Санкт-Петербургский политехнический университет имени Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики
Прикладной Математики и Информатики

Отчет
по лабораторным работам №10
по дисциплине
«Математическая статистика»

Выполнил:

Студент: КУАССИ СИЕМО ТИТ ГЕДЕОН
Группа: 5030102/00101

Принял:

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

Санкт-Петербург
2023 г.

Содержание

Список иллюстраций

1 Постановка задачи

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

$$y = f(\beta, x) \tag{1}$$

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

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

Если функция f является линейной, то можно записать

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m \tag{2}$$

В общем случае результаты измерений величин $x_1, x_2, ..., x_m$ и y являются интервальнозначными

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

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]:

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

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

2 Теория

Данные выборки. Имеется выборка данных $\mathbf{X}_1$ с интервальной неопределённостью. Число отсчётов в выборке равно 200.

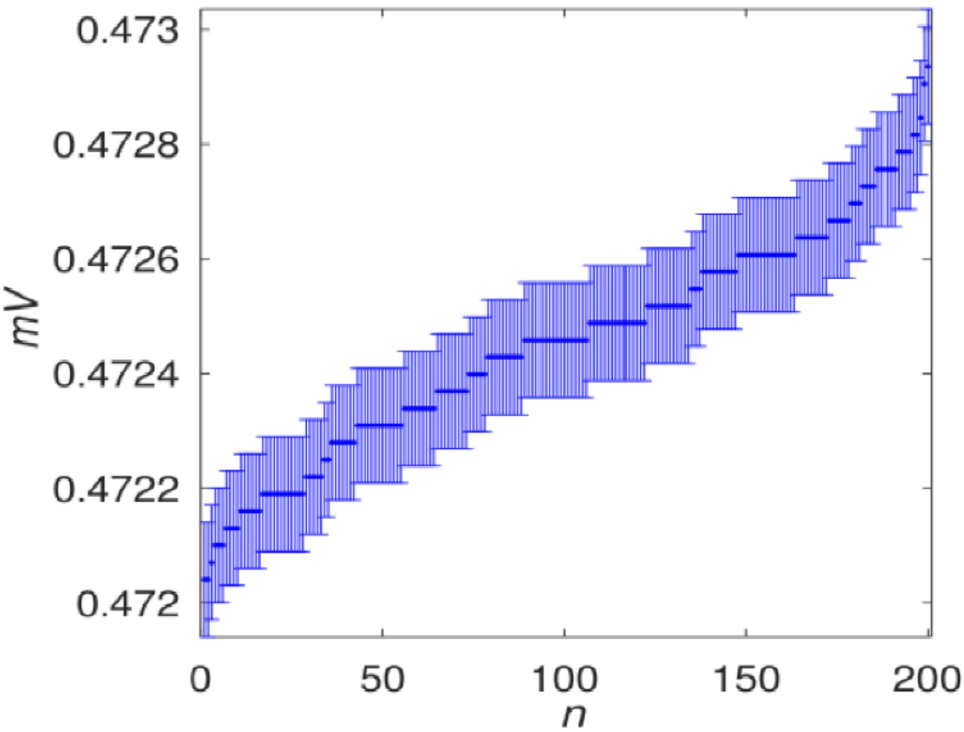

Рис. 1: Диаграмма рассеяния выборки X_1 с уравновешенным интервалом погрешности

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 $\omega = (\omega_1, ..., \omega_n)$, необходимая для совместности задачи построения зависимости $x = \beta_0 + \beta_1 * i$ может быть найдена решением задачи условной оптимизации

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

при ограничениях

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

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

$$[tau, w, yint] = DataLinearModel(input1, epsilon0) \tag{6}$$

В (6) входами программы служат значения $mid\mathbf{X}_1$ и величин неопределённости ϵ, а выходами tau — значения параметров регресии 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

$$\beta_0 = tau(1) = 4.7203e - 01, \tag{7}$$

$$\beta_1 = tau(2) = 4.0915e - 06. \tag{8}$$

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

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

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

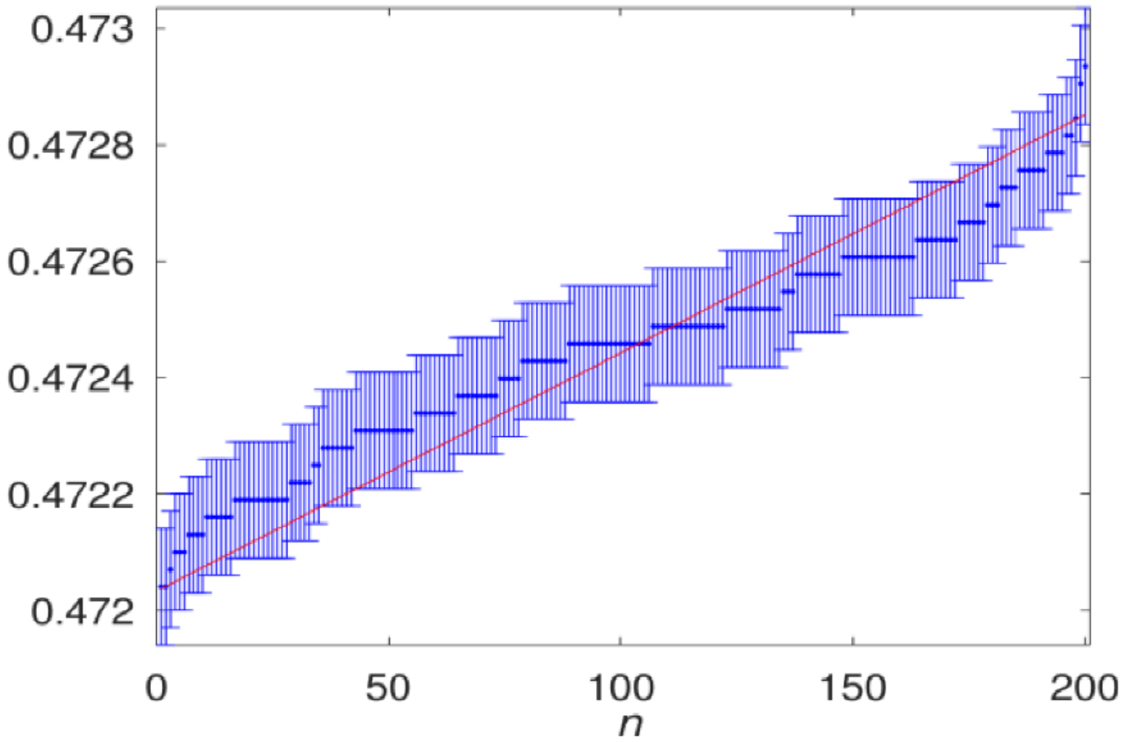

Рис. 2: Диаграмма рассеяния выборки X_1 и регрессионная прямая по модели (4) и (5)

(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:

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

при ограничениях

$$\begin{cases} mid\, x_i - \omega_i \epsilon_i \le \beta_0 + \beta_1 * i \le mid\, x_i + \omega_i \epsilon_i, \\ \omega_i \ge 0, \end{cases} \quad i = 1, ..., n. \tag{11}$$

Отличие постановки от (4) и (5) состоит в том, что интервалы измерений могут как расширяться в случае $\omega_i \ge 1$, так и сужаться при $0 \le \omega_i \le 1$. 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

$$[tau, w, yint] = DataLinearModelZ(input1, epsilon0) \tag{12}$$

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

На Рис. 3 красным цветом приведена регрессионная прямая.

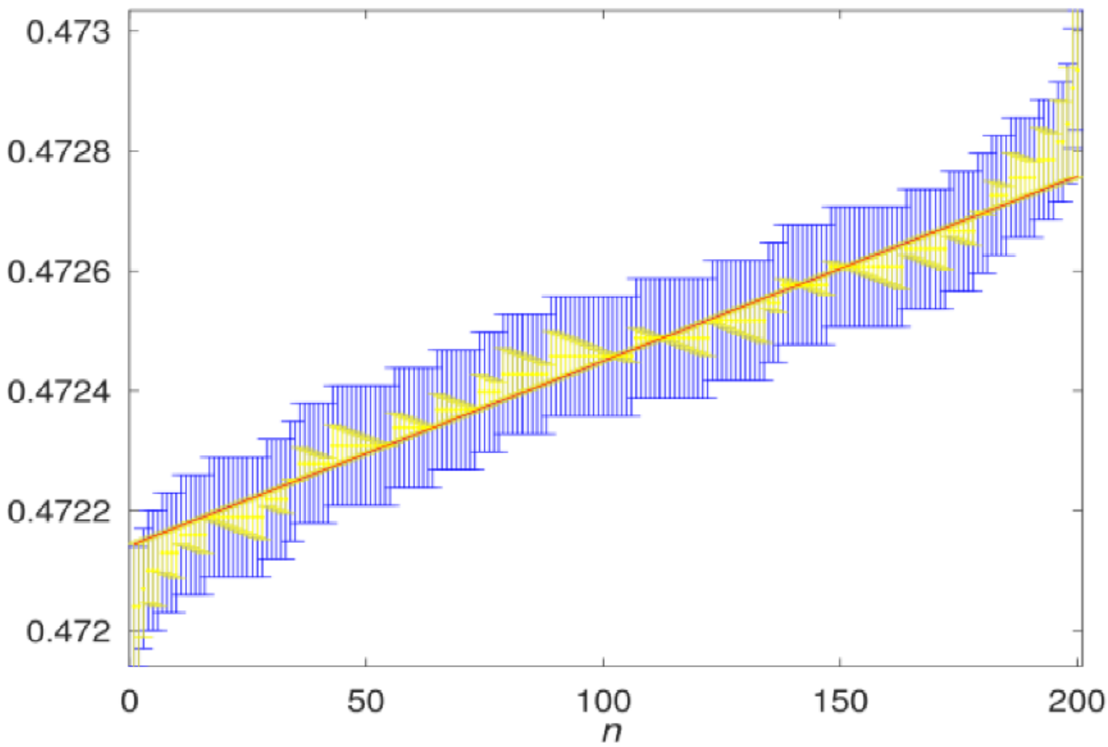

Рис. 3: Диаграмма рассеяния выборки X_1 и регрессионная прямая по модели (10) и (11)

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.

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

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 соотвествует исходным данным. Например:

$$\Psi = \arg_i \omega_i \geq \alpha \tag{14}$$

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

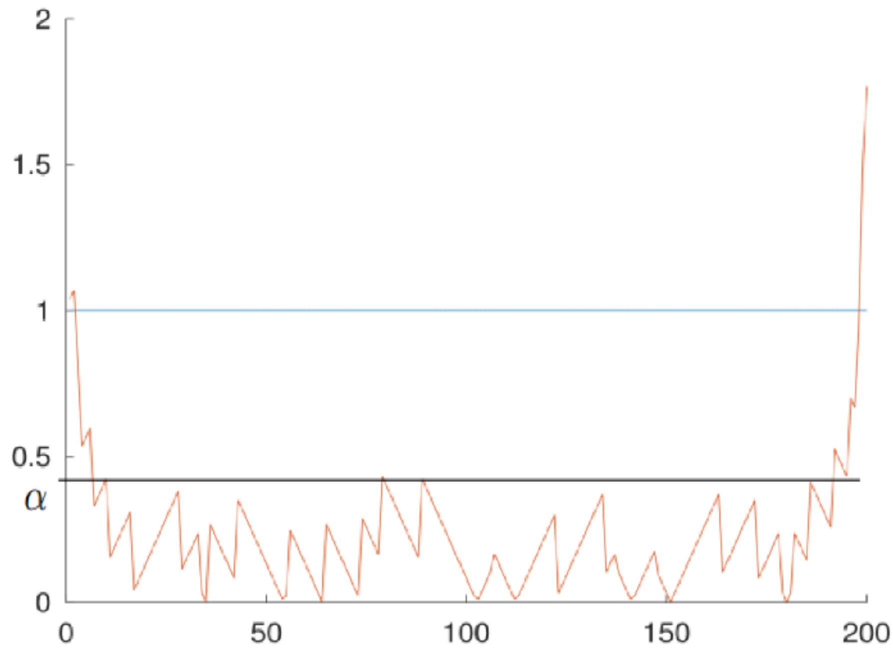

Рис. 4: Векторы ω_1 и ω_0

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

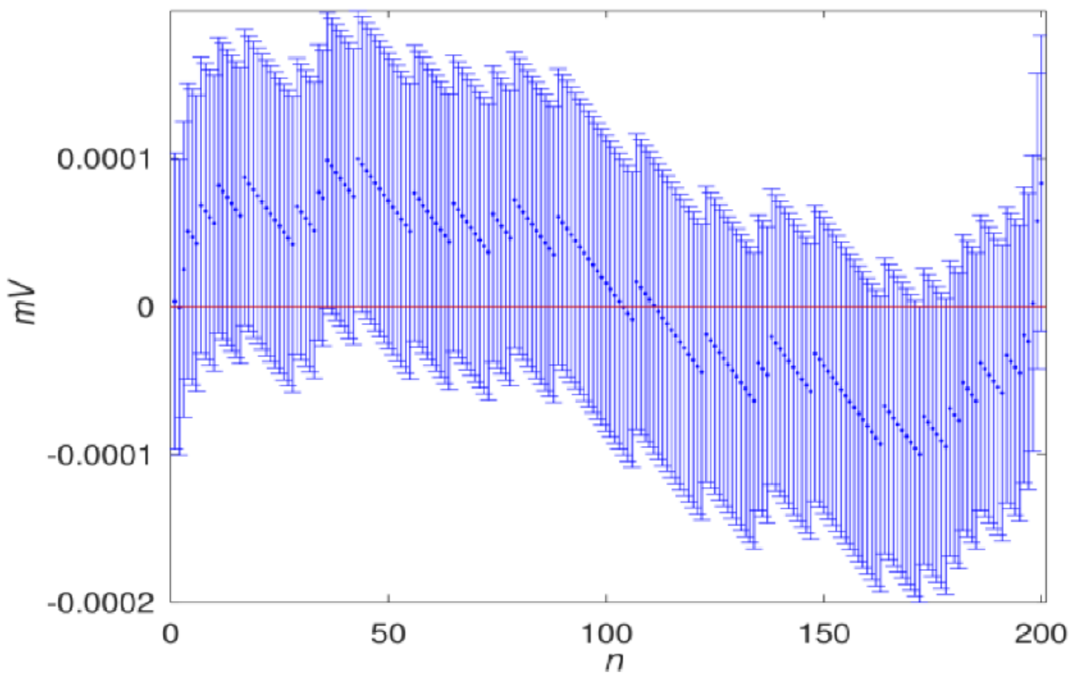

Рис. 5: Диаграмма рассеяния по модели (4) и (5)

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

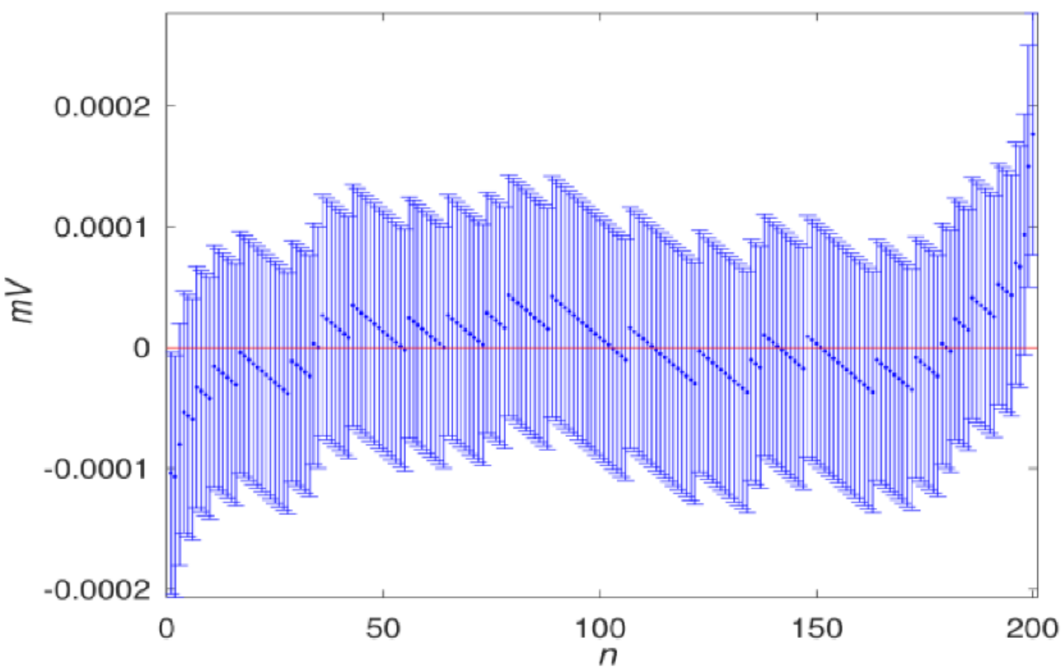

Рис. 6: Диаграмма рассеяния по модели (10) и (11)

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.

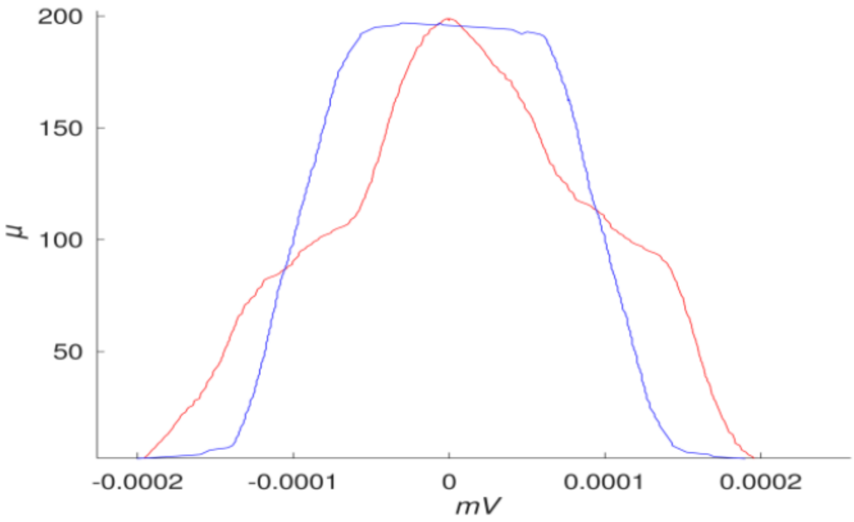

Рис. 7: Частоты элементарных подинтервалов регрессионных остатков выборки X_1 по модели (4) и (5) — красный график, и (10) и (11) — синий график

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

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

$$mode\,\mathbf{X}^1 = ... \tag{15}$$

$$Ji(\mathbf{X})^1 = ... \tag{16}$$

$$\vdots \tag{17}$$

$$mode\,\mathbf{X}^2 = ... \tag{18}$$

$$Ji(\mathbf{X})^2 = ... \tag{19}$$

$$\vdots \tag{20}$$

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 разниться.

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:

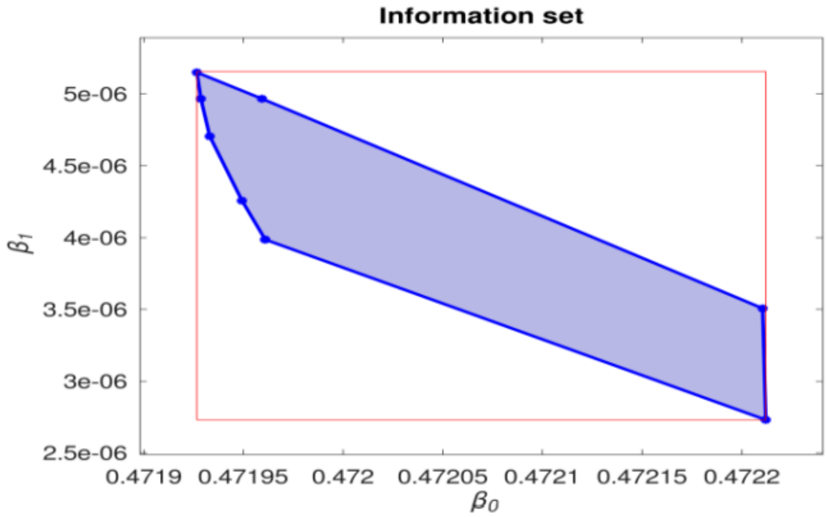

Рис. 8: Информационное множество по модели (10) и (11), интервальная оболочка — красный брус

Проведём вычисление параметров линейной регрессии по данным интервальной выборки $\mathbf{X}_1$ 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 регресии.

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

$$mid\,\beta_0 = [4.7193e - 01, 4.7221e - 01], \tag{21}$$

$$mid\,\beta_1 = [2.7304e - 06, 5.1571e - 06]. \tag{22}$$

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

$$x(k) = mid\,\beta_0 + mid\,\beta_1 * k, \tag{23}$$

где k — номер измерения.

На Рис. 9 приведён коридор совместных зависимостей для модели (23). Визуально видно, что внутри коридор 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.

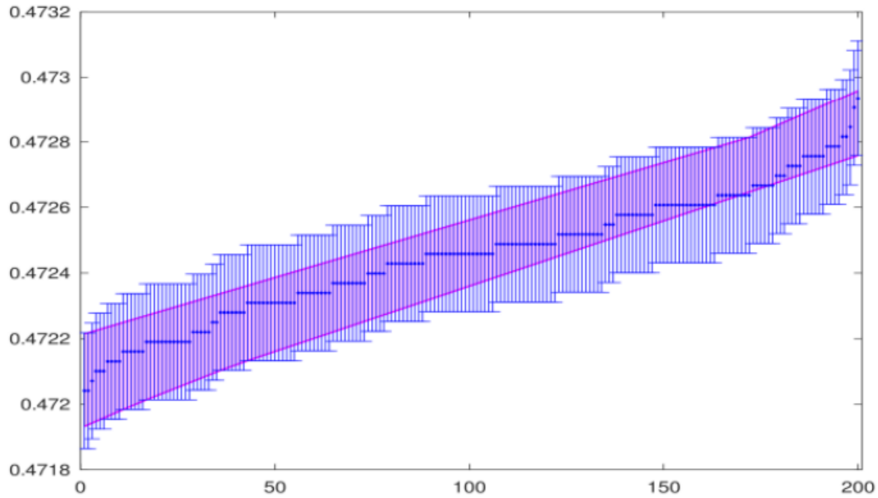

Рис. 9: Коридор совместных зависимостей (23).

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 $\mathbf{X}_1$.

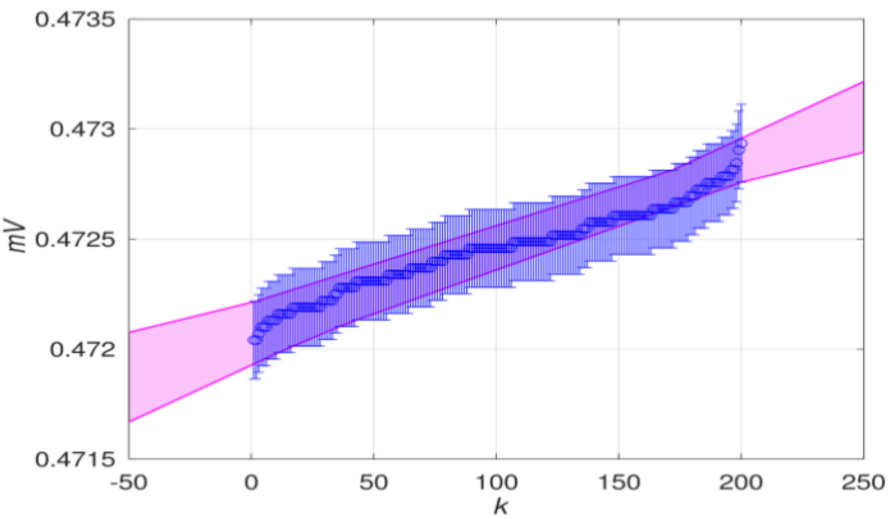

Рис. 10: Коридор совместных зависимостей (23). Построение прогноза

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 регресионных остатков свидетельствуют о том, что линейные регрессионные модели не вполне точно отражают характер зависимости для интервальной выборки $\mathbf{X}_1$. 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.

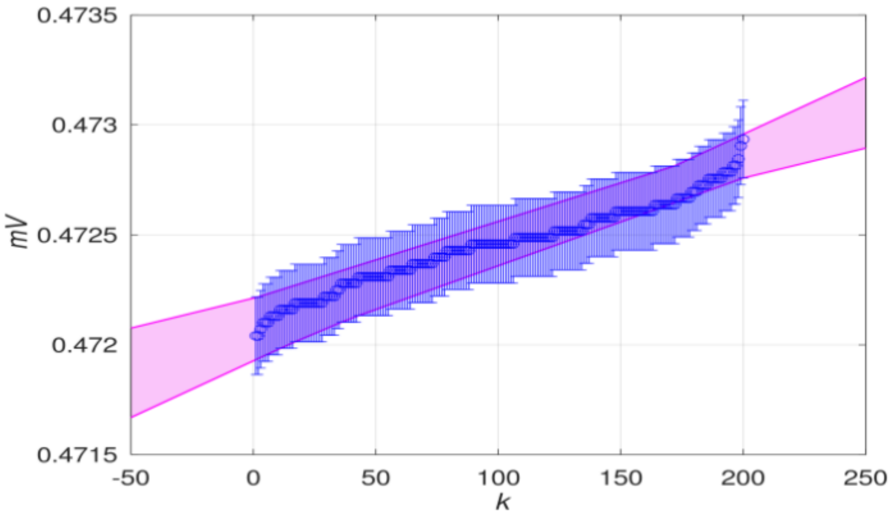

Рис. 11: Кусочно-линейная регрессионная зависимость

На Рис. 11 показан пример построения кусочно-линейная регрессионной зависимости и коридора совместных зависимостей. После вычитания модели, можно переходить к анализу отстатков регресии 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 Реализация

3.1 Описание

Данная лабораторная работа была выполнена с использованием языка программирования Python с использованием следующих библиотек:

- math - использование математических функций
- matplotlib версии 3.7.1 - построение графиков
- numpy версии 1.24.2 - использование многомерных массивов
- prettytable версии 3.6.0 - вывод таблиц в консоли
- scipy версии 1.10.1 - статические распределения и функции
- seaborn версии 0.12.2 - посроение графиков, визуализация
- statsmodels - дополнение к scipy, использование статистических вычислений, включая описательную статистику, оценку и вывод статистических моделей

Отчёт подготовлен с помощью языка LaTEX в редакторе Overleaf.

3.2 Ссылка на репозиторий

https://github.com/STGK10/Mathstatistics2023-Labs/tree/main - GitHub репозиторий

4 Результаты

4.1 Данные выборки

Данные для выборки взяты из файла $Channel\_1\_400nm\_2mm.csv$, $\varepsilon = 10^{-4}$

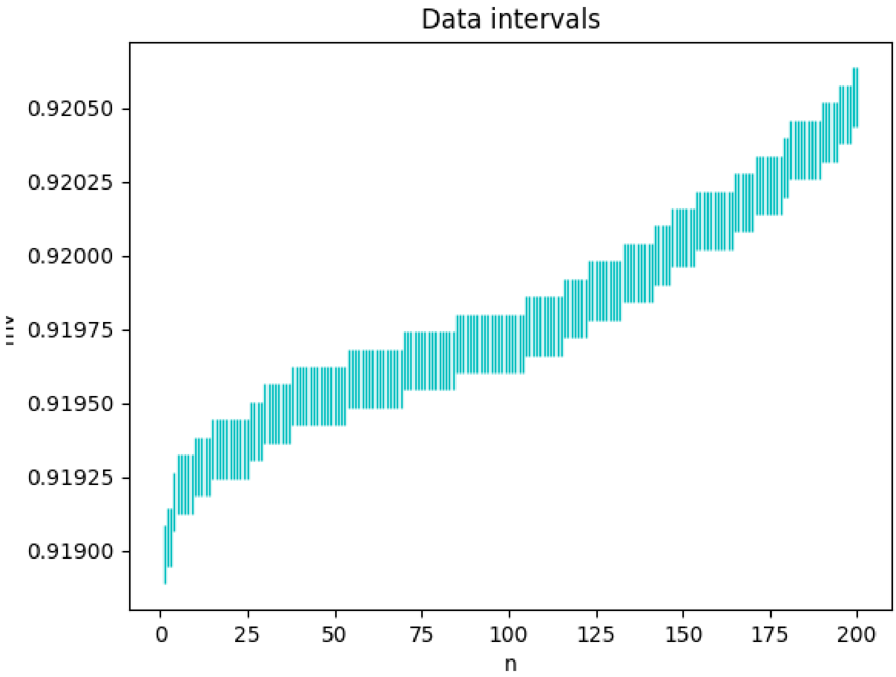

Рис. 12: Диаграмма рассеяния выборки X_1 с уравновешенным интервалом погрешности

4.2 Варьирование неопределённости измерений

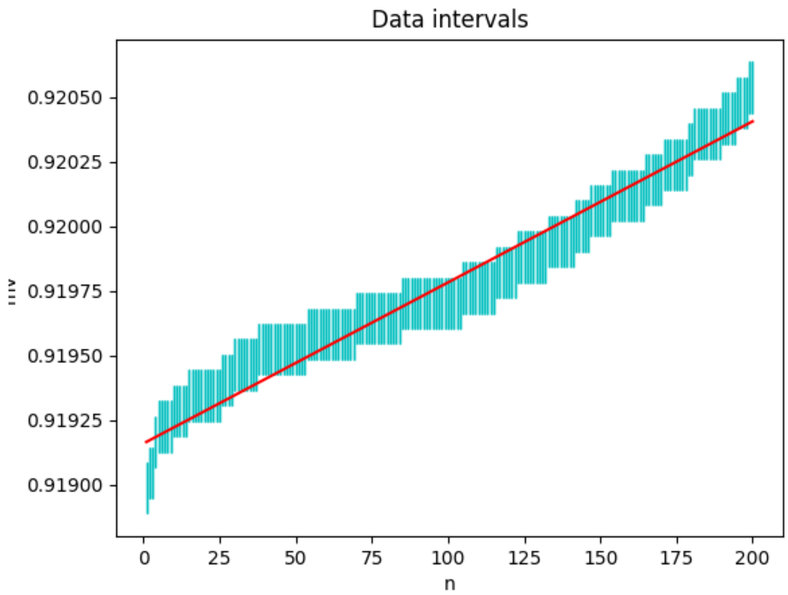

Рис. 13: Диаграмма рассеяния выборки X_1 и регрессионная прямая по модели (4) и (5)

$$\sum_{i=1}^{n} \omega_i = 200,\ \beta_0 = 0.91916, \beta_1 = 6.2333 \cdot 10^{-6}$$

4.3 Варьирование неопределённости измерений с расширением и сужением интервалов

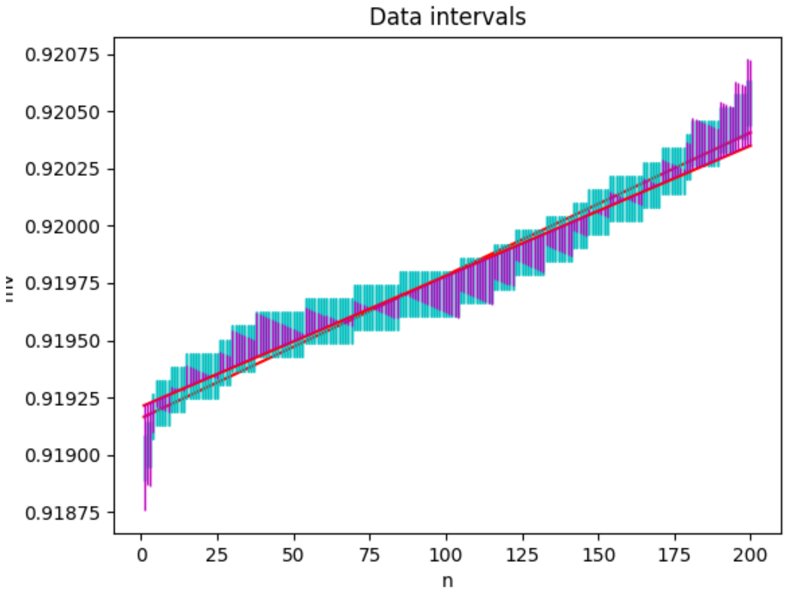

Рис. 14: Диаграмма рассеяния выборки X_1 и регрессионная прямая по модели (10) и (11)

$$\sum_{i=1}^{n} \omega_i = 98.559,\ \beta_0 = 0.91921, \beta_1 = 5.6971 \cdot 10^{-6}$$

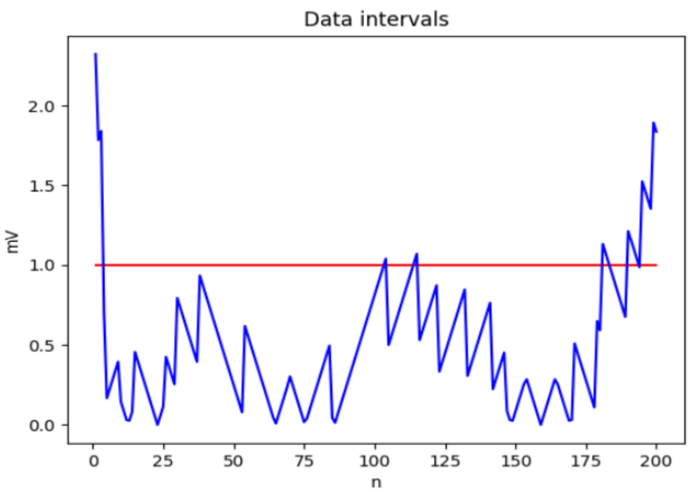

Рис. 15: Векторы ω_0 и ω_1

4.4 Анализ регресионных остатков

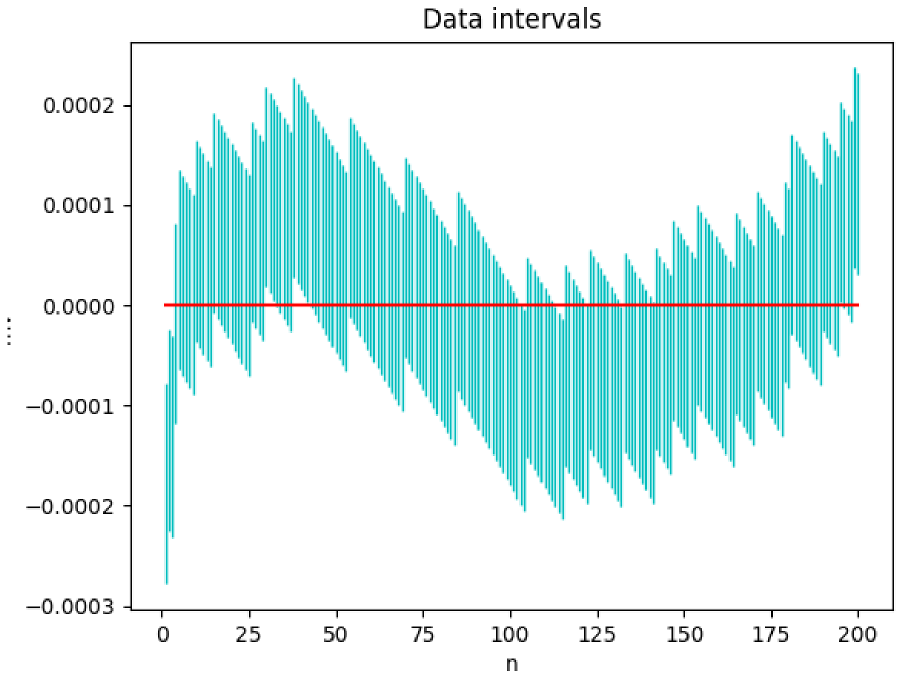

Рис. 16: Диаграмма рассеяния по модели (4) и (5)

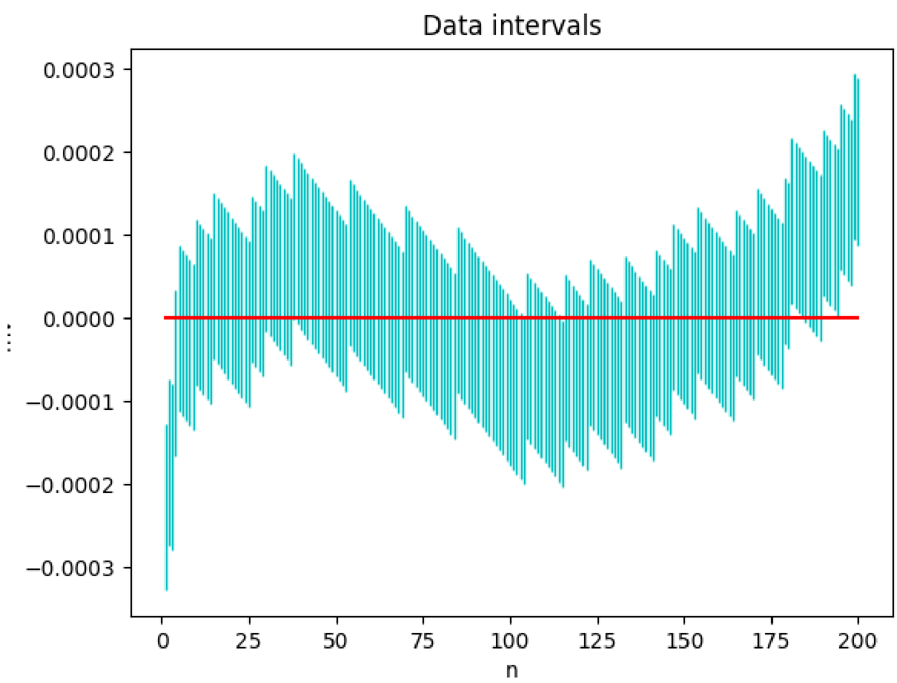

Рис. 17: Диаграмма рассеяния регрессионных остатков выборки X_1 по (10) и (11)

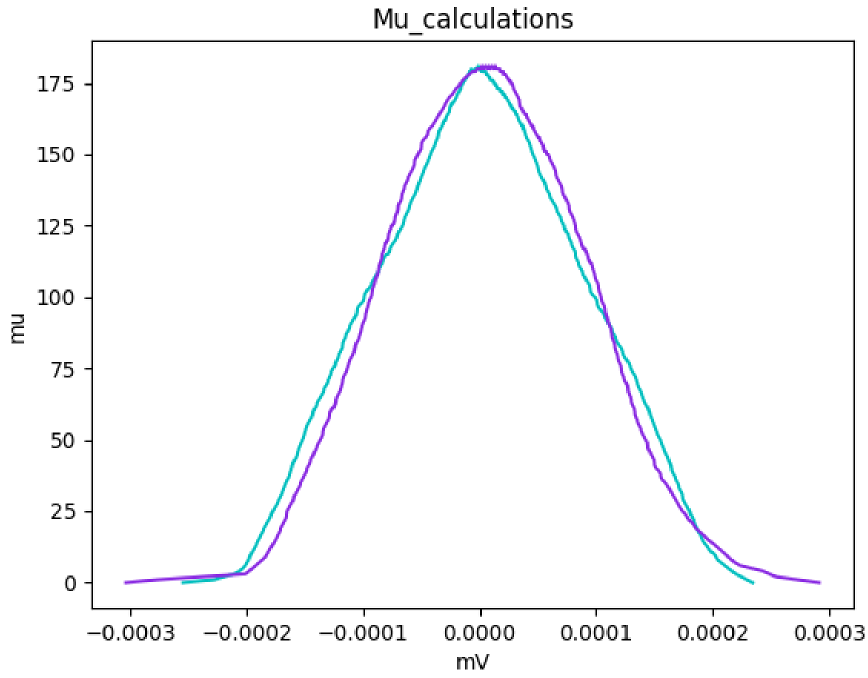

Рис. 18: Частоты элементарных подинтервалов регрессионных остатков выборки X_1 по модели (4) и (5) — синий график, и (10) и (11) — фиолетовый график.

Меры совместности регрессионных остатков: $mode X_0 = [-0.000007, 0.000007] J_i(X_0) = 0.5335$
$mode X_1 = [-0.000008, 0.000008] J_i(X_1) = 0.6808$
Здесь X_0, X_1 – регрессионные остатки выборки X_1, вычисленные с использованием разных условий оптимизации.

4.5 Информационное множество задачи

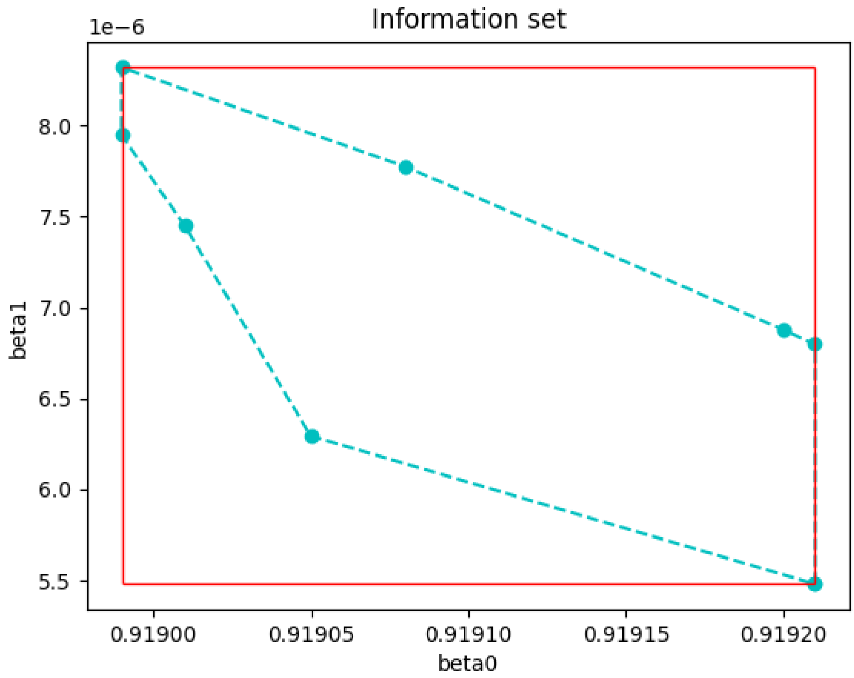

Рис. 19: Информационное множество по модели (10) и (11), интервальная оболочка — красный брус

4.6 Коридор совместных зависимостей

$$\operatorname{mid} \boldsymbol{\beta}_0 = [0.91899, 0.91921]$$
$$\operatorname{mid} \boldsymbol{\beta}_1 = [5.4802 \cdot 10^{-6}, 8.3358 \cdot 10^{-6}]$$

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

$$\operatorname{mid} x(k) = mid\beta_0 + mid\beta_1 \cdot k$$

где k — номер измерения.

4.7 Построение прогноза внутри и вне области данных

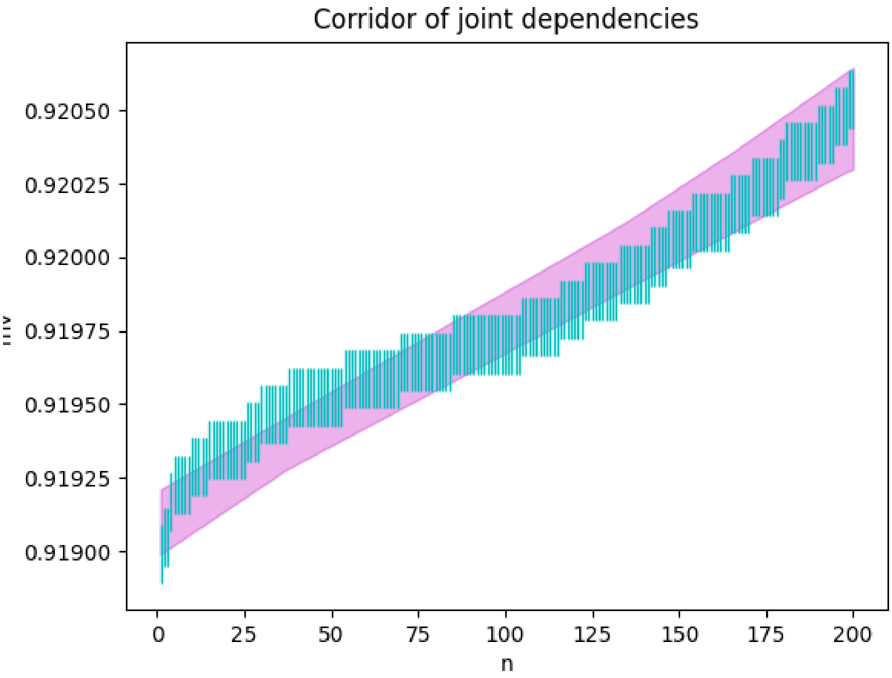

Рис. 20: Коридор совместных зависимостей (23)

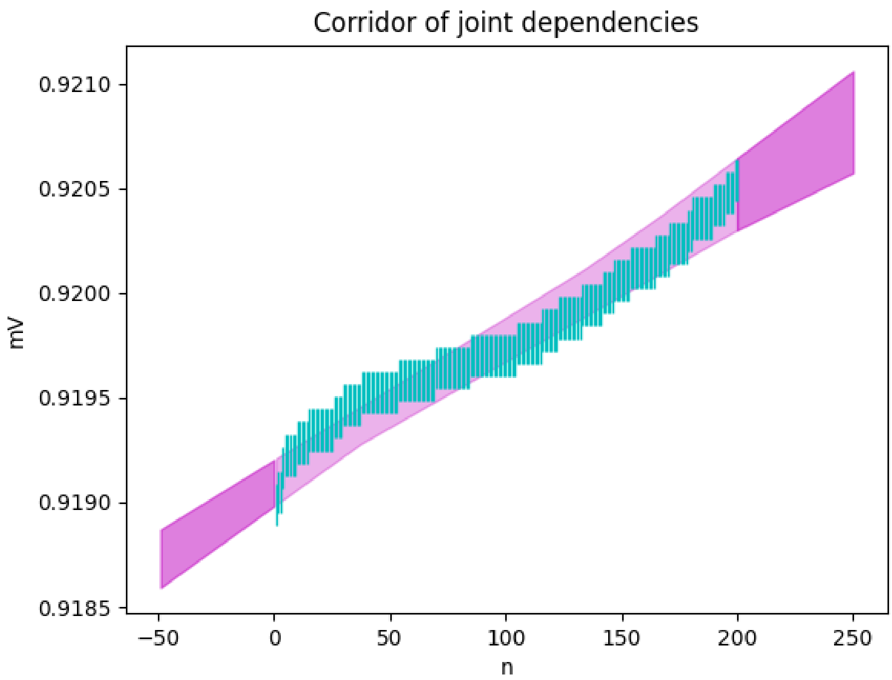

Рис. 21: Коридор совместных зависимостей (23). Построение прогноза

5 Обсуждение

5.1 Варьирование неопределённости измерений

Все компоненты вектора ω оказались равны 1, то есть, расширения интервалов измерений не понадобилось. Таким образом, величина (4) равна числу элементов выборки. Недостатком полученного решения с единичными значениями ω_i является неучёт расстояний точек регрессионной зависимости до данных интервальной выборки. Таким образом, прямая с параметрами (7) и (8) «не чувствует» отклонений измерений от прямой на концах выборки — неопределённости измерений достаточно велики, чтобы покрыть этот эффект.

5.2 Варьирование неопределённости измерений с расширением и сужением интервалов

Величина меры (4) уменьшилась более, чем в 2 раза. Таким образом, постановка задачи с возможностью одновременного увеличения и уменьшения радиусов неопределённости измерений позволяет более гибко подходить к задаче оптимизации.

5.3 Анализ регрессионных остатков

По результатам вычислений для регрессионных остатков можно сделать вывод, что мода регрессионных остатков по модели с $\omega_i \geq 0$ представляет собой более широкую окрестность нуля. Это означает, что регрессия по этой модели качественнее, нежели по модели $\omega_i \geq 1$.

5.4 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.

5.5 Коридор совместных зависимостей

На Рис. 20 приведён коридор совместных зависимостей для модели (2.54). Визуально видно, что внутри коридор совместных зависимостей можно провести множество прямых.

5.6 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.

Список литературы

[1] Histogram. URL: https://en.wikipedia.org/wiki/Histogram

[2] Вероятностные разделы математики. Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д. — Спб.: «Иван Федоров», 2001. — 592 с., илл.

[3] Box plot. URL: https://en.wikipedia.org/wiki/Box_plot

[4] Анатольев, Станислав (2009) «Непараметрическая регрессия», Квантиль, №7, стр. 37-52.

[5] М.З.Шварц. Данные технологических испытаний оборудования для калибровки фотоприемников солнечного излучения. 2022.