0443
1.9	[-3.8890, -3.8070]	0.0409
2.0	[-4.6035, -4.3512]	0.1261
3.0	[-11.7709, -9.5904]	1.0902
5.0	[-26.1059, -20.0688]	3.0185

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, 3, 5. Именно эти точки могут полностью самостоятельно определить модель.

Заключение

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

Было получено информационное множество для параметров линейной модели, построен коридор совместности и обнаружены граничные точки коридора совместности. Так как по изначальным данным y = f(x) – зависимость была задана неоднозначно, задача линейной регрессии решалась для x = g(y) и дополнительно был сделал обратный переход.

По полученной модели были вычислены прогнозы за пределами области измерений. Результаты совпали с ожидаемыми.

Приложение:

Ссылка на проект с кодом реализации:

https://github.com/DimaDemyanov/Stochastic-models

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