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 группы 5040102/00201 Н.В. Суханов
Руководитель
доцент, к.ф.-м.н. А.Н. Баженов

Санкт-Петербург
2022

Оглавление

Постановка задачи 3
Исходные данные 3
Модель регрессии 4
Информационное множество и коридор совместности 5
Прогноз дальнейших значений 7
Выводы 8
Реализация 8
Использованная литература 9

Постановка задачи

Построить линейную модель интервальной регрессии для несовместных данных. Рассмотреть параметры модели, информационное множество и коридор совместности, построить прогноз за пределы выборки.

Исходные данные

В качестве исходных данных рассмотрим зависимость «медианного числа обусловленности» матрицы длин хорд [1] от числа элементов разбиения. Медианное число обусловленности будем определим, как отношение максимального сингулярного числа к медиане сингулярных чисел. Для матриц с нечётным n оно равно числу обусловленности матрицы, которая получится при отбрасывании из исходной всех столбцов, которым соответствуют сингулярные числа меньше медианного.

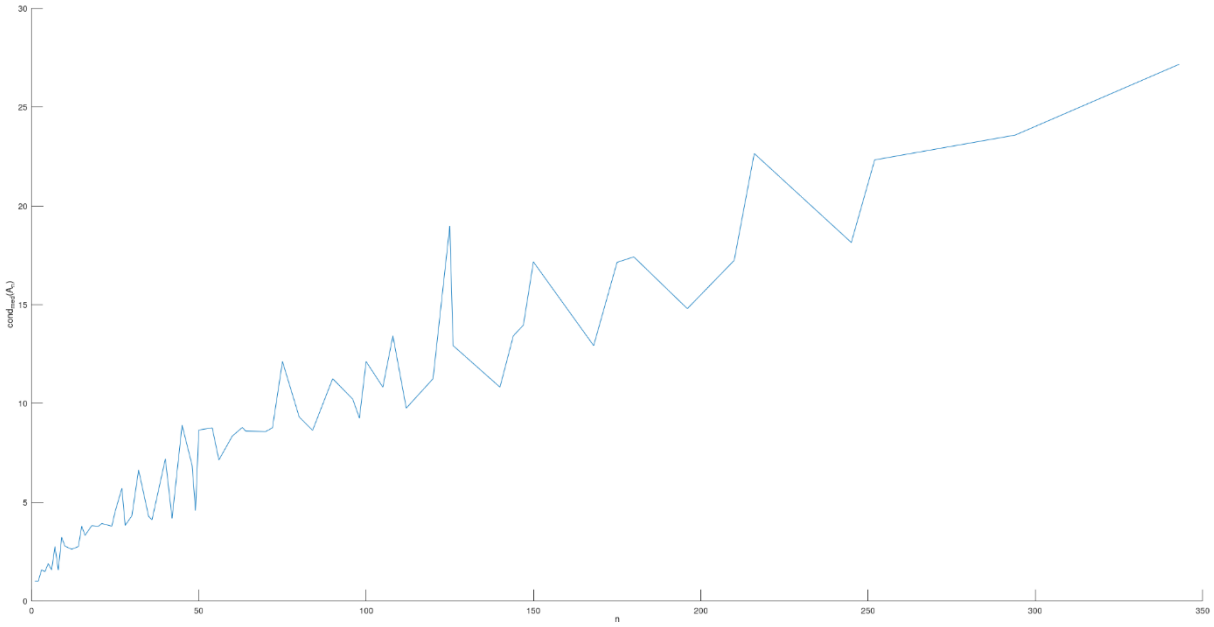

Рис. 1 Исходные данные

На Рис. 1 приведена исходная выборка целиком. Для построения модели ограничимся первыми двадцатью значениями (Рис. 2).

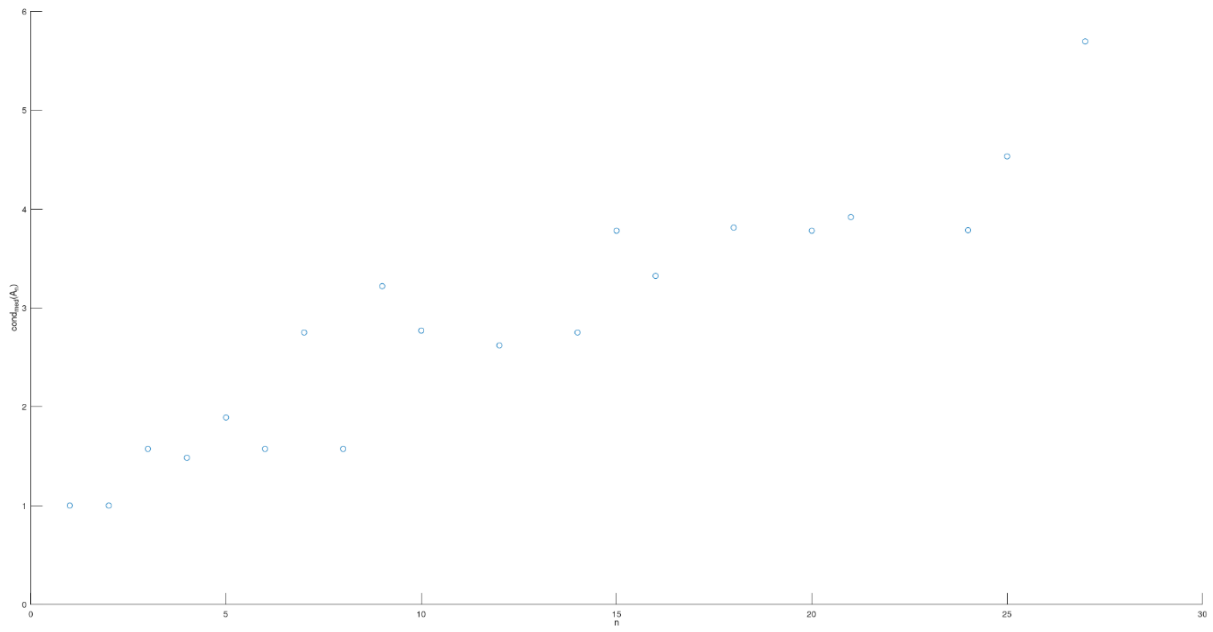

Рис. 2 Данные для построения модели

Модель регрессии

Интервализуем данные с радиусом 0.1. На данном этапе выбор радиуса не является принципиальным, так как в дальнейшем для большинства элементов нам придётся его увеличить, чтобы добиться совместности модели. Для начала построим по центрам данных точечную линейную регрессию в смысле метода наименьших квадратов (Рис. 3).

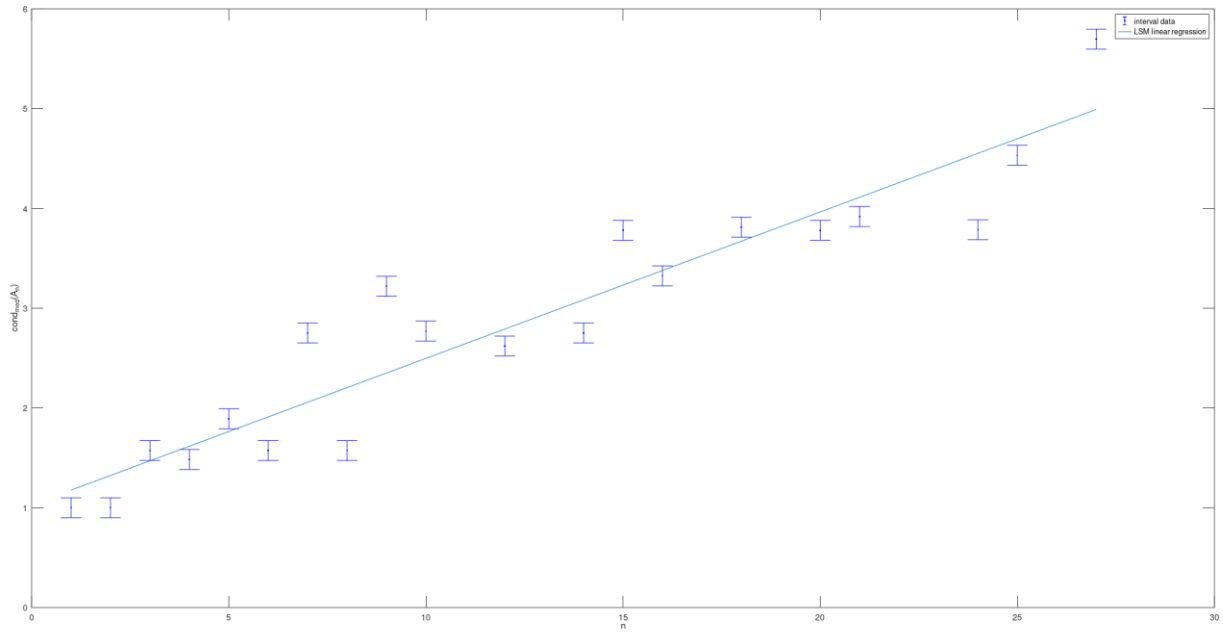

Рис. 3 МНК - линейная регрессия

Линейная регрессия задаётся уравнением $y = \beta_1 + \beta_2 x$ с параметрами $\beta_1 = 1.0301,\ \beta_2 = 0.1467.$

Информационное множество для данной выборки пусто. Чтобы построить совместную интервальную регрессионную модель, увеличим радиусы элементов. Веса в соответствии с которыми необходимо увеличить радиусы можно найти как решение задачи линейного программирования [2]:

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

$$mid\, \boldsymbol{y}_i - \omega_i \cdot rad\, \boldsymbol{y}_i \le X\beta \le\ mid\, \boldsymbol{y}_i + \omega_i \cdot rad\, \boldsymbol{y}_i,$$

$$\omega_i \ge 1\,, i = 1, \dots, n.$$

Здесь X – матрица линейной регрессии:

$$X^{n\times 2} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Расширим радиусы элементов выборки: $rad\, \boldsymbol{y}_i' = rad\, \boldsymbol{y}_i \cdot \omega_i$. Чтобы информационное множество не вырождалось, искусственно увеличим радиусы ещё на 10%. Также из решения задачи ЛП мы получили ещё одну точечную оценку для параметров регрессии: $\beta_1 = 0.98,\ \beta_2 = 0.1451.$

Информационное множество и коридор совместности

На Рис. 4 представлено информационное множество нашей модели. Звездочками обозначены некоторые точечные оценки для параметров регрессии: зелёным цветом – решение задачи ЛП, красным – результат точечной МНК – регрессии, чёрным – центроид (среднее по вершинам информационного множества). Заметим, что МНК – оценка не попала в информационное множество.

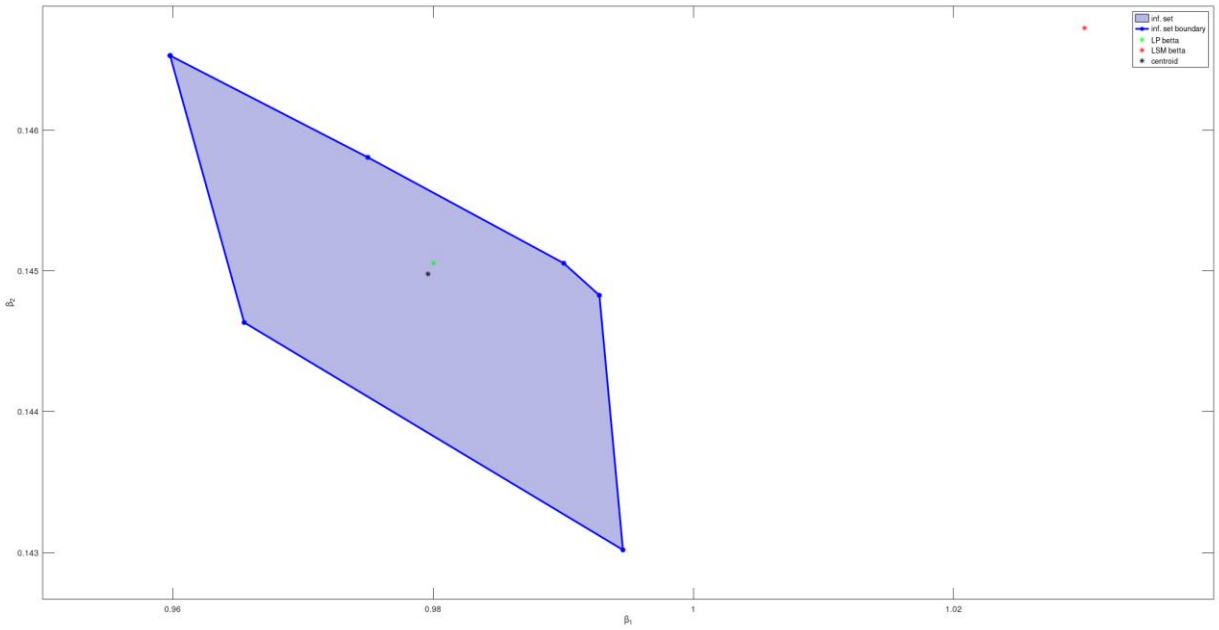

Рис. 4 Информационное множество

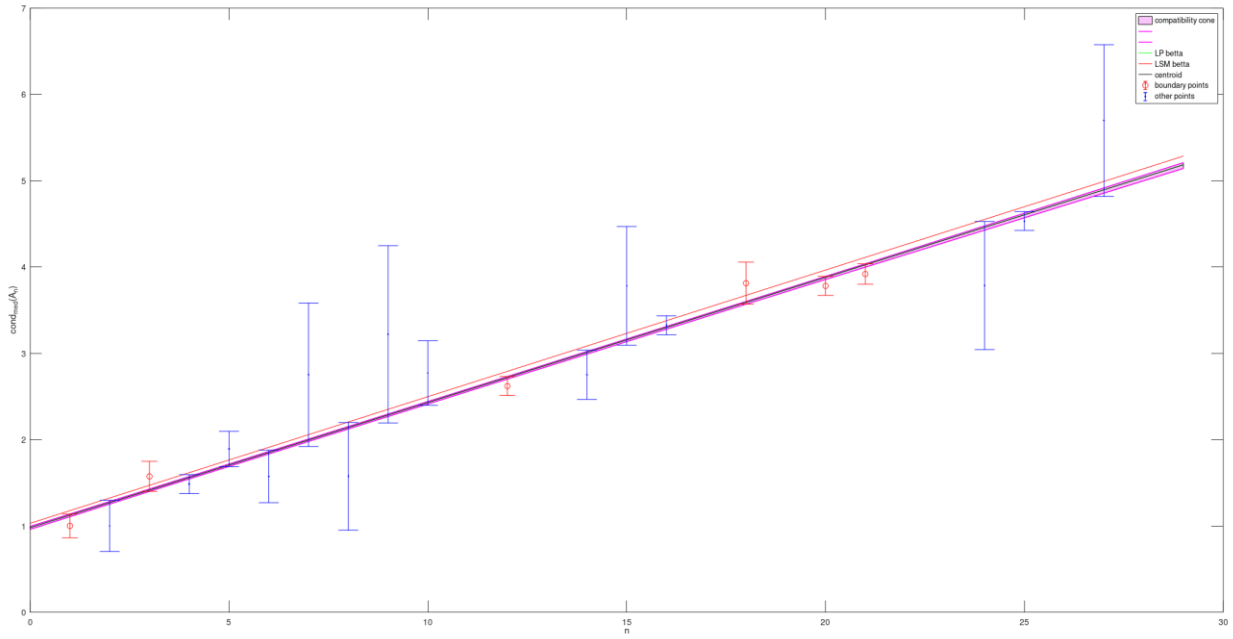

Рис. 5 Коридор совместности

На Рис. 5 представлен коридор совместности модели. Красным выделены граничные элементы выборки, на которые опирается коридор, именно они определяют модель. Граничными оказались элементы с номерами 1, 3, 11, 15, 16, 17, заметим, что их количество совпадает с количеством вершин информационного множества. Подробнее рассмотрим участок коридора совместности рядом с первым элементом (Рис. 6), чтобы убедиться, что точечные оценки соотносятся с коридором также, как и с информационным множеством.

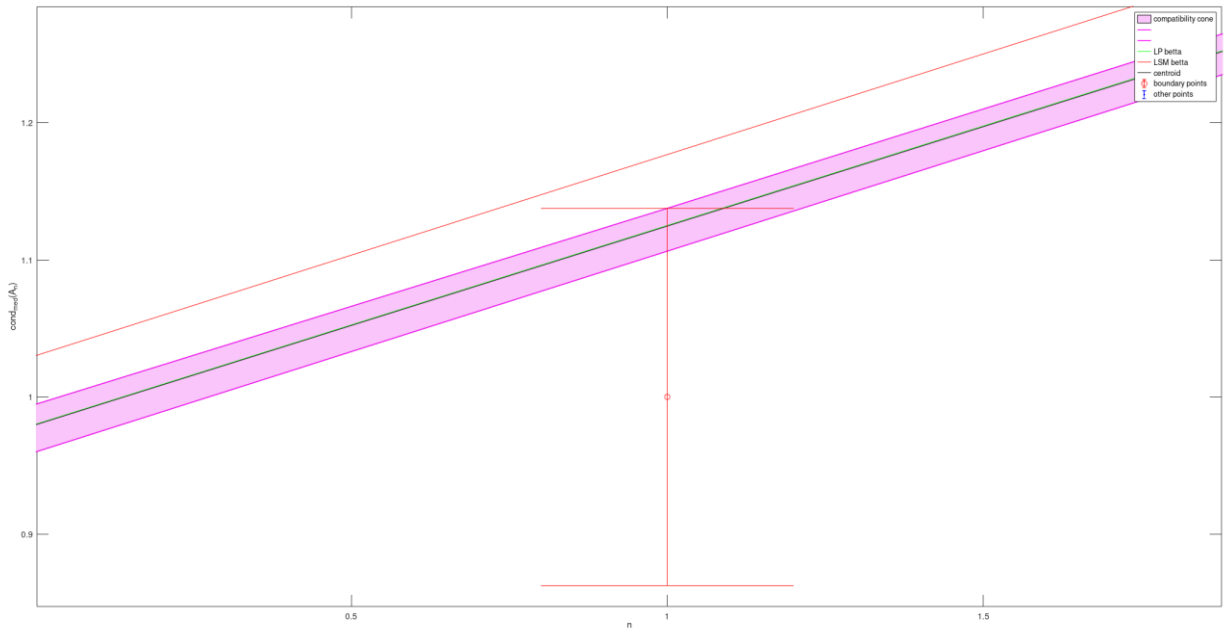

Рис. 6 Коридор совместности в окрестности первой точки

Прогноз дальнейших значений

С помощью построенной модели попробуем спрогнозировать значения для следующих десяти элементов выборки.

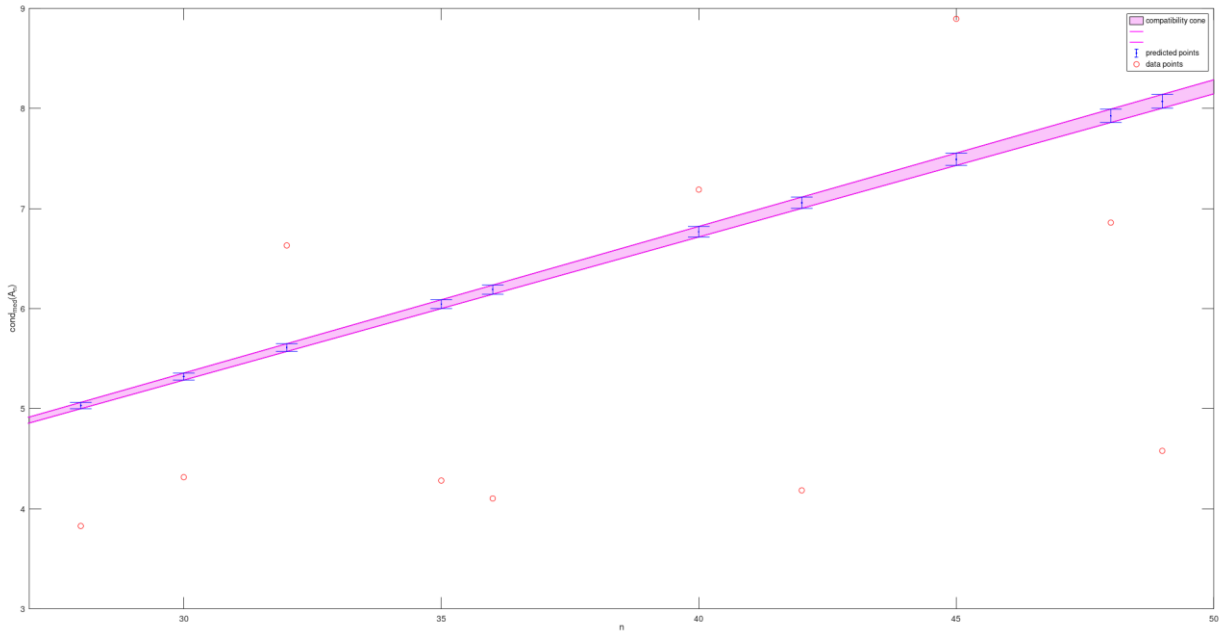

Рис. 7 Прогноз значений

Из Рис. 7 видно, что ни одно из реальных значений для следующих десяти элементов выборки не лежит в коридоре совместности и соответственно не попало ни в один из спрогнозированных интервалов. Также можно заметить, что радиус прогнозируемых элементов, отвечающий за неопределённость прогноза, возрастает по мере удаления от элементов исходной выборки, по которой строилась модель.

Выводы

В ходе работы была построена линейная интервальная регрессионная модель данных, а также несколько точечных оценок параметров регрессии. Для того, чтобы добиться совместности модели, радиусы элементов были скорректированы путём решения вспомогательной задачи линейного программирования. Для получившейся модели были рассмотрены её информационное множество и коридор совместности, был построен прогноз дальнейших элементов выборки.

По результатам работы можно заключить, что зависимость медианного числа обусловленности матрицы длин хорд от числа элементов разбиения плохо описывается линейной моделью (как, вероятно, и полиномиальными моделями более высоких порядков). Построенная модель позволяет описать глобальный тренд на возрастание числа обусловленности, но не может уловить более сложные закономерности в данных, в частности, их существенную немонотонность. По этим же причинам модель не позволяет делать точные прогнозы дальнейших значений или предположения о том, являются ли элементы выборки выбросами.

Реализация

Все приведённые выше расчёты и построения были выполнены в среде GNU Octave с использованием функций для построения интервальной регрессии [3]. Код проекта и исходные данные выложены на GitHub:
https://github.com/NikitaSukhanov/Stochastic_models_and_data_analysis
Данные были сгенерированы на основе цилиндрической модели плазмы [4].

Использованная литература

1. Н.В. Суханов. Отчёт по летней практике «Моделирование вращения плазмы» https://drive.google.com/file/d/1TIfBR3f5uOjEgtACsoVzOIiclgUAdEEh/view?usp=sharing
2. 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.
3. С.И.Жилин. Примеры анализа интервальных данных в Octave https://github.com/szhilin/octave-interval-examples
4. Н.В. Суханов. Plasma3D https://github.com/NikitaSukhanov/Plasma3D