

Санкт-Петербургский политехнический университет Петра Великого
Физико-механический институт
Кафедра «Прикладная математика»

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №4

по дисциплине
«Математическая статистика»

« Задача восстановления линейной зависимости »

Выполнила студентка
группы 5030102/00101

Еремина Ксения Игоревна

Проверил
Доцент, к.ф.-м.н.

Баженов Александр Николаевич

Санкт-Петербург
2023

Содержание

1	Постановка задачи	3
2	Теория	4
3	Результаты	16
3.1	Диаграмма рассеяния	16
3.2	Варьирование неопределенности измерений	16
3.3	Варьирование неопределённости измерений с расширением и сужением интервалов	17
3.4	Анализ регрессионных остатков	19
3.5	Информационное множество задачи	21
3.6	Коридор совместных зависимостей	22
3.7	Построение прогноза внутри и вне области данных	24
4	Обсуждение	25
4.1	Варьирование неопределенности измерений	25
4.2	Варьирование неопределенности измерений с расширением и сужением интервалов	25
4.3	Анализ регрессионных остатков	25
4.4	Информационное множество задачи	25
4.5	Коридор совместных зависимостей	25
4.6	Построение прогноза внутри и вне области данных	25
5	Реализация	26
6	Приложение	26

Список иллюстраций

1	Диаграмма рассеяния выборки \mathbf{X}_1 с уравновешенным интервалом погрешности.	4
2	Диаграмма рассеяния выборки \mathbf{X}_1 и регрессионная прямая по модели (4) и (5).	6
3	Диаграмма рассеяния выборки \mathbf{X}_1 и регрессионная прямая по модели (10) и (11)	7
4	Векторы ω_1 и ω_0 .	8
5	Диаграмма рассеяния по модели (4) и (5).	9
6	Диаграмма рассеяния регрессионных остатков выборки \mathbf{X}_1 по (10) и (11).	10
7	Частоты элементарных подинтервалов регрессионных остатков выборки \mathbf{X}_1 по модели (4) и (5) — красный график, и (10) и (11) — синий график.	11
8	Информационное множество по модели (10) и (11), интервальная оболочка — красный брус.	12
9	Коридор совместных зависимостей (23).	13
10	Коридор совместных зависимостей (23). Построение прогноза.	14
11	Кусочно-линейная регрессионная зависимость.	15
12	Диаграмма рассеяния выборки \mathbf{X}_1 с уравновешенным интервалом погрешности	16
13	Диаграмма рассеяния выборки \mathbf{X}_1 и регрессионная прямая по модели (4) и (5)	17
14	Диаграмма рассеяния выборки \mathbf{X}_1 и регрессионная прямая по модели (10) и (11)	18
15	Векторы ω_0 и ω_1	19
16	Диаграмма рассеяния по модели (4) и (5)	20
17	Диаграмма рассеяния регрессионных остатков выборки \mathbf{X}_1 по (10) и (11)	20
18	Частоты элементарных подинтервалов регрессионных остатков выборки \mathbf{X}_1 по моделям (4), (5) и (10), (11)	21
19	Информационное множество по модели (10) и (11), интервальная оболочка — желтый брус	22
20	Коридор совместных зависимостей (23)	23
21	Коридор совместных зависимостей (23). Построение прогноза.	24

1 Постановка задачи

Дадим общую формулировку задачи восстановления функциональной зависимости. Пусть некоторая величина y является функцией от независимых переменных x_1, x_2, \dots, x_m :

$$y = f(\beta, x) \quad (1)$$

где $x = (x_1, x_2, \dots, x_m)$ является вектором независимых переменных, $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ — вектор параметров функции. Заметим, что переменные x_1, x_2, \dots, x_m также называются входными, а переменные y_1 — выходной.

Задача восстановления функциональной зависимости заключается в том, чтобы, располагая набором значений x и y , найти такие $\beta_1, \beta_2, \dots, \beta_p$ в выражении (1), которые соответствуют конкретной функции f из параметрического семейства.

Если функция f является линейной, то можно записать

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m \quad (2)$$

В общем случае результаты измерений величин x_1, x_2, \dots, x_m и y являются интервальнозначными

$$x_1^{(k)}, x_2^{(k)}, \dots, x_m^{(k)}, y^{(k)}.$$

Индекс k пробегает значения от 1 до n , равного полному числу измерений.

Определение 2.2.1 Брусом неопределенности k -го измерения функциональной зависимости будем называть интервальный вектор-брус, образованный интервальными результатами измерений с одинаковыми значениями индекса k [1]:

$$(x_{k1}, x_{k2}, \dots, x_{km}, y_k) \subset \mathbb{R}^{m+1}, k = 1, 2, \dots, n. \quad (3)$$

Брус неопределенности измерения является прямым декартовым произведением интервалов неопределенности независимых переменных и зависимой переменной.

2 Теория

Данные выборки. Имеется выборка данных X_1 с интервальной неопределённостью. Число отсчётов в выборке равно 200.

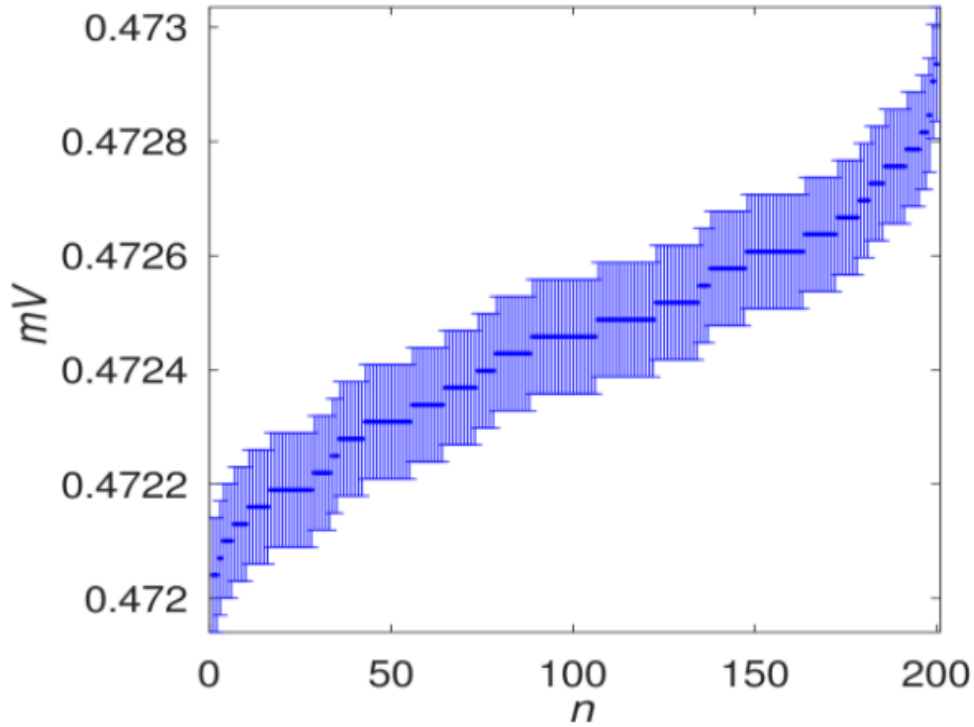


Рис. 1: Диаграмма рассеяния выборки X_1 с уравновешенным интервалом погрешности.

На Рис. 1 представлены данные с прибора [23] с учётом погрешности измерительного прибора.

Построим линейную модель данных и посмотрим, насколько удачно она описывает линейный тренд.

Варьирование неопределённости измерений. Если величину коррекции каждого интервального наблюдения выборки выражать коэффициентом его уширения $\omega_i \geq 1$, а общее изменение выборки характеризовать суммой этих коэффициентов, то минимальная коррекция выборки в виде вектора коэффициентов $\omega = (\omega_1, \dots, \omega_n)$, необходимая для совместности задачи построения зависимости $x = \beta_0 + \beta_1 * i$ может быть найдена решением задачи условной оптимизации

$$\min_{\omega, \beta} \sum_{i=1}^n \omega_i \quad (4)$$

при ограничениях

$$\begin{cases} mid x_i - \omega_i \epsilon_i \leq \beta_0 + \beta_1 * i \leq mid x_i + \omega_i \epsilon_i, \\ \omega_i \geq 1, \end{cases} \quad i = 1, \dots, n. \quad (5)$$

Результирующие значения коэффициентов ω_i , строго превосходящие единицу, указывают на наблюдения, которые требуют уширения интервалов неопределённости для обеспечения совместности данных и модели.

Проведём вычисление параметров линейной регрессии по данным интервальной выборки \mathbf{X}_1 с использованием программ С.И.Жилина [8] и оформленных применительно к задаче на [23]. Синтаксис вызова программы

$$[tau, w, yint] = DataLinearModel(input1, epsilon0) \quad (6)$$

В (6) входами программы служат значения $mid \mathbf{X}_1$ и величин неопределённости ϵ , а выходами tau — значения параметров регрессии β_0, β_1 w — вектор весов расширения интервалов.

На Рис. 2 красным цветом приведена регрессионная прямая.

Вычисления с использованием программы (6) дают следующие результаты для регрессионных коэффициентов

$$\beta_0 = tau(1) = 4.7203e - 01, \quad (7)$$

$$\beta_1 = tau(2) = 4.0915e - 06. \quad (8)$$

Все компоненты вектора ω оказались равны 1, то есть, расширения интервалов измерений не понадобилось. Таким образом, величина (4) равна числу элементов выборки.

$$\min_{\omega, \beta} \sum_{i=1}^n \omega_i = 200 \quad (9)$$

Недостатком полученного решения с единичными значениями ω_i является неучёт расстояний точек регрессионной зависимости до данных интервальной выборки. Таким образом, прямая с параметрами

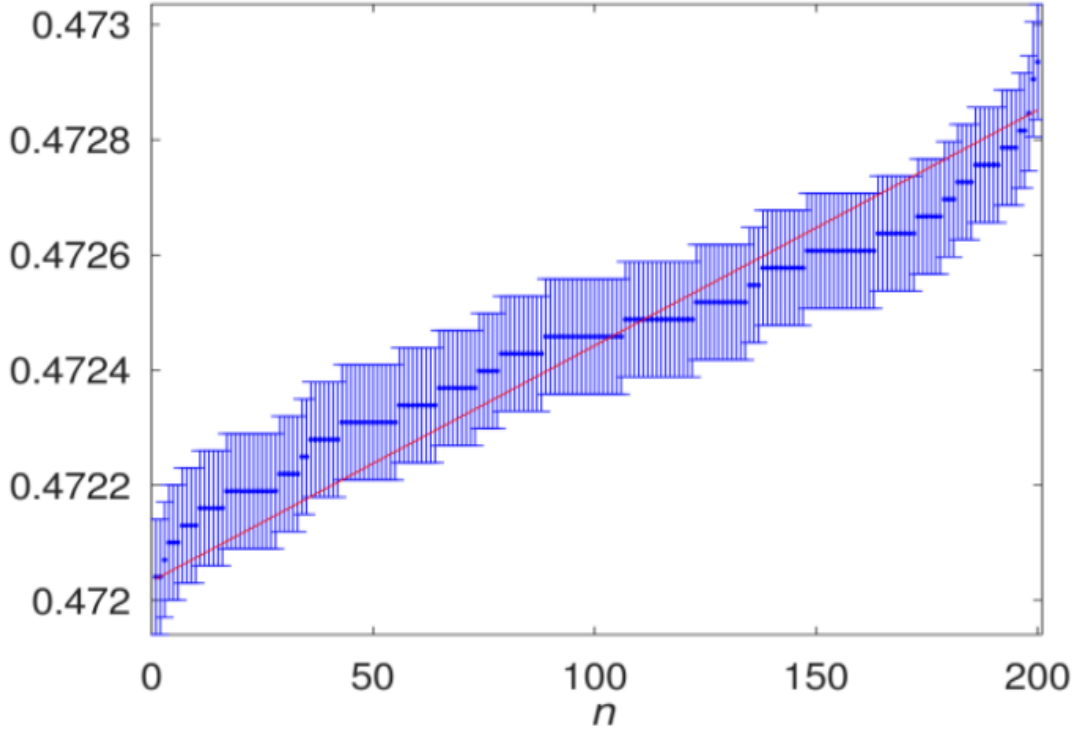


Рис. 2: Диаграмма рассеяния выборки \mathbf{X}_1 и регрессионная прямая по модели (4) и (5).

(7) и (8) «не чувствует» отклонений измерений от прямой на концах выборки — неопределённости измерений достаточно велики, чтобы покрыть этот эффект.

Варьирование неопределённости измерений с расширением и сужением интервалов. Выясним, что даёт решение задачи оптимизации другим способом, с расширением и сужением интервалов.

Поставим задачу условной оптимизации следующим образом:

$$\min_{\omega, \beta} \sum_{i=1}^n \omega_i \quad (10)$$

при ограничениях

$$\begin{cases} mid x_i - \omega_i \epsilon_i \leq \beta_0 + \beta_1 * i \leq mid x_i + \omega_i \epsilon_i, \\ \omega_i \geq 0, \end{cases} \quad i = 1, \dots, n. \quad (11)$$

Отличие постановки от (4) и (5) состоит в том, что интервалы измерений могут как расширяться в случае $\omega_i \geq 1$, так и сужаться при $0 \leq \omega_i \leq 1$. Вычисление параметров линейной регрессии по данным интервальной выборки \mathbf{X}_1 производится как и в случае (6) с использованием программ С.И.Жилина [8] и оформленных применительно к задаче на [23]. Синтаксис

вызова программы

$$[\tau, w, y_{int}] = \text{DataLinearModelZ}(\text{input1}, \text{epsilon0}) \quad (12)$$

Входы и выходы функции DataLinearModelZ такие же, как и для DataLinearModelZ (6).

На Рис. 3 красным цветом приведена регрессионная прямая.

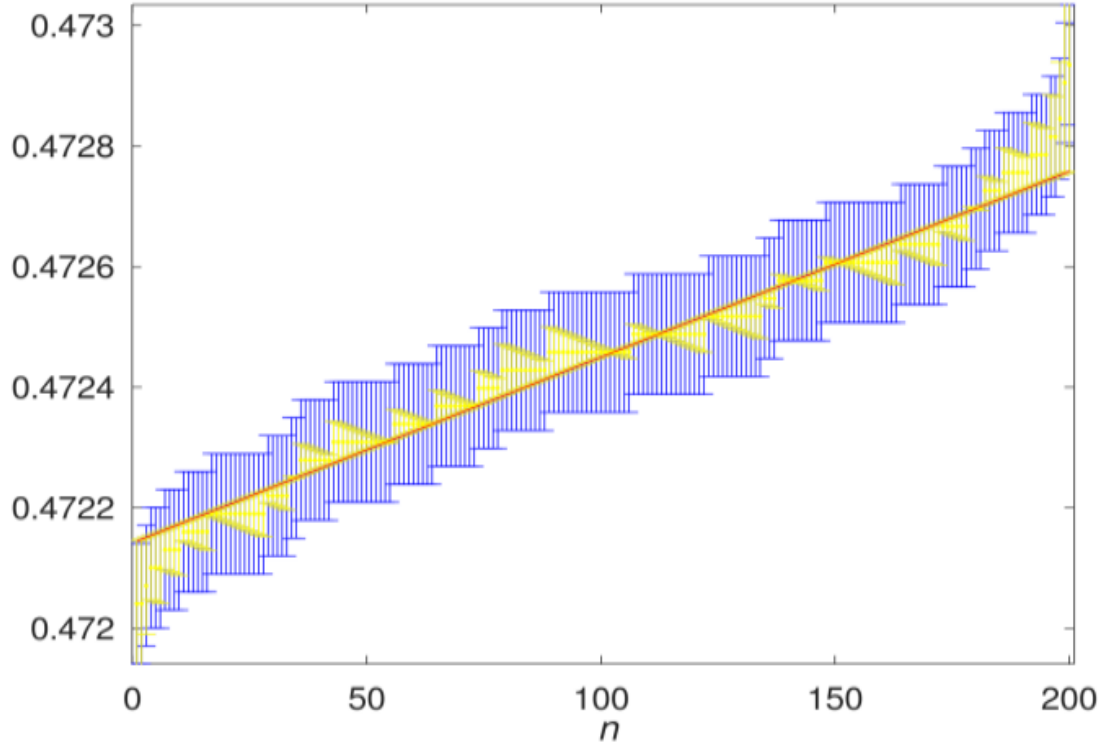


Рис. 3: Диаграмма рассеяния выборки \mathbf{X}_1 и регрессионная прямая по модели (10) и (11)

Жёлтым цветом на Рис. 3 показаны скорректированные интервалы выборки \mathbf{X}_1 . Небольшая часть интервалов на границах области расширилась, а большинство интервалов в диапазоне замеров примерно от 20 до 180 — сузилось.

Величина меры (4) уменьшилась более, чем в 4 раза.

$$\min_{\omega, \beta} \sum_{i=1}^n \omega_i = 45.7 \leq 200 \quad (13)$$

Таким образом, постановка задачи с возможностью одновременного увеличения и уменьшения радиусов неопределённости измерений позволяет более гибко подходить к задаче оптимизации.

На Рис. 4 приведены графики векторов ω_0 и ω_1 , полученных при использовании двух рассмотренных подходов.

В конкретном случае график вектора ω_0 для постановки задачи оптимизации (10) и (11) содержит большое количество информации.

Например, задавшись каким-то порогом α : $0 < \alpha \leq 1$, можно выделить области входного аргумента Ψ , в которых регрессионная зависимость хуже соответствует исходным данным. Например:

$$\Psi = \arg_i \omega_i \geq \alpha \quad (14)$$

Для конкретного примера имеем две области Ψ в начале и конце области данных.

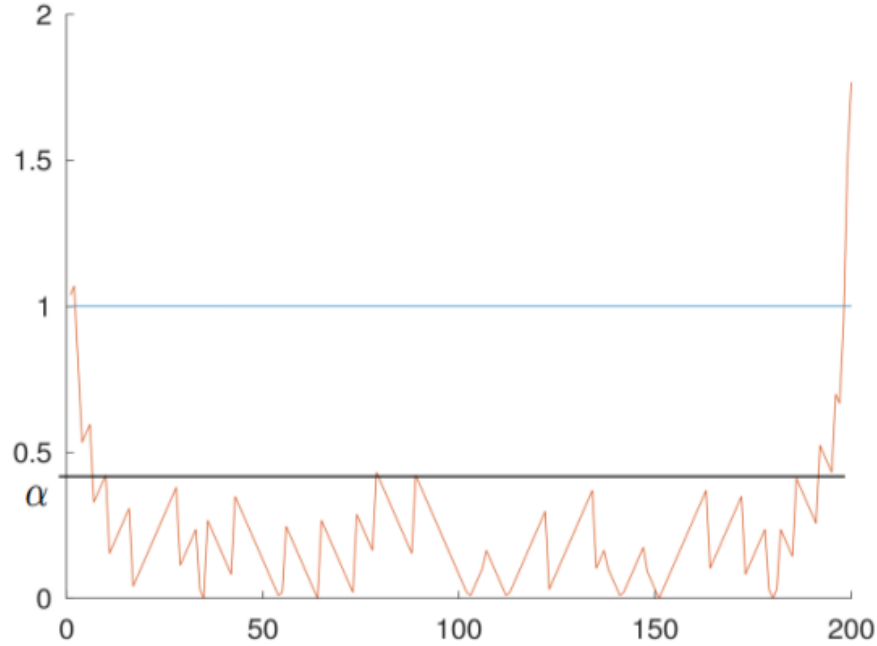


Рис. 4: Векторы ω_1 и ω_0 .

Для объективного использования этого приёма параметр α можно брать, например, из анализа гистограммы распределения вектора ω .

Использование выделения «подозрительных» областей даёт основу для других приёмов. Например, для построения кусочно-линейной регрессионной зависимости.

Анализ регрессионных остатков. В теоретико-вероятностной математической статистике анализ регрессионных остатков — один из приёмов оценки качества регрессии.

Приведём пример пояснения этого приёма. «Если выбранная регрессионная модель хорошо описывает истинную зависимость, то остатки должны быть независимыми, нормально распределёнными случайными величинами с нулевым средним, и в их значениях должен отсутствовать тренд. Анализ регрессионных остатков — это процесс проверки выполнения этих условий.» <https://wiki.loginom.ru/articles/discrepancy.html>

В случае интервальных выборок мы не задаёмся вопросом о виде распределения остатков, а будем использовать те возможности которые появляются при описании объектов и результатов вычислений в виде интервалов.

На Рис. 5 приведена диаграмма рассеяния регрессионных остатков выборки \mathbf{X}_1 по модели (4) и (5).

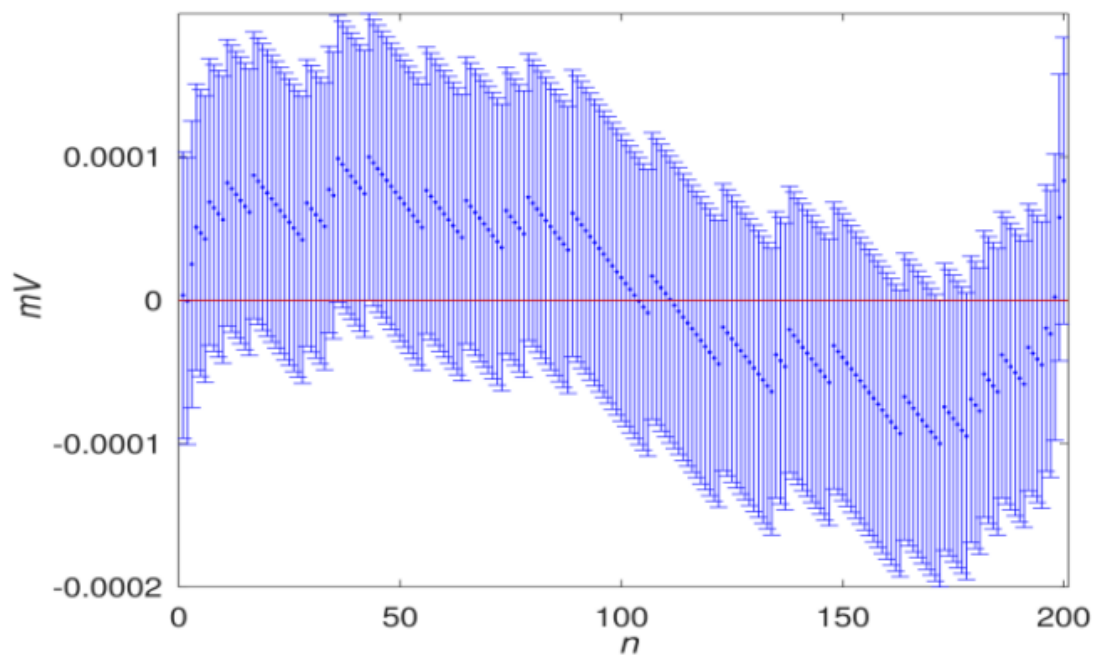


Рис. 5: Диаграмма рассеяния по модели (4) и (5).

На Рис. 6 приведена диаграмма рассеяния регрессионных остатков выборки \mathbf{X}_1 по модели (10) и (11).

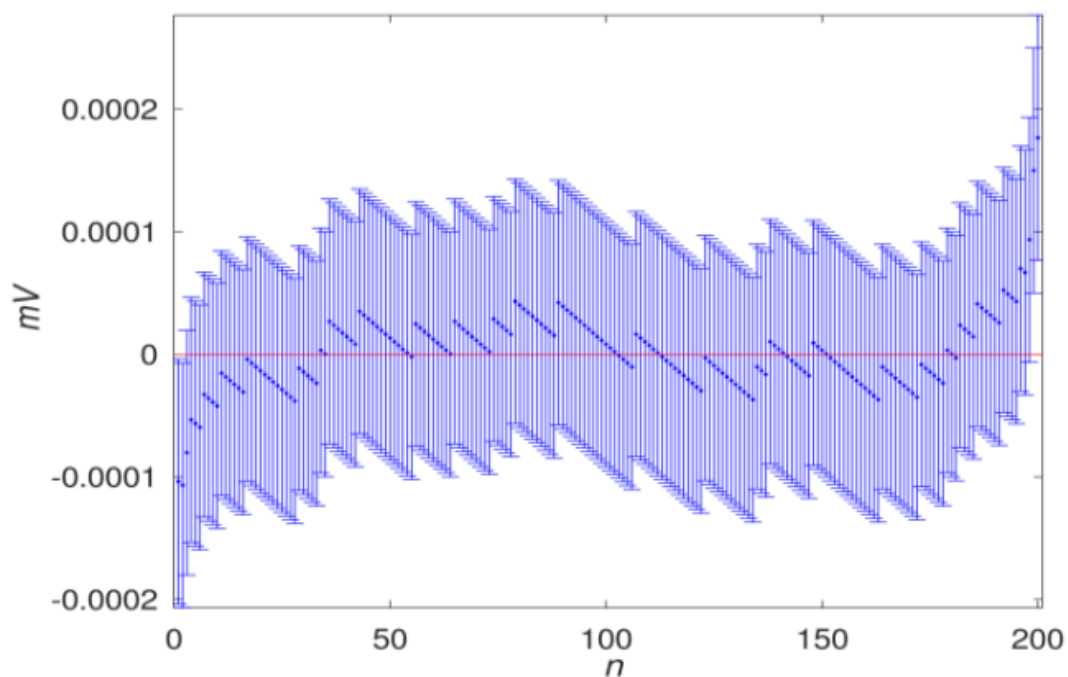


Рис. 6: Диаграмма рассеяния регрессионных остатков выборки \mathbf{X}_1 по (10) и (11).

Из сравнения Рис. 5 и На Рис. 6 видно, что интервальные выборки остатков получились с весьма разными свойствами. Формально диаграмма рассеяния на первом рисунке 'уже, то есть внешняя оценка более компактная. В то же время вторая диаграмма рассеяния выглядит более естественно.

На Рис. 7 приведены графики частот элементарных подинтервалов при вычислении интервальной моды для двух моделей.

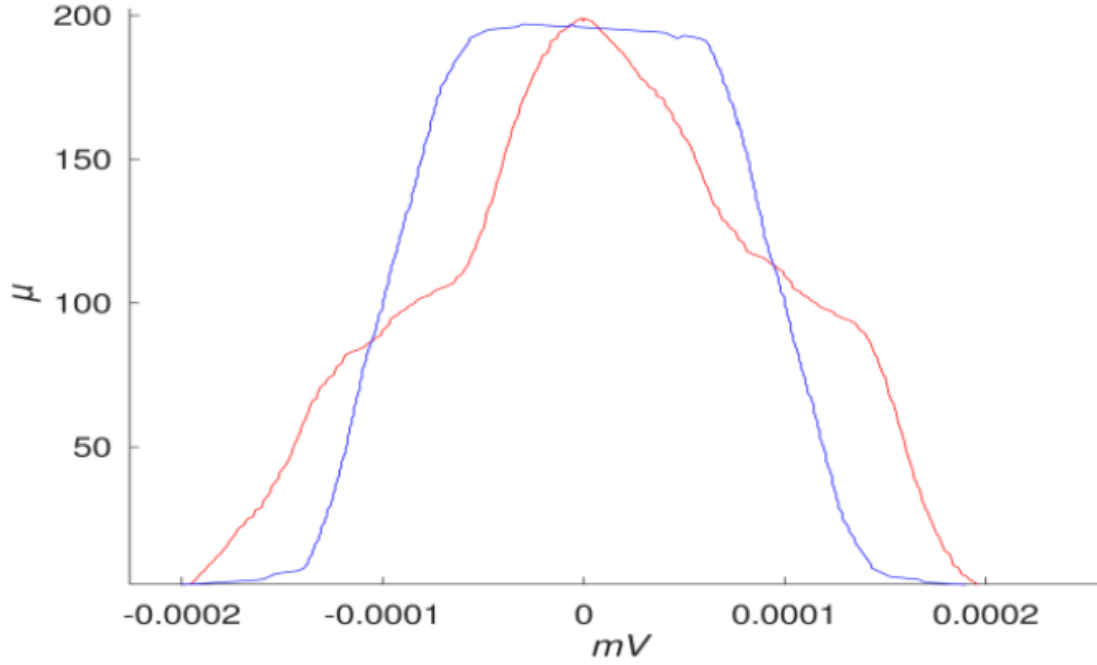


Рис. 7: Частоты элементарных подинтервалов регрессионных остатков выборки \mathbf{X}_1 по модели (4) и (5) — красный график, и (10) и (11) — синий график.

Как и в случае анализа диаграмм рассеяния, второй график выглядит более естественно. Его внутренняя оценка существенно шире, что соответствует большей устойчивостью к возмущениям данных.

К остаткам можно применить и другие меры совместности оценки постоянной величины, описанные ранее.

$$mode \mathbf{X}^1 = \dots \quad (15)$$

$$Ji(\mathbf{X})^1 = \dots \quad (16)$$

$$\vdots \quad (17)$$

$$mode \mathbf{X}^2 = \dots \quad (18)$$

$$Ji(\mathbf{X})^2 = \dots \quad (19)$$

$$\vdots \quad (20)$$

здесь $\mathbf{X}^{1,2}$ — регрессионные остатки выборки \mathbf{X}_1 , вычисленные с использованием разных условий оптимизации.

Информационное множество задачи. Интервальные оценки параметров.

Один из главных вопросов при построении регрессии – оценивание её параметров. В зависимости от прикладных целей характер и назначение искомых оценок могут существенно различаться.

Внешняя интервальная оценка параметра определяется минимальным и максимальным значениями, которых может достигать значение параметра в информационном множестве.

В совокупности интервальные оценки параметров задают брус, описанный вокруг информационного множества и именуемый внешней интервальной оболочкой информационного множества:

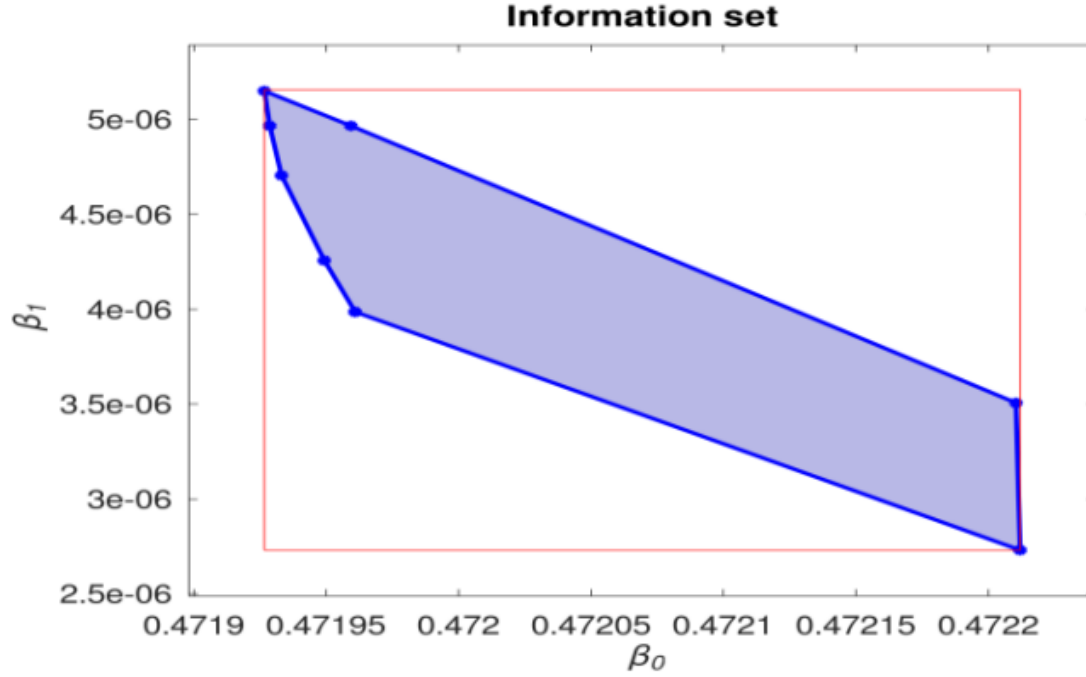


Рис. 8: Информационное множество по модели (10) и (11), интервальная оболочка — красный брус.

Проведём вычисление параметров линейной регрессии по данным интервальной выборки X_1 с использованием программ С.И.Жилина [8].

Синтаксис вызова программ:

Решение задачи линейного программирования

$$SS = ir_problem(A, x, max(w0) * epsilon, lb);$$

Вершины информационного множества задачи построения интервальной регрессии

$$vertices = ir_beta2poly(SS);$$

Внешние интервальные оценки параметров модели $y = \beta_1 + \beta_2 * x$

$$b_{int} = ir_outer(SS).$$

Входами программы служат значения $mid \mathbf{X}_1$ и величин неопределённости ϵ , умноженные на расчётное уширение по модели (10) и (11), матрица A , составленная из нулевой и первой степеней номеров замеров, параметры условной оптимизации. Структура SS содержит значения параметров регрессии.

Коридор совместных зависимостей. Информационное множество задачи определяется в пространстве параметров. Каждая его точка задаёт зависимость в пространстве переменных. Множество всех таких моделей именуется коридором совместных зависимостей.

Выше мы нашли внешние интервальные оценки параметров модели

$$mid \beta_0 = [4.7193e - 01, 4.7221e - 01], \quad (21)$$

$$mid \beta_1 = [2.7304e - 06, 5.1571e - 06]. \quad (22)$$

Подставляя значения (21) и (22) в уравнение регрессии, получаем

$$x(k) = mid \beta_0 + mid \beta_1 * k, \quad (23)$$

где k — номер измерения.

На Рис. 9 приведён коридор совместных зависимостей для модели (23). Визуально видно, что внутри коридора совместных зависимостей можно провести множество прямых.

Построение прогноза внутри и вне области данных. Одним из способов использования регрессионной модели является предсказание значений выходной переменной для заданных значений входной. С помощью построенной выше модели (23) можно получить прогнозные значения выходной переменной в точках эксперимента.

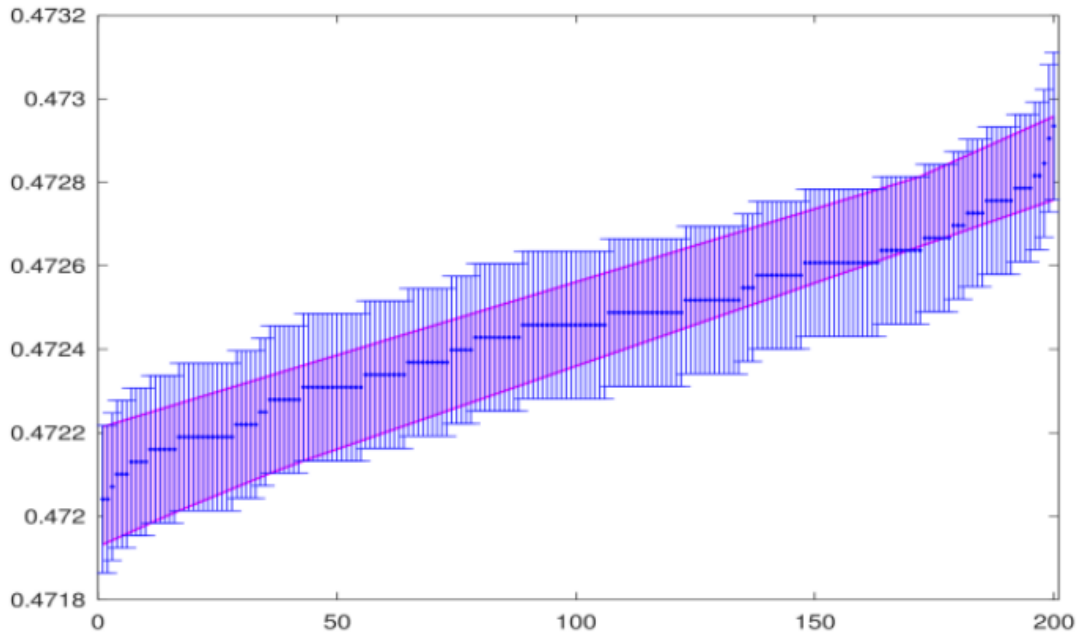


Рис. 9: Коридор совместных зависимостей (23).

Ценность модели также заключается в возможности её употребления для предсказания выходной переменной в точках, где измерения не производились.

Расширив область определения аргумента для модели (23), можно получить оценки для значений выходной переменной (экстраполяция). На Рис. 10 сплошной заливкой дан прогноз в том числе за пределами данных интервальной выборки X_1 .

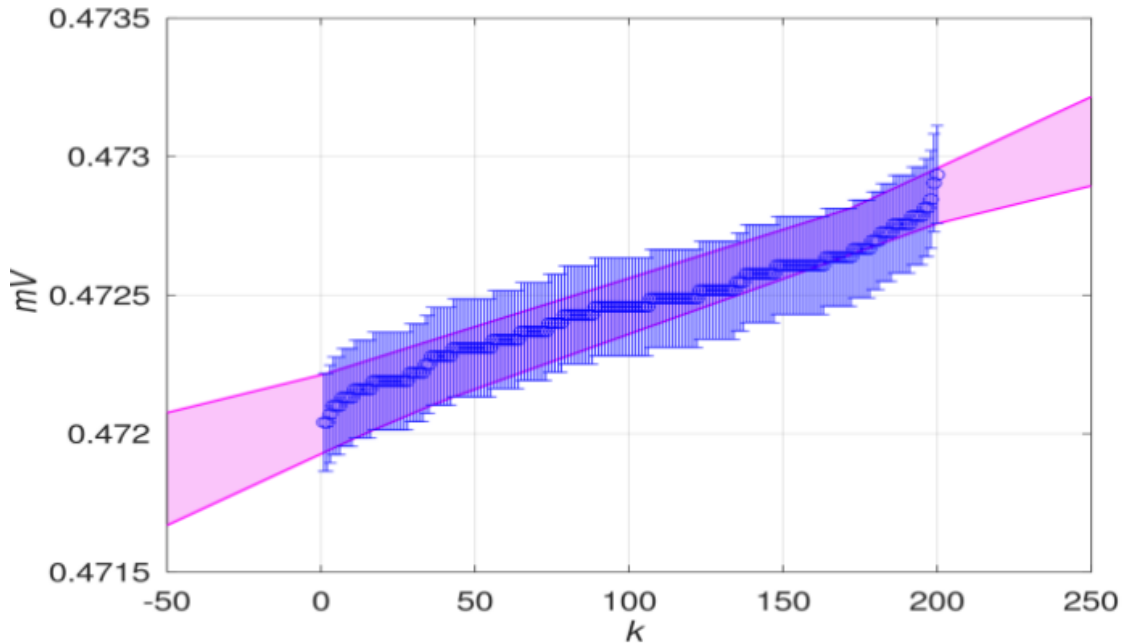


Рис. 10: Коридор совместных зависимостей (23). Построение прогноза.

Следует обратить внимание, что величина неопределённости прогнозов растёт по мере удаления от области, в которой производились исходные измерения. Это обусловлено видом коридора зависимостей, расширяющимся за пределами области измерений, и согласуется со здравым смыслом.

Уточнение структуры модели. Кусочно-линейная регрессионная зависимость.

Рис. 5 и Рис. 6 регрессионных остатков свидетельствуют о том, что линейные регрессионные модели не вполне точно отражают характер зависимости для интервальной выборки X_1 . Наиболее простым способом учёта этого факта является использование кусочно-линейная регрессионной зависимости.

В разделе «Варьирование неопределённости измерений» были вычислены векторы весов ω расширения неопределённости измерений для достижения совместности — см. Рис. 4. Резкое возрастание весов ω на границах области определения свидетельствует о несоответствии данных и модели. Эти точки и можно взять как «угловые» для определения линейных участков.

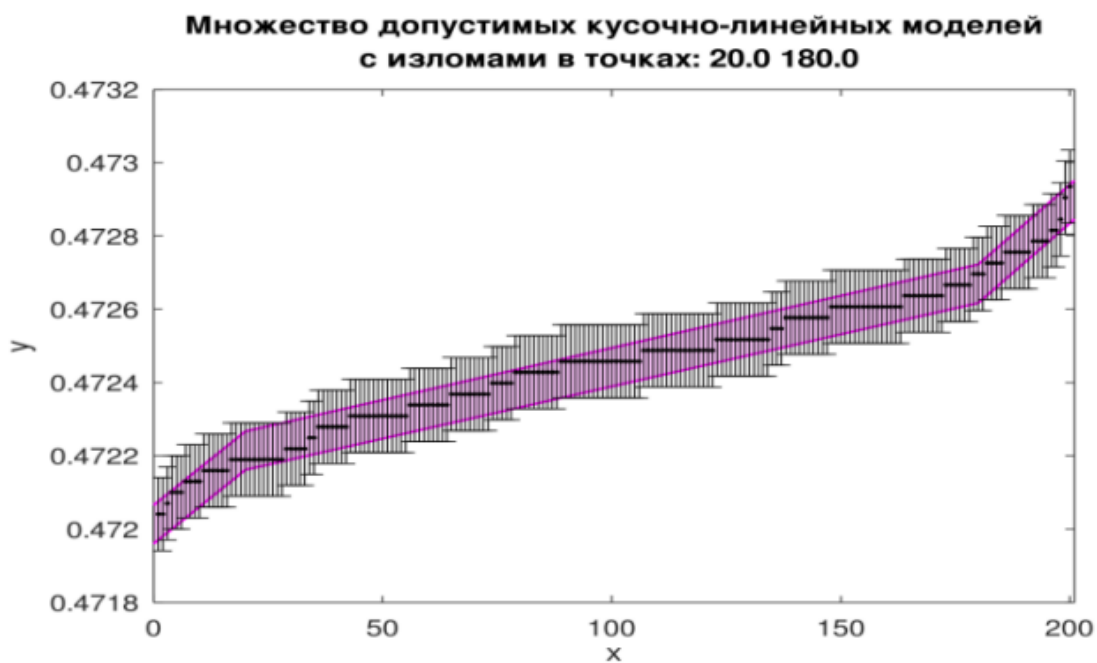


Рис. 11: Кусочно-линейная регрессионная зависимость.

На Рис. 11 показан пример построения кусочно-линейная регрессионной зависимости и коридора совместных зависимостей. После вычитания модели, можно переходить к анализу остатков регрессии и другим приёмам анализа.

В более общей постановке ставится задача автоматического определения точек излома [29], [30]. Имеется программное обеспечение С.И.Жилина, реализующее идеи этого подхода.

3 Результаты

3.1 Диаграмма рассеяния

Имеется выборка данных \mathbf{X}_1 с интервальной неопределенностью. Число отсчетов в выборке равно 200. Данные для выборки взяты из файла Channel_1_600nm_0.23mm.csv, погрешность прибора $\epsilon = 10^{-4}$.

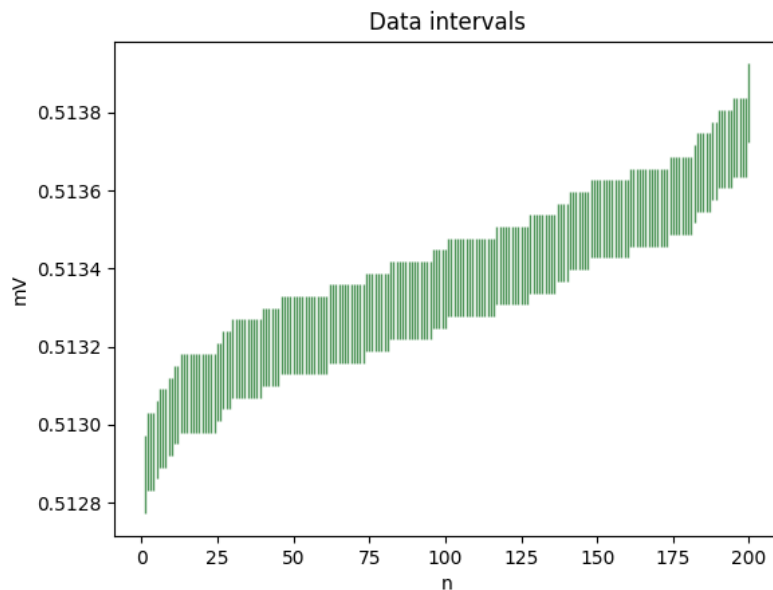


Рис. 12: Диаграмма рассеяния выборки \mathbf{X}_1 с уравновешенным интервалом погрешности

3.2 Варьирование неопределенности измерений

Проведем вычисление параметров линейной регрессии по данным интервальной выборки \mathbf{X}_1 с использованием программ С.И.Жилина [8] и оформленных применительно к задаче на [23], используя программу (6).

На Рис.13 красным цветом приведена регрессионная прямая.

Вычисления с использованием программы (6) дают следующие результаты для регрессионных коэффициентов

$$\beta_0 = \text{tau}(1) = 5.1295e - 01$$

$$\beta_1 = \text{tau}(2) = 4.0874e - 06$$

Все компоненты вектора ω оказались равными 1, то есть расширение интервалов измерений не понадобилось. Таким образом, величина (4) равна числу элементов выборки.

$$\sum_{i=1}^n \omega_i = 200$$

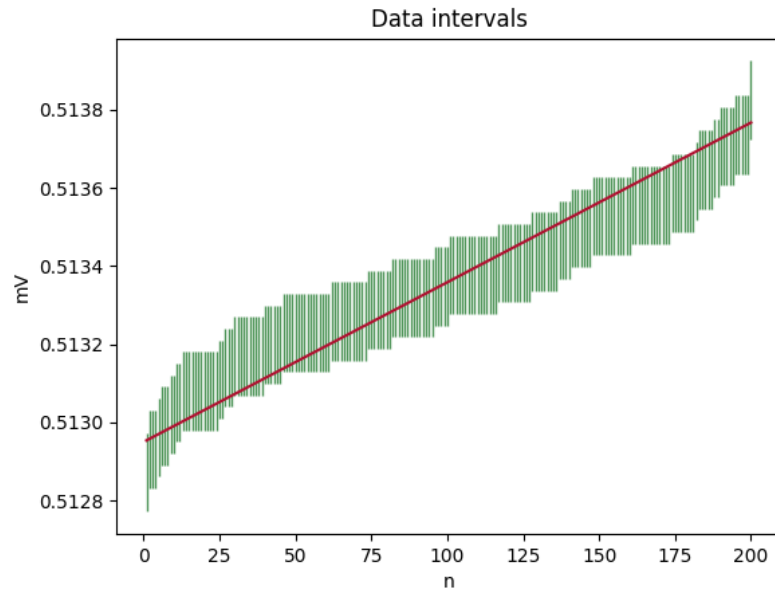


Рис. 13: Диаграмма рассеяния выборки \mathbf{X}_1 и регрессионная прямая по модели (4) и (5)

3.3 Варьирование неопределённости измерений с расширением и сужением интервалов

Вычисление параметров линейной регрессии по данным интервальной выборки \mathbf{X}_1 производится как и в случае (6) с использованием программ С.И.Жилина [8] и оформленных применительно к задаче на [23]. Синтаксис вызова программы - (12)

На Рис.14 красным цветом приведена регрессионная прямая.

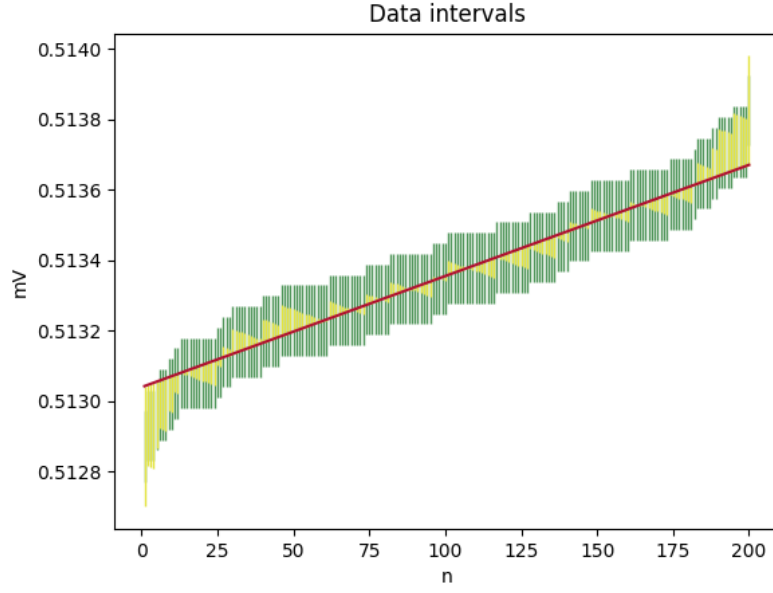


Рис. 14: Диаграмма рассеяния выборки \mathbf{X}_1 и регрессионная прямая по модели (10) и (11)

Желтым цветом на Рис.14 показаны скорректированные интервалы выборки \mathbf{X}_1 .

$$\beta_0 = \tau(1) = 5.1304e - 01$$

$$\beta_1 = \tau(2) = 3.1553e - 06$$

Величина меры (4) уменьшилась

$$\sum_{i=1}^n \omega_i = 43.733 < 200$$

На Рис.15 приведены графики векторов ω_0 и ω_1 , полученных при использовании двух рассмотренных подходов.

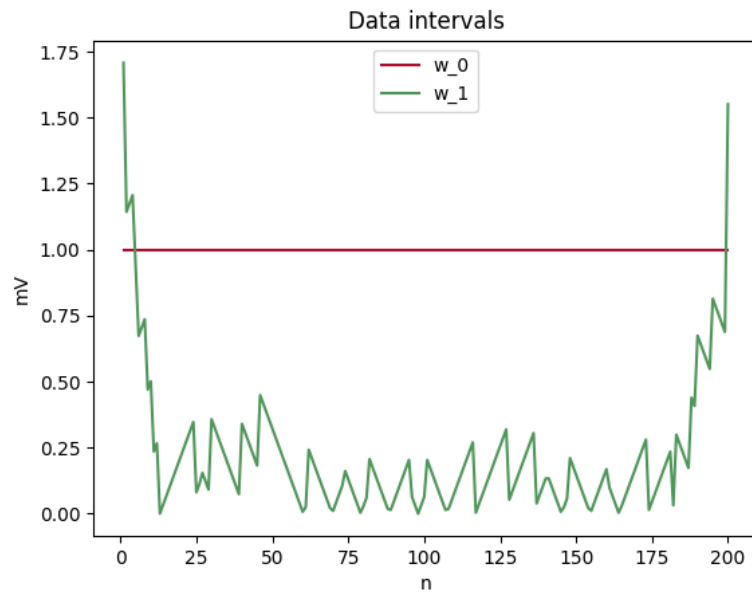


Рис. 15: Векторы ω_0 и ω_1

3.4 Анализ регрессионных остатков

На Рис. 16 приведена диаграмма рассеяния регрессионных остатков выборки \mathbf{X}_1 по модели (4) и (5).

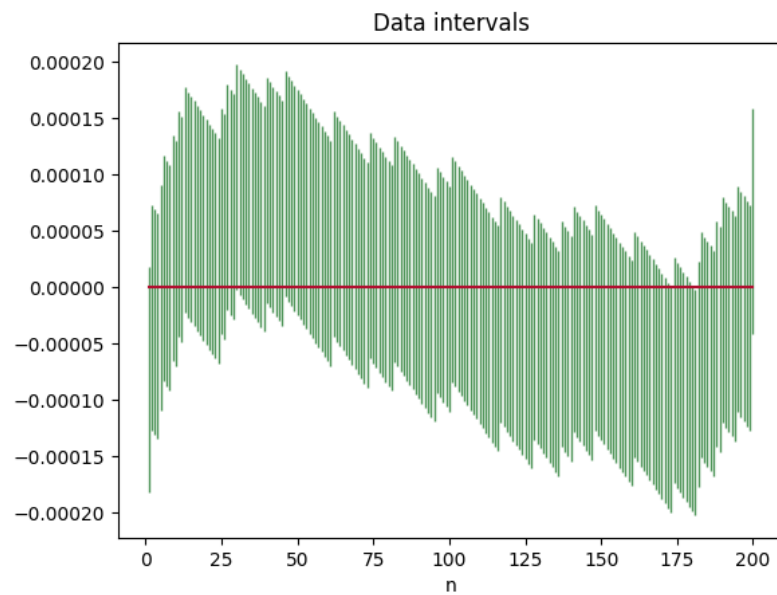


Рис. 16: Диаграмма рассеяния по модели (4) и (5)

На Рис. 17 приведена диаграмма рассеяния регрессионных остатков выборки \mathbf{X}_1 по модели (10) и (11).

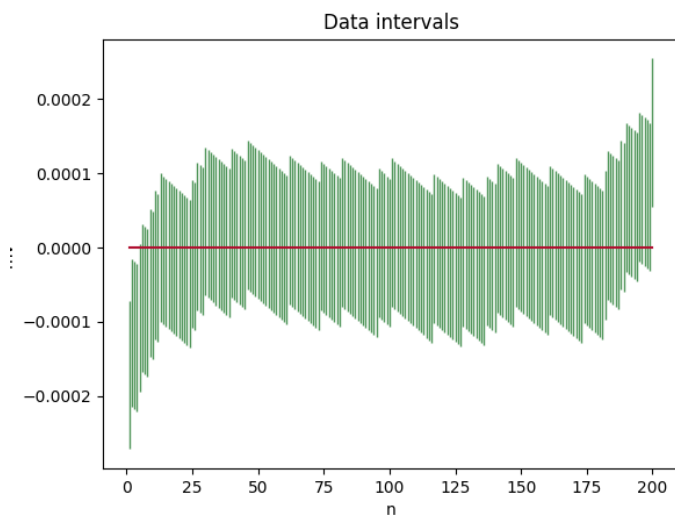


Рис. 17: Диаграмма рассеяния регрессионных остатков выборки \mathbf{X}_1 по (10) и (11)

На Рис.18 приведены графики частот элементарных подинтервалов при вычислении интервальной моды для двух моделей

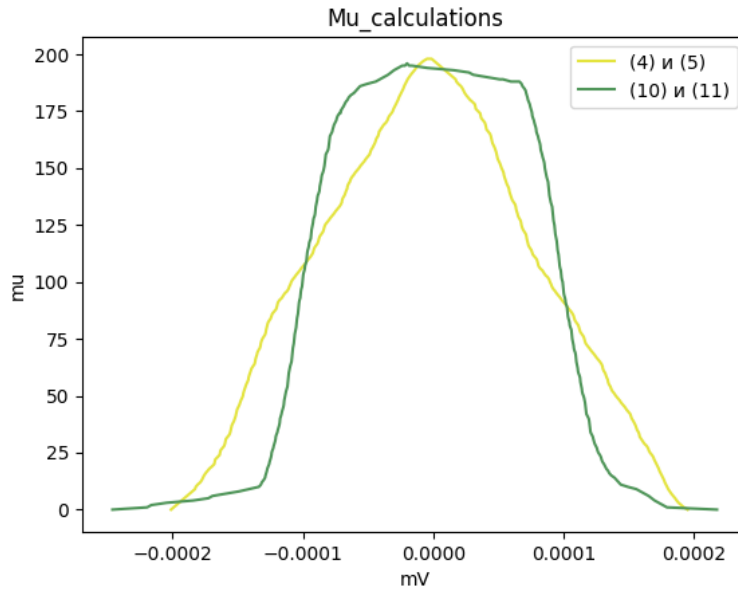


Рис. 18: Частоты элементарных подинтервалов регрессионных остатков выборки X_1 по моделям (4), (5) и (10), (11)

3.5 Информационное множество задачи

В совокупности интервальные оценки параметров задают брус, описанный вокруг информационного множества и именуемый внешней интервальной оболочкой информационного множества:

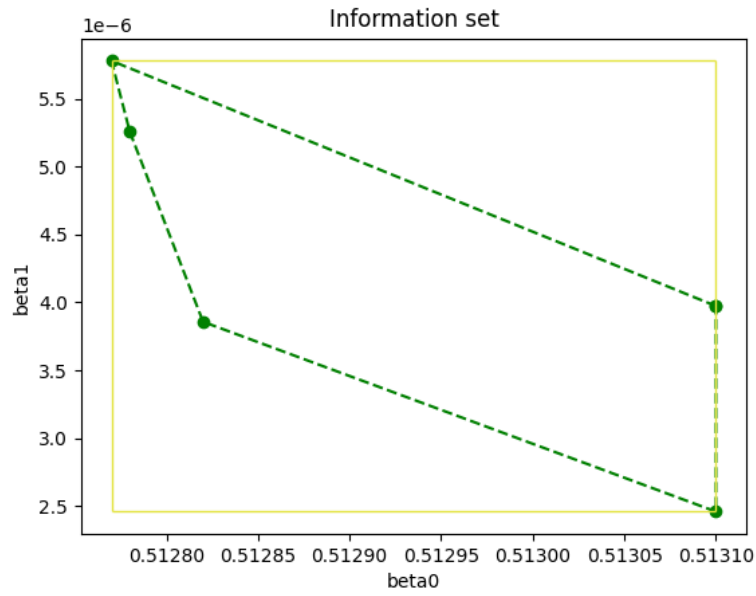


Рис. 19: Информационное множество по модели (10) и (11), интервальная оболочка — желтый брус

3.6 Коридор совместных зависимостей

$$\beta_0 = [5.1277e - 01, 5.131e - 01]$$

$$\beta_1 = [2.4572e - 06, 5.7817e - 06]$$

Подставляя значения (21) и (22) в уравнение регрессии, получаем

$$x(k) = \beta_0 + \beta_1 k,$$

где k — номер измерения.

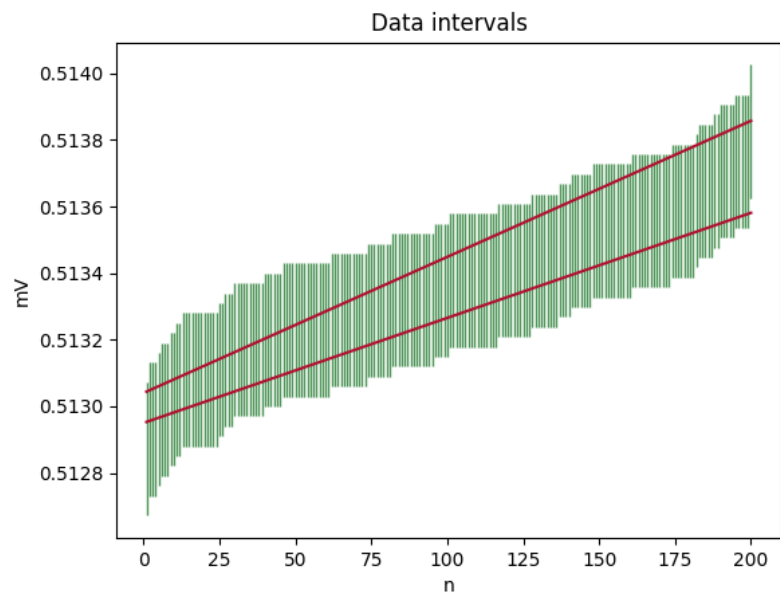


Рис. 20: Коридор совместных зависимостей (23)

3.7 Построение прогноза внутри и вне области данных

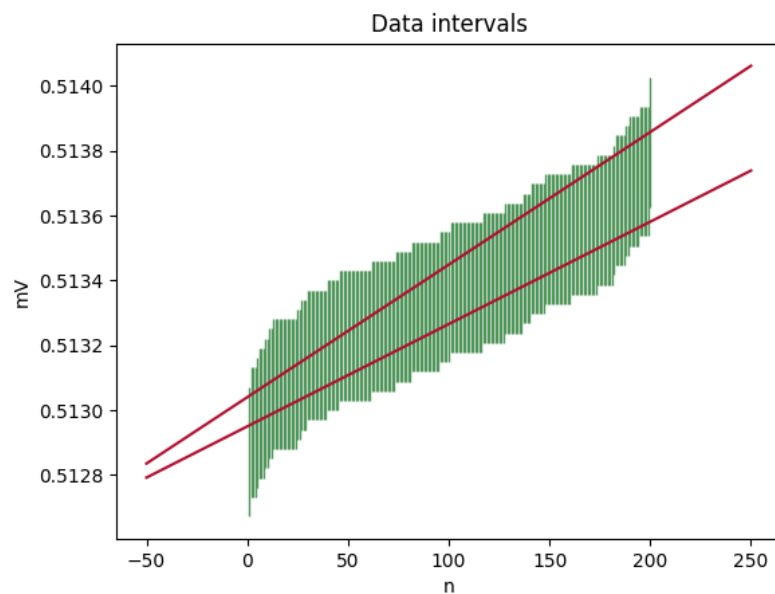


Рис. 21: Коридор совместных зависимостей (23). Построение прогноза.

4 Обсуждение

4.1 Варьирование неопределенности измерений

Для модели регрессии с $\omega_i \geq 1$ видим, что все $\omega_i = 1$, а регрессионная прямая действительно пересекает каждый отрезок без необходимости увеличения какого-либо из них.

4.2 Варьирование неопределенности измерений с расширением и сужением интервалов

Для модели регрессии с $\omega_i \geq 0$ видим, что для большинства интервалов $\omega_i < 1$, однако в начале и конце имеются выбросы $\omega_i \approx 1.5$. Также из рисунка видно, что регрессионная прямая пересекает уже не все интервалы. Это объясняется тем, что некоторые из них были увеличены, и регрессионная прямая пересекает измененные интервалы (желтые), притом пересекая увеличенный интервал она вовсе не обязана пересечь исходный.

4.3 Анализ регрессионных остатков

Из сравнения Рис. 16 и Рис. 17 видно, что интервальные выборки остатков получились с весьма похожими свойствами.

4.4 Информационное множество задачи

Внешняя интервальная оценка параметра определяется минимальным и максимальным значениями, которых может достигать значение параметра в информационном множестве. В совокупности интервальные оценки параметров задают брус, описанный вокруг информационного множества и именуемый внешней интервальной оболочкой информационного множества.

4.5 Коридор совместных зависимостей

По результатам построения коридора совместных зависимостей получено, как нетрудно видеть, множество, любая прямая, лежащая в котором, будет являться совместной регрессионной зависимостью для данной интервальной выборки.

4.6 Построение прогноза внутри и вне области данных

Следует обратить внимание, что величина неопределённости прогнозов растёт по мере удаления от области, в которой производились исходные измерения. Это обусловлено видом коридора зависимостей, расширяющимся за пределами области измерений, и согласуется со здравым смыслом.

5 Реализация

Лабораторная работа выполнена на языке Python версии 3.8.10 в среде разработки Google Colab. Использовались дополнительные библиотеки:

1. csv
2. scipy
3. numpy
4. matplotlib
5. math

Также использовался пакет GNU Octave 8.2.0. Использованы сторонние функции:

- <https://github.com/AlexanderBazhenov/Solar-Data>
- <https://github.com/szhilin/octave-interval-examples>

6 Приложение

Код программы GitHub: <https://github.com/KsenErem/MatStat>