Санкт-Петербургский политехнический университет
Высшая школа прикладной математики и
вычислительной физики,
Физико-механический институт

Направление подготовки
01.03.02 «Прикладная математика и информатика»

Отчет по курсовому проекту
по дисциплине «Интервальный анализ»
на тему «Интервальная регрессия несовместных групп данных»

Выполнил студент гр. 5030102/80201
Кирпиченко С. Р.
Руководитель
Баженов А. Н.

Санкт-Петербург
2022

Содержание

Страница

Список иллюстраций

Список таблиц

1 Постановка задачи

Для предоставленных данных

x	y
0	30
64	30
128	26
192	24
256	17
320	11
384	7
448	0
384	6
320	7
256	11
192	14
128	20
64	25
0	29

Таблица 1: Исходные данные

Рассматривается задача построения интервальной регрессии $\mathbf{X}\beta = \mathbf{y}$. Необходимо провести линейный регрессионный анализ и добиться совместности выборки посредством введения в рассмотрение интервальных величин. Необходимо провести вычисления и привести иллюстрации:

- Обработки данных посредством введения интервальных величин в отклик
- Обработки данных посредством введения брусов совместности, добиться сильной линейной совместности интервальных данных
- Обработки данных как единого целого с помощью метода квадратных подсистем

2 Теория

2.1 Понятие накрывающей и совместной выборки

Брус неопределенности измерения называется накрывающим, если он гарантированно содержит истинные значения измеряемых величин входных и выходных переменных зависимости. Иначе брус называется ненакрывающим.

Накрывающей выборкой называется совокупность накрывающих измерений. Если в выборке содержится хотя бы одно ненакрывающее измерение, то в таком случае она называется ненакрывающей.

Данное определение тесно связано с понятием совместности. Если имеются точные измерения предикторных переменных и некоторые неопределенности в измерении отклика, то в таком случае построенная нами линейная регрессия называется совместной (согласованной) с данными, если ее график проходит через все отрезки неопределенности.

2.2 Слабая и сильная совместность

Функциональная зависимость называется слабо совместной с данными, если ее график проходит через каждый брус неопределенности измерений хотя бы для одного значения аргумента.

Функциональная зависимость называется сильно совместной с данными, если ее график проходит через каждый брус неопределенности измерений для любого значения аргумента из интервалов неопределенности входных переменных.

2.3 Алгоритм достижения совместности входных данных

Для решения поставленной задачи будет использован аппарат линейного программирования. Каждая точка исходных данных будет интерпретирована как середина отрезка неопределенности, который будет расширяться для достижения совместности выборки. В ходе решения составленной задачи линейного программирования будут получены коэффициенты регрессии β и величины $\operatorname{rad}\mathbf{y}_i = w_i$, минимизированные по норме из пространства ℓ_1.

Формальная постановка задачи:

$$\begin{cases} x_i\beta_1 + \beta_0 \le y_i + w_i, & i = \overline{1,n} \\ x_i\beta_1 + \beta_0 \ge y_i - w_i, & i = \overline{1,n} \\ f(\beta, w) = \sum w_i \to \min \end{cases} \tag{1}$$

В результате решения симплекс методом находятся $n + 2$ параметра.

2.4 Алгоритм достижения сильной совместности интервальных брусов

Идейно аналогичен предыдущему алгоритму, вводятся дополнительные величины для расширения неопределенности входных данных по координате предиктора. Для сохранения линейности постановки необходимо задавать коэффициент наклона прямой β_1 извне. Минимизируется сумма коэффициентов расширения брусьев.

Формальная постановка задачи:

$$\begin{cases} (x_i - q_i)\beta_1 + \beta_0 \le y_i + w_i, & i = \overline{1,n} \\ (x_i + q_i)\beta_1 + \beta_0 \ge y_i - w_i, & i = \overline{1,n} \\ f(\beta_0, q, w) = \sum w_i + \sum q_i \to \min \end{cases} \tag{2}$$

Постановка знаков после величин x_i обусловлена отрицательной корреляцией исходных данных. Итого в результате решения задачи получается $2n + 1$ параметр.

2.5 Алгоритм обработки данных в целом

Для использования данного метода необходимо перейти из $\mathbb{IR}$ в полную интервальную арифметику $\mathbb{KR}$. Две несовместные «ветви» исходных данных ($\mathbf{y}_k^1$ и $\mathbf{y}_k^2$) рассматриваются вместе.

Алгоритм:

1. Составляем вектор минимумов по включению $\mathbf{y}_k = \mathbf{y}_k^1 \bigwedge \mathbf{y}_k^2 = [\max\{\underline{\mathbf{y}}_k^1, \underline{\mathbf{y}}_k^2\}, \min\{\overline{\mathbf{y}}_k^1, \overline{\mathbf{y}}_k^2\}]$
2. Решаем задачу нахождения максимума совместности: $\mathbf{X}\beta \subseteq \mathbf{y}$
3. Получив переопределенную ИСЛАУ, воспользуемся методом квадратных подсистем. Исходная ИСЛАУ разбивается на несколько более мелких $\mathbf{X}^{(1)}\beta = \mathbf{y}^{(1)}, \ldots, \mathbf{X}^{(m)}\beta = \mathbf{y}^{(m)}$ с квадратными матрицами $\mathbf{X}^{(i)}$.
4. Данные системы решаются известными численными методами (в данной работе - субдифференциальным методом Ньютона), после чего строится пересечение полученных оценок множеств решений $\Xi = \bigcap_{i=1}^{m} \Xi_i$.
5. Множество Ξ является оценкой информационного множества для исходных несовместных данных.

3 Реализация

Для осуществления вычислений и визуализации результатов использовалась среда Octave с пакетом интервальной арифметики interval и сторонней библиотекой полной интервальной арифметики kinterval. Для решения задач линейного программирования использовалась функция glpk.

4 Результаты

Как было описано выше, исходные данные отрицательно коррелированы. Можно выделить две квазилинейных «ветви», изображенных синим и красным цветом на графике ниже.

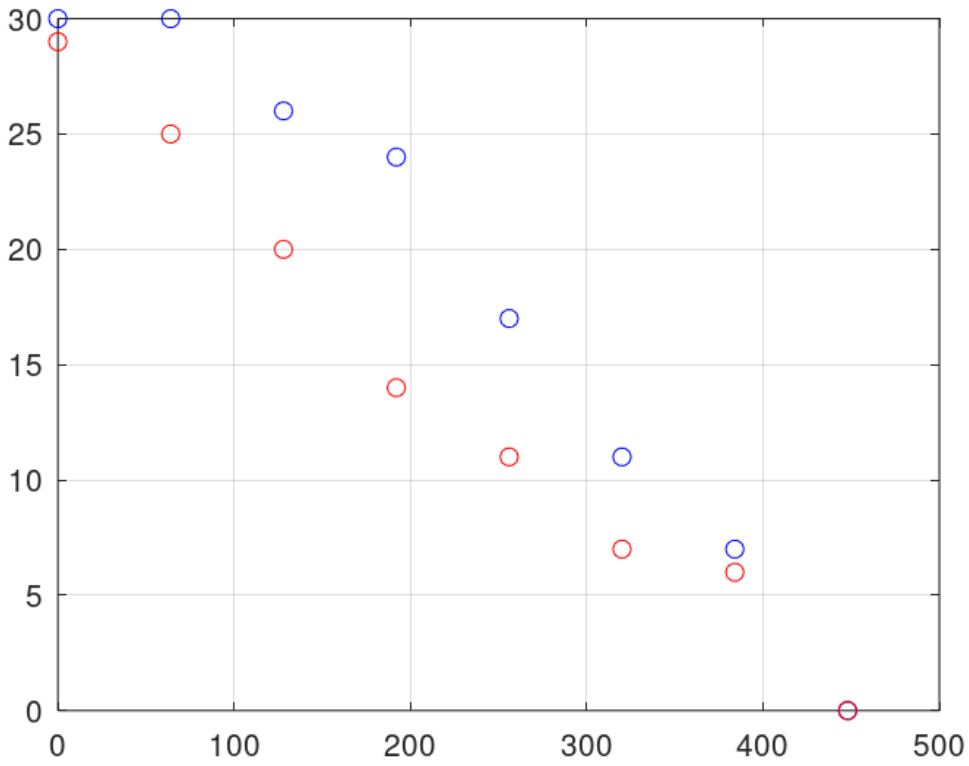

Рис. 1: Исходные точечные данные

4.1 Обработка данных посредством введения интервальных величин в отклик

Рассмотрим выделенные выше ветви данных по отдельности и воспользуемся алгоритмом 1. Результаты приведены на графике ниже. Цветовые обозначения ветвей сохранены.

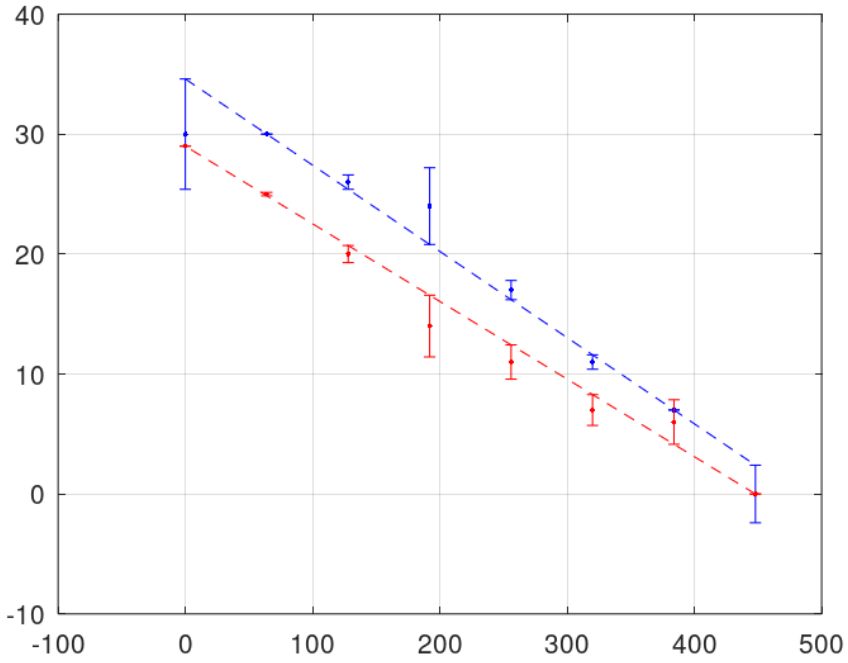

Рис. 2: Достижение совместности данных для разных ветвей

Для синей ветви данных получены следующие результаты: $\beta_0 = 34.6, \beta_1 \approx -0.072, ||w||_1 =$ 12.2. Для красной ветви: $\beta_0 = 29,\ \beta_1 \approx -0.065,\ ||w||_1 \approx 8$.

Обработаем всю выборку и добьемся ее совместности.

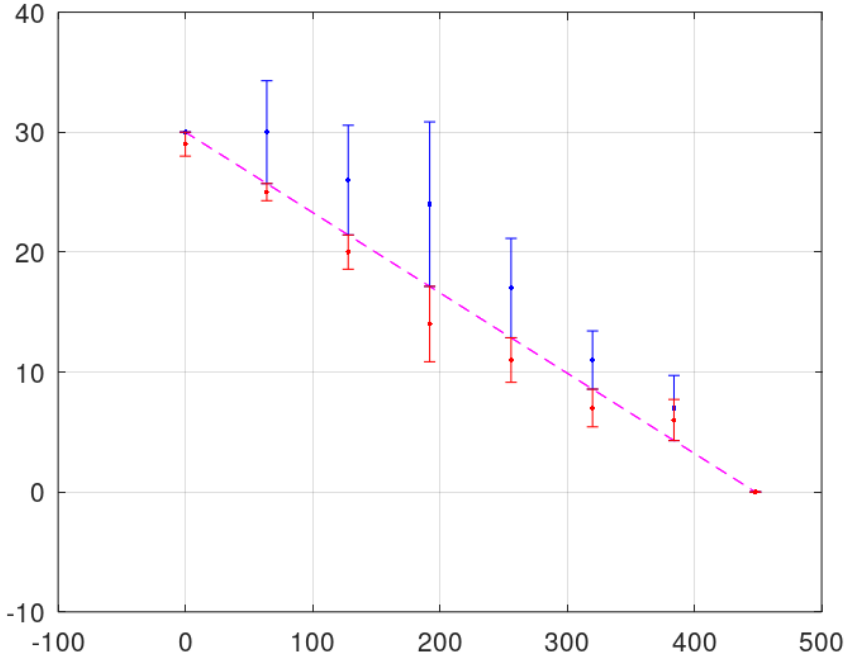

Рис. 3: Достижение совместности данных для всей выборки

Численные результаты: $\beta_0 = 30,\ \beta_1 \approx -0.067,\ ||w||_1 \approx 36.43$.

4.2 Обработка данных посредством введения брусов совместности

Алгоритму 2 нет смысла искусственно расширять брусы по координате предиктора, так как в случае точечной матрицы X совместная выборка (например, полученная в предыдущем пункте) сильно совместна по определению. Ввиду этого введем неопределенность в координату x каждой исходной точки так, что $\operatorname{rad}\mathbf{x}_i = \frac{1}{4}(x_2 - x_1) = 16$. Координаты брусов $\mathbf{y}$ получены следующим образом: $\mathbf{y}_k = [\min\{y_k^1, y_k^2\}, \max\{y_k^1, y_k^2\}]$. Коэффициент β_1 возьмем из результатов построения регрессии в предыдущем пункте, $\beta_1 \approx -0.067$.

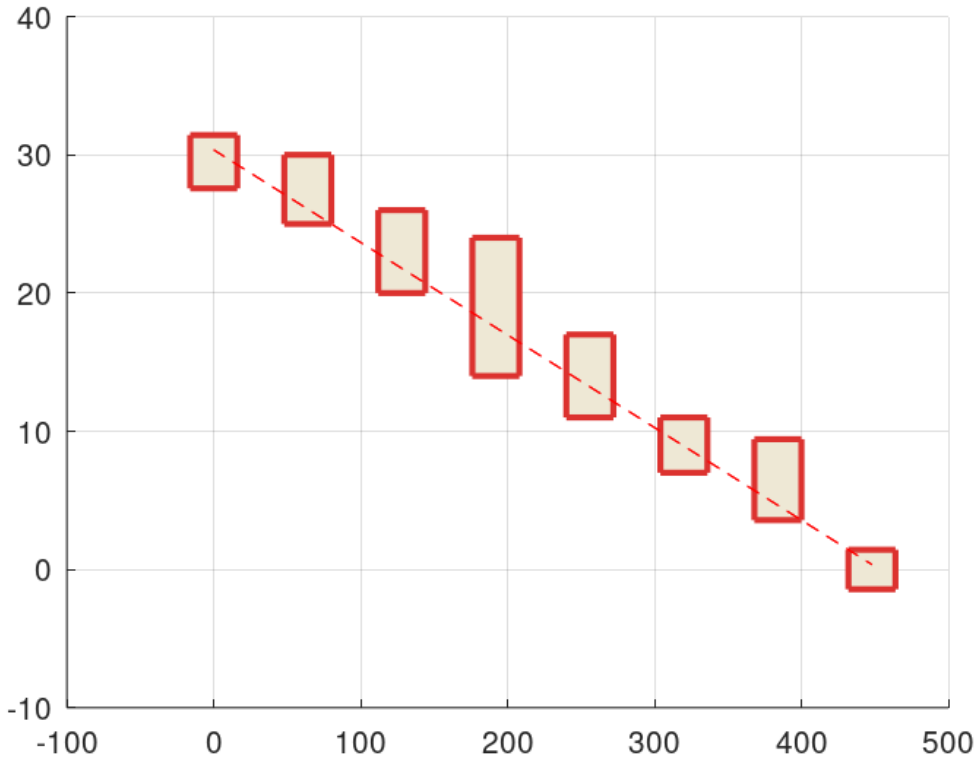

Рис. 4: Сильная линейная совместность для интервальных предиктора и отклика

Численные результаты: $\beta_0 \approx 30.36$, $||w||_1 \approx 5.29$, $||q||_1 = 0$.

4.3 Обработка данных как единого целого с помощью метода квадратных подсистем

Для построения регрессии данным методом в данные была добавлена неопределенность: $\operatorname{rad}\mathbf{x}_i = 16$, $\operatorname{rad}\mathbf{y}_i = 1$. Результат операции взятия минимумов по включению:

x	y
[-16,16]	[29,30]
[48,80]	[29,26]
[112,144]	[25,21]
[176,208]	[23,15]
[240,272]	[16,12]
[304,336]	[10,8]
[368,400]	[6,7]
[432,464]	[-1,1]

Таблица 2: $\mathbf{y}_k^1 \bigwedge \mathbf{y}_k^2$

Для решения переопределенной системы она была разбита на 6 частей: ИСЛАУ 2×2, каждая смещена на строку ниже относительно предыдущей. Решение получено с помощью субдифференциального метода Ньютона из библиотеки kinterval.

Таблица результатов решения ИСЛАУ методом квадратных подсистем:

Номера строк исходной ИСЛАУ	Ξ_i
$1, 2$	$\begin{pmatrix}[0, -0.083333] \\ [29, 30]\end{pmatrix}$
$2, 3$	$\begin{pmatrix}[-0.0625, -0.078125] \\ [34, 29.75]\end{pmatrix}$
$3, 4$	$\begin{pmatrix}[-0.03125, -0.09375] \\ [29.5, 31.5]\end{pmatrix}$
$4, 5$	$\begin{pmatrix}[-0.10938, -0.046875] \\ [45.75, 23.25]\end{pmatrix}$
$5, 6$	$\begin{pmatrix}[-0.09375, -0.0625] \\ [41.5, 27]\end{pmatrix}$
$6, 7$	$\begin{pmatrix}[-0.0625, -0.015625] \\ [31, 12.75]\end{pmatrix}$

Таблица 3: Результаты метода квадратных подсистем

Пересечение $\Xi = \bigcap_{i=1}^{m} \Xi_i$ обозначено на следующем графике красным цветом.

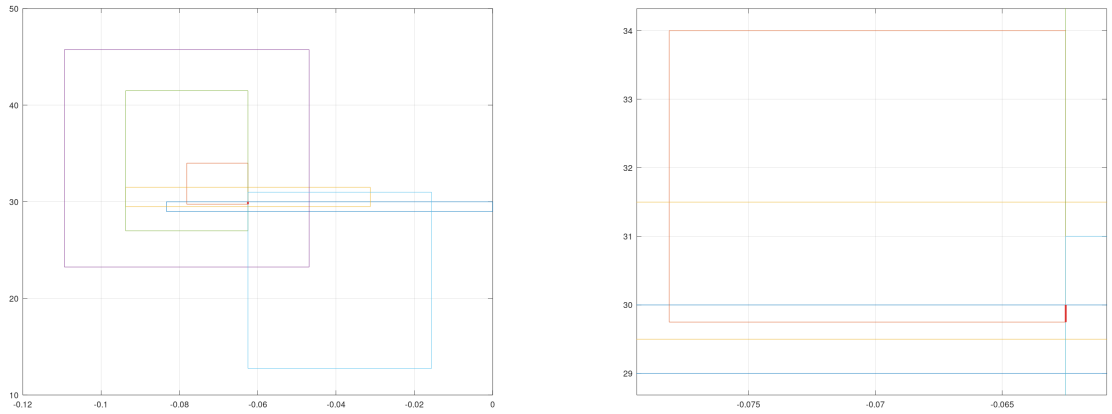

Рис. 5: Брусы Ξ_i, полученные методом квадратных подсистем и множество Ξ

Численные результаты: $\Xi \approx \begin{pmatrix}[-0.0625,-0.0625] \\ [29.75,30]\end{pmatrix}$. Ввиду узости полученной оценки информационного множества был выбран только один набор параметров регрессии $\beta = \operatorname{mid} \Xi = \begin{pmatrix} -0.0625 \\ 29.875 \end{pmatrix}$. На следующем графике изображены исходные точечные данные, брусья, полученные после операции взятия минимумов по включению и полученная регрессионная прямая.

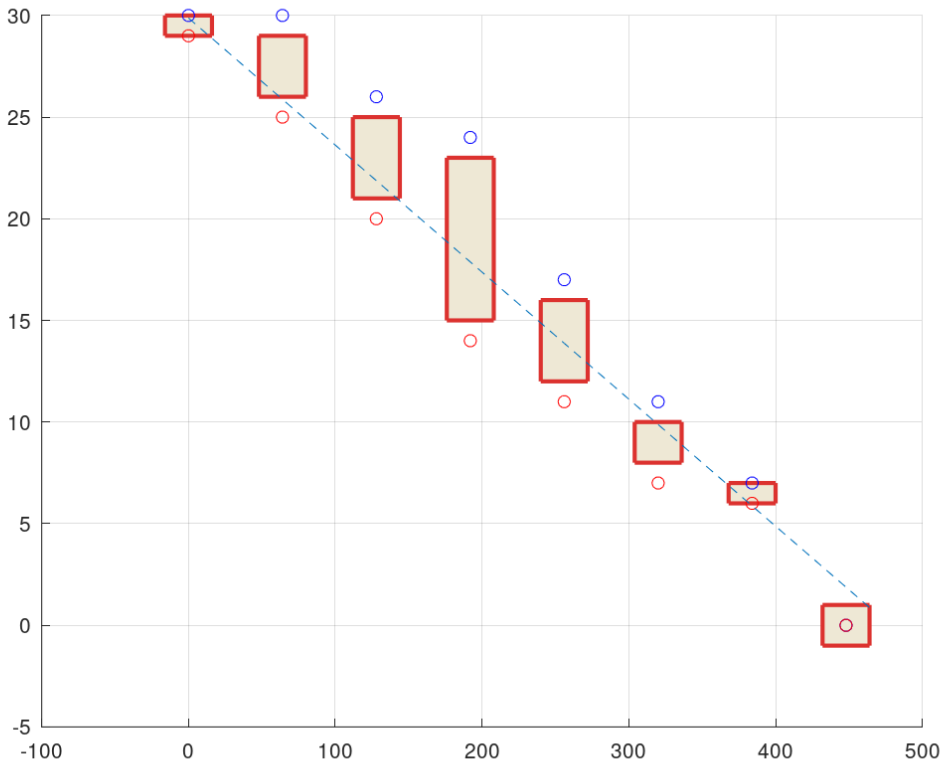

Рис. 6: Результат построений с помощью обработки данных в $\mathbb{KR}$

5 Обсуждение

1. Как уже было описано, исходя из данных на графике 1 можно сделать вывод о характере входных данных: в исходной выборке имеются две квазилинейные подвыборки, обе отрицательно коррелированы.

2. При достижении совместности этих ветвей по отдельности были получены практически параллельные прямые, смещенные друг относительно друга.

3. По графику 2 видно, что алгоритм 1 достиг своей цели: обе прямые подходят через полученные интервалы, совместность обоих выборок была достигнута минимальным суммарным расширением отклика. Информационное множество в обоих случаях состоит из одной точки, так как в обоих ветвях видны по 2 точечных величины, следовательно, прямая, проходящая через них, задается единственным образом.

4. По графику 2 и численным результатам соответствующей процедуры видно, что нижняя ветвь содержит меньше отклонений от линейной зависимости: для достижения совместности в ее отклики пришлось вносить меньшую неопределенность.

5. Ситуация с рассмотрением выборки в целом (график 3) аналогична: информационное множество точечное - есть два точечных отклика. Можно отметить, что расширение, которое пришлось внести, значительно больше, чем сумма расширений для двух ветвей по отдельности.

6. В нижнюю ветвь все так же было внесено меньше неопределенности, чем в верхнюю: красные интервалы на графике 3 заметно уже синих.

7. По графику 4 видно, что полученная зависимость действительно является сильно совместной - алгоритм 2 достиг нужного результата.

8. Сравнивая коэффициенты β_0, полученные после обработки выборки алгоритмами 1 и 2, приходим к выводу, что полученная с помощью алгоритма 1 регрессия весьма оптимальна: даже при внесении дополнительной неопределенности в исходные данные прямая не нуждается в большой коррекции для обеспечения сильной совместности.

9. В ходе применения алгоритма 2 исходные брусья были в небольшой степени расширены по отклику. По предиктору расширение внесено не было.

10. Полученное в ходе обработки выборки третьим способом информационное множество получилось очень узким, практически точечным по параметру β_1. Это может быть объяснено как не самым удачным выбором матриц в методе квадратных подсистем (на графике 5 видно, что один из брусьев почти что вырождает пересечение по первой координате), так и не лучшей совместностью исходных данных.

11. По графику 6 можно сделать вывод, что полученная регрессия обладает слабой совместностью с брусьями $\mathbf{y}_k^1 \bigwedge \mathbf{y}_k^2$.

12. Полученные тремя подходами результаты схожи: различие по параметру β_1 составляет около 7%, по параметру β_0 - не более 2%.

Исходный код

С исходным кодом программы и отчета можно ознакомиться в репозитории `https://github.com/Stasychbr/IntervalArith`.

Список литературы

[1] А. Н. Баженов. Лекции по интервальному анализу. СПбПУ. 2021 https://cloud.mail.ru/public/VSFh/gJgtFVynE

[2] А. Н. Баженов. Интервальный анализ. Основы теории и учебные примеры: учебное пособие. - СПб. 2020 - 78 с.

[3] С. И. Жилин. Библиотека kinterval. Код в Octave. Альфа версия. 2020.