Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

Отчёт
по лабораторной работе №10
по дисциплине
«Математическая статистика»

Выполнил студент:
Желудев К.И.
группа: 5030102/00101

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

Санкт-Петербург
2023 г.

Содержание

Список иллюстраций

1 Постановка задачи

Имеется выборка данных с интервальной неопределенностью. Число отсчетов в выборке равно 200. Используется модель данных с уравновешенным интервалом погрешности.

$x = \overset{\circ}{x} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} = [-\epsilon, \epsilon]$ для некоторого $\epsilon > 0$,

Здесь $\overset{\circ}{x}$ – данные некоторого прибора, $\epsilon = 10^{-4}$ – погрешность прибора.
Необходимо [1]:

- Иллюстрировать данные выборки
- Построить диаграмму рассеяния
- Построить линейную регрессионную зависимость варьированием неопределенности изменений с расширением и без сужения интервалов
- Построить линейную регрессионную зависимость варьированием неопределенности изменений с расширением и сужением интервалов
- Произвести анализ регрессионных остатков
- Построить информационное множество по модели
- Проиллюстрирвоать коридор совместных зависимостей
- Построить прогноз вне области данных

Файл с данными интервальной выборки "Channel_1_700nm_0.2.csv" расположен по следующей ссылке: **исходные данные**
Данные взяты из архива, расположенного по следующей ссылке: **архив с данными интервальных выборок** (название использованного файла "Канал 1_700nm_0.2.csv")

2 Теория

2.1 Постановка задачи восстановления функциональной зависимости

Пусть некоторая величина y является функцией от независимых переменных $x_1, \ldots, x_m$:

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

Где $x = (x_1, \ldots, x_m)$ является вектором независимых переменных, $\beta = (\beta_1, \ldots, \beta_p)$ – вектор параметров функции. Заметим, что переменные $x_1, \ldots, x_m$ также называются входными, а переменная y – выходной.
Задача восстановления функциональной зависимости заключается в том, чтобы, располагая набором значений x и y, найти такие $\beta_1, \ldots, \beta_p$ в выражении (1), которые соответствуют конкретной функции f из параметрического семейства.
Если функция f является линейной, то можно записать

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

В общем случае результаты измерений величин $x_1, \ldots, x_m$ и y являются интервальнозначными

$$x_1^{(k)}, \ldots, x_m^{(k)}, y^{(k)}, k \in \overline{1..n}$$

n – число измерений.
Брусом неопределенности k - го измерения функциональной зависимости будем называть интервальный вектор-брус, образованный интервальными результатами измерений с одинаковыми значениями индекса k

$$(x_{k1}, \ldots, x_{km}, y_k) \subset R^{m+1} \tag{3}$$

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

2.2 Варьирование неопределенности изменений

Если величину коррекции каждого интервального наблюдения $y_i = [\overset{\circ}{y}_i - \epsilon_i, \overset{\circ}{y}_i + \epsilon_i]$ выборки S_n выражать коэффициентом его уширения $w_i \geq 1$, а общее изменение выборки характеризовать суммой этих коэффициентов, то минимальная коррекция выборки в виде вектора коэффициентов $w^* = (w_1^*, \ldots, w_n^*)$, необходимая для совместности задачи построения $y = f(x, \beta)$ может быть решена решением задачи условной оптимизации:

Найти:

$$\min_{w,\beta} \sum_{i=1}^{n} w_i \tag{4}$$

При ограничениях:

$$\begin{cases} \text{mid}x_i - w_i\epsilon_i \leq \beta_0 + \beta_i * i \leq \text{mid}x_i + w_i\epsilon_i, \\ w_i \geq 1 \end{cases} \tag{5}$$

$i \in \overline{1..n}$

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

2.3 Варьирование неопределенности изменений с расширением и сужением интервалов

Поставим задачу условной оптимизации следующим образом:
Найти:

$$\min_{w,\beta} \sum_{i=1}^{n} w_i \tag{6}$$

При ограничениях:

$$\begin{cases} \text{mid}x_i - w_i\epsilon_i \leq \beta_0 + \beta_i * i \leq \text{mid}x_i + w_i\epsilon_i, \\ w_i \geq 0 \end{cases} \tag{7}$$

$i \in \overline{1..n}$

Отличие постановки от (4) и (5) состоит в том, что интервалы измерений могут как расширяться в случае $w_i \geq 1$, так и сужаться при $0 \leq w_i < 1$

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

$$\Psi = \arg_i w_i \geq \alpha \tag{8}$$

Для объективного использования этого приема параметр α можно брать, например, из анализа гистограммы распределения веткора w.
Использование выделения «подозрительных» областей даёт основу для других приемов. Например, для построения кусочно-линейной регрессионной зависимости.

2.4 Анализ регрессионных остатков

В теоретико-вероятностной математической статистике анализ регрессионных остатков один из приемов оценки качества регрессии.
Приведем пример пояснения этого приема. «Если выбранная регрессионная модель хорошо описывает истинную зависимость, то остатки должны быть независимыми, нормально распределенными случайными величинами с нулевым средним, и в значениях должен отсутствовать

тренд. Анализ регрессионных остатков – это процесс проверки выполнения этих условий» [2][с. 658-659].
В случае интервальных выборок мы не задаемся вопросом о виде распределения остатков, а будем использовать те возможности, которые появляются при описании объектов и результатов вычислений в виде интервалов.

2.5 Информационное множество задачи

Информационным множеством задачи восстановления функциональной зависимости является множество всех значений параметров функции β, получаемое в результате обработки интервальных значений входных переменных x и выходной переменной y.

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 минимальными и максимальными 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.

3 Результаты

3.1 Данные выборки

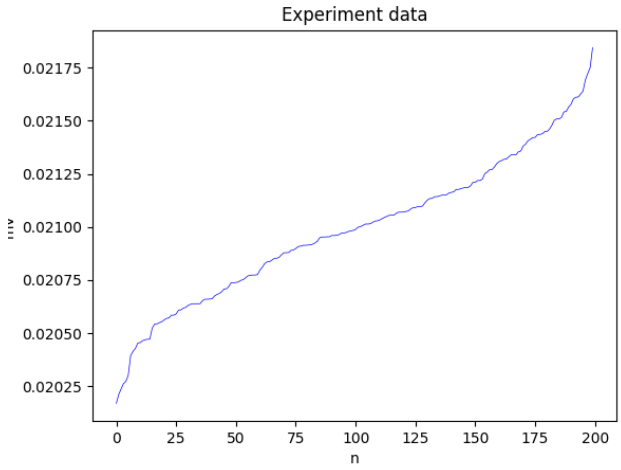

Рис. 1: Данные выборки X_1

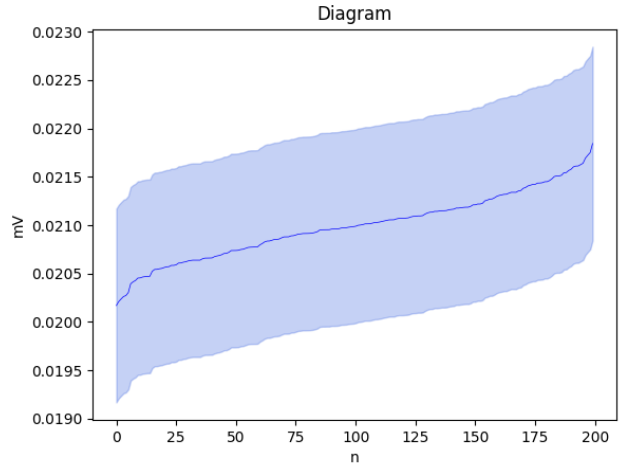

Рис. 2: Диаграмма рассеяния выборки X_1 с уравновешанным интервалом неопределенности

3.2 Варьирование неопределенности изменений

При решении задачи линейного программирования (4), (5) были получены следующие результаты:

$$\beta_0 = 0.020, \beta_1 = 6.504 \cdot 10^{-6}$$

$$w_1 = (w_1^1, \ldots, w_1^n), \sum_{i=1}^{n} w_1^i = 202.831$$

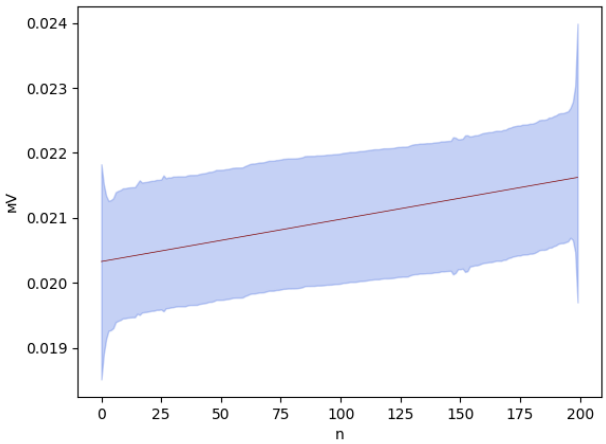

Рис. 3: Диаграмма рассеяния выборки X_1 и регрессионная прямая по модели

Красным цветом обозначена регрессионная прямая модели.

3.3 Варьирование неопределенности изменений с расширением и сужением интервалов

При решении задачи линейного программирования (6), (7) были получены следующие результаты:

$$\beta_0 = 0.020, \beta_1 = 5.354 \cdot 10^{-6}$$

$$w_0 = (w_0^1, \ldots, w_0^n), \sum_{i=1}^{n} w_0^i = 77.636$$

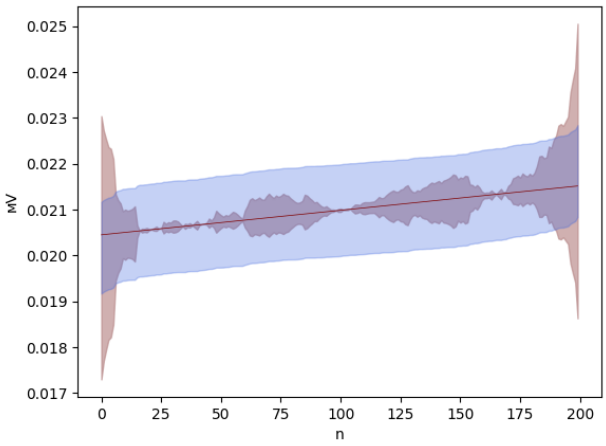

Рис. 4: Диаграмма рассеяния выборки X_1 и регрессионная прямая по модели

Розовым цветом обозначены скорректированные интервалы выборки X_1, голубым - первоначальные интервалы. Регрессионная прямая обозначена красным цветом.

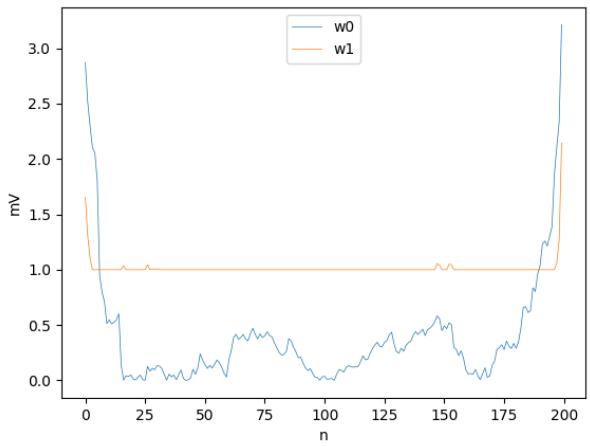

Рис. 5: Векторы w_0 и w_1

3.4 Анализ регрессионных остатков

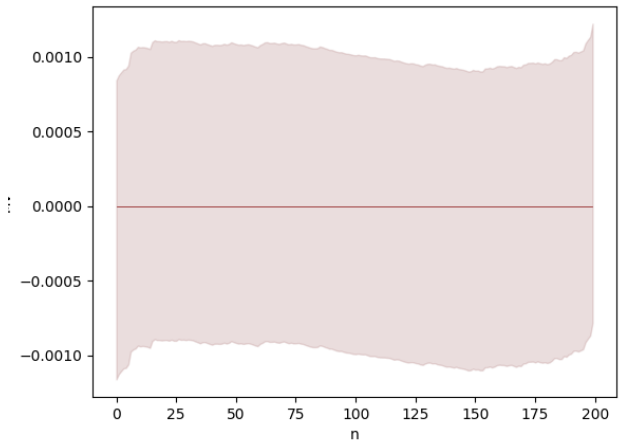

Рис. 6: Диаграмма рассеяния регрессионных остатков для модели без сужения интервалов

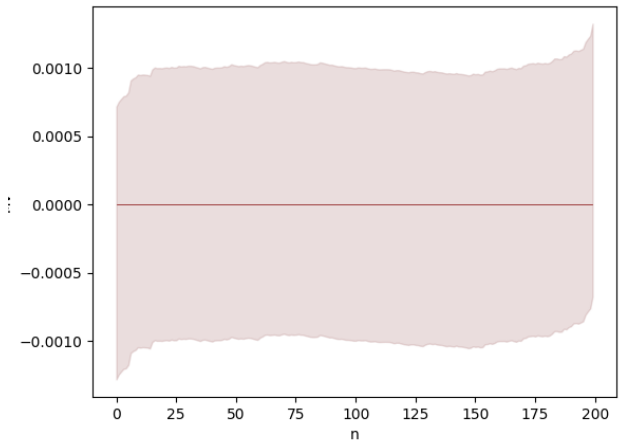

Рис. 7: Диаграмма рассеяния регрессионных остатков для модели с сужением и расширением интервалов

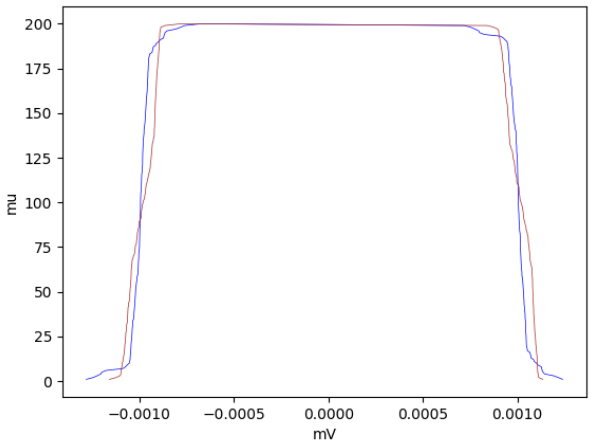

Рис. 8: Частоты элементарных подинтервалов регрессионных остатков при вычислении моды для двух моделей

Красный график – график частот элементарных подинтервалов регрессионных остатков при вычислении моды модели без сужения интервалов.
Синий график – график частот элементарных подинтервалов регрессионных остатков при вычислении моды модели с сужением и расширением интервалов.

Меры совместности регрессионных остатков:

$$\text{mode } X^0 = [-0.0007, 0.0007] \quad J_i(X^0) = 0.5335$$
$$\text{mode } X^1 = [-0.0008, 0.0008] \quad J_i(X^1) = 0.6808$$

Здесь X^0, X^1 – регрессионные остатки выборки X_1, вычисленные с использованием разных условий оптимизации.

3.5 Информационное множество задачи

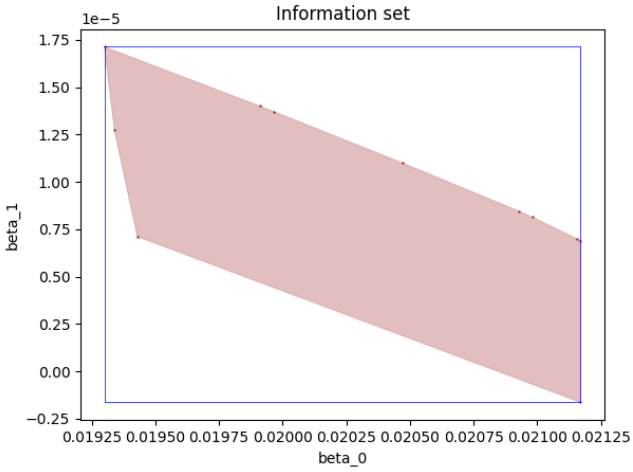

Рис. 9: Информационное множество задачи по заданной модели – красный брус, интервальная оболочка – синий брус

3.6 Коридор совместных зависимостей

Внешние интервальные оценки параметров модели:

$$\text{mid}\beta_0 = [0.0193, 0.0212]$$

$$\text{mid}\beta_1 = [-1.6382e-06, 1.7129e-05]$$

Подставляя эти значения в уравнение регрессии, получаем:

$$x(k) = \text{mid}\beta_0 + \text{mid}\beta_1 \cdot k \tag{9}$$

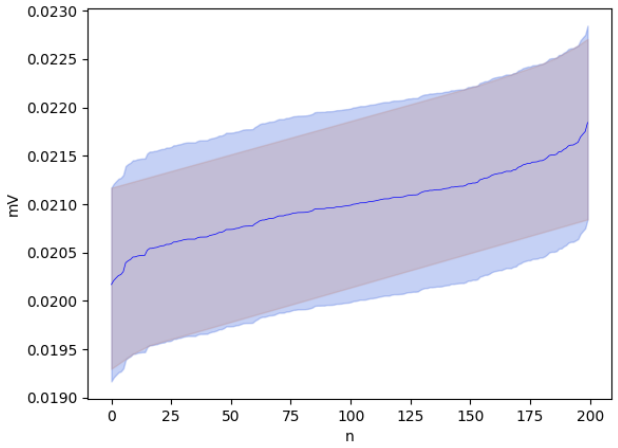

Рис. 10: Коридор совместных зависимостей

3.7 Прогноз вне области данных

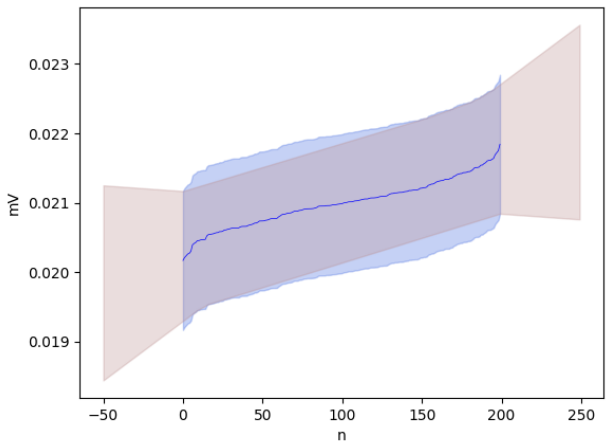

Рис. 11: Коридор совместных зависимостей. Построение прогноза

4 Обсуждение

4.1 Варьирование неопределенности изменений

Почти все компоненты вектора w оказались равны 1, то есть, расширения интервалов измерений почти не понадобилось.
Недостатком полученного значения с единичными значениями w_1^i является неучет расстояний точек регрессионной зависимости до данных интервальной выборки. Прямая, построенная по данной модели, "не чувствует"отклонений измерений от прямой на концах выборки – неопределенности измерений достаточно велики, чтобы покрыть этот эффект.

4.2 Варьирование неопределенности изменений с расширением и сужением интервалов

В результате построения модели первые 6 интервалов и последние 10 интервалов расширились, остальные интервалы сузились. Это наглядно показано на графике 4. Сумма компонент вектора w уменьшилась примерно в 3 раза. Таким образом, постановка задачи с возможностью одновременного расширения и сужения радиусов неопределенности измерений позволяет более гибко подходить к задаче оптимизации.
Из графиков векторов w_0 и w_1 5 можно сделать вывод, что график вектора w_0 содержит большее количество информации, чем график вектора w_1.

4.3 Анализ регрессионных остатков

Диаграммы рассеяния регрессионных остатков для модели только с расширением интервалов (6) и для модели с сужением и расширением интервалов (7) выглядит почти совпадающими из-за высокой степени совместности исходной выборки.
Данный вывод подтверждает и график частот элементарных подинтервалов регрессионных остатков при вычислении моды моделей (8). Графики очень похожи, однако график для частот элементарных подинтервалов регрессионных остатков при вычислении моды модели с расширением и сужением интервалов имеет немного более широкую внутреннюю оценку, что соответствует большей устойчивости к возмущениям данных.
Обе интервальные выборки остатков имеют довольно высокий коэффициент Жаккара, что свидетельствует о высокой степени совместности выборки.

4.4 Информационное множество задачи

Информационное множество задачи представляет представляет собой многогранный брус, а интервальные оценки, именуемые интервальной оболочкой, задают брус, описанный вокруг многогранного множества.

4.5 Коридор совместных зависимостей

Из внешнего вида коридора совместных зависимостей можно сделать вывод, что внутри его можно провести множество прямых и что он покрывает почти всю интервальную выборку.

4.6 Прогноз вне области данных

Была расширена область определения аргумента для модели и построен прогноз за пределами интервальной выборки. На основе полученных результатов можно сделать вывод, что величина неопределенности растет по мере удаления от области, в которой производились исходные измерения. Это обусловлено видом коридора зависимостей, расширяющимся за пределами области измерений.

5 Реализация

Лабораторная работа выполнена с использованием языка программирования Python 3.10 в среде разработки PyCharm Community с использованием библиотек

- pandas v.2.0.1
- matplotlib v.3.7.1
- numpy v.1.24.3
- scipy v.1.10.1
- pypoman v. 1.0.0

6 Приложение

Ссылка на репозиторий GitHub: `https://github.com/krzhld/mathstat`

Литература

[1] Баженов А.Н., Карпова А.А. Математические методы в физике. Курс для аспирантов ФТИ им. А.Ф.Иоффе. Практикум. – 2023. – 40 с.

[2] Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. – 816 с.