Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

Отчёт
по лабораторным работам №9-10
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

1 Постановка задачи **3**

2 Теория **4**
2.1 Представление данных 4
2.2 Линейная регрессия 4
2.2.1 Описание модели 4
2.2.2 Метод наименьших модулей 4
2.3 Предварительная обработка данных 5
2.4 Коэффициент Жаккара 5
2.5 Процедура оптимизации 6

3 Результаты **7**

4 Обсуждение **11**
4.1 Множители коррекции w 11
4.2 Коэффициент Жаккара 11
4.3 Гистограмма объединённых данных при оптимальном значении R 11

5 Реализация **12**

6 Приложение **13**

Список иллюстраций

1 Схема установки 3
2 Исходные данные из экспериментов 7
3 Интервальное представление исходных данных 7
4 Линейная модель дрейфа данных 8
5 Гистограммы значений множителей коррекции w 8
6 Скорректированные модели данных 9
7 Гистограммы скорректированных данных 9
8 Значение коэффициента Жаккара от калибровочного множителя от R_{21} 10
9 Гистограмма объединённых данных при оптимальном значении R_{21} 10

1 Постановка задачи

Исследование из области солнечной энергетики [1]. На рис. 1 показана схема установки для исследования фотоэлектрических характеристик.

Рис. 1: Схема установки

Калибровка датчика ФП1 производится по эталону ФП2. Зависимость между квантовыми эффективностями датчиков предполагается одинаковой для каждой пары измерений

$$QE_2 = \frac{I_2}{I_1} * QE_1, \tag{1}$$

где QE_2, QE_1 – квантовые эффективности эталонного и исследуемого датчиков, I_2, I_1 – измеренные токи.

Исходные данные. Даны 2 выборки данных с интервальной неопределенностью. Одна из них относится к эталонному датчику ФП2, другая – к исследуемому датчику ФП1.
Задача. Требуется определить коэффициент калибровки

$$R_{21} = \frac{I_2}{I_1} \tag{2}$$

методом линейной регрессии на множестве интервальных данных и коэффициента Жаккара.

2 Теория

2.1 Представление данных

В первую очередь представим данные таким образом, чтобы применить понятия статистики данных с интервальной неопределенностью. Один из распространённых способов получения интервальных результатов в первичных измерениях – это «обинтерваливание» точечных значений, когда к точечному базовому зачению $\dot{x}$, которое считывается по показаниям измерительного прибора, прибавляется *интервал погрешности* ε:

$$\mathbf{x} = \dot{x} + \varepsilon \tag{3}$$

Интервал погрешности зададим как

$$\varepsilon = [-\xi; \xi]$$

В конкретных измерениях примем $\xi = 10^{-4}$ мВ.
Согласно терминологии интервального анализа, рассматриваемая выборка – это вектор интервалов, или интервальный вектор $x = (x_1, x_2, ..., x_n)$.
Информационным множеством в случае оценивания единичной физической величины по выборке интервальных данных будет также интервал, который называют информационным интервалом. Неформально говоря, это интервал, содержащий значения оцениваемой величины, которые «совместны» с измерениями выборки («согласуются» с данными этих измерений).

2.2 Линейная регрессия

2.2.1 Описание модели

Регрессионную модель описания данных называют простой линейной, если заданный набор данных аппроксимируется прямой с внесённой добавкой в виде некоторой нормально распределенной ошибки:

$$y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i, i \in \overline{1, n} \tag{4}$$

где
$\{x_i\}_{i=1}^n$ – заданные значения,
$\{y_i\}_{i=1}^n$ – параметры отклика,
$\{\varepsilon_i\}_{i=1}^n$ – независимые, центрированные, нормально распределённые случайные величины с неизвестной дисперсией δ, суть предполагаемые погрешности,
β_0, β_1 – параметры, подлежащие оцениванию.

В данной модели мы считаем, что у заданных значений нет погрешности (пренебрегаем ей). Считаем, что основная погрешность получается при измерении y_i.

2.2.2 Метод наименьших модулей

Для наиболее точного приближения входных с фотоприемников данных y_i аппроксимирующей функции $f(x_i)$ используется метод наименьших модулей. Этот метод основывается на минимизации l^1-нормы разности последовательностей:

$$\|f(x_i) - y_i\|_{l^1} \to min \tag{5}$$

В данном случае ставится задача линейного программирования, решение которой дает нам коэффициенты β_0 и β_1, а также вектор множителей коррекции данных w, на который стоит домножить погрешности наших интервальных данных.
По итогу получается следующая задача линейного программирования:

$$\sum_{i=1}^{n} w_i \to min \tag{6}$$

При ограничениях:

$$\beta_0 + \beta_1 * x_i - w_i * \xi \leq y_i, i \in \overline{1, n} \tag{7}$$

$$\beta_0 + \beta_1 * x_i + w_i * \xi \geq y_i, i \in \overline{1, n} \tag{8}$$

$$w_i \geq 1, i \in \overline{1, n} \tag{9}$$

2.3 Предварительная обработка данных

Из последующих результатов ясно, что для оценки коэффициента калибровки необходима предварительная обработка данных. Для этого можем задаться линейной моделью дрейфа.

$$Lin(n) = A + B \cdot n, n \in \overline{1, N}, \tag{10}$$

Поставив задачу линейного программирования, воспользуемся методом наименьших модулей (2.2.2) и найдем коэффициенты A, i и вектор w множителей коррекции данных для каждого из фотоприемников ФП1 и ФП2.
Для данных с первого фотоприемника: $A = 4.748, B = 9.173 \cdot 10^{-6}$
Для данных со второго фотоприемника: $A = 5.182, B = 1.105 \cdot 10^{-5}$
В последствии множитель коррекции данных необходимо применить к погрешностям выборки, чтобы получить данные, которые согласовывались с линейной моделью дрейфа:

$$I^f(n) = \dot{x}(n) + \varepsilon \cdot w(n), n \in \overline{1, N} \tag{11}$$

В итоге необходимо построить «спрямленные» данные выборки: получить их можно путем вычитания из исходных данных линейную компоненту:

$$I^c(n) = I^f(n) - B \cdot n, n \in \overline{1, N} \tag{12}$$

2.4 Коэффициент Жаккара

В различных областях анализа данных в науках о Земле, биологии, информатике используют множество мер сходства множеств. Иначе их называют коэффициентами сходства. Нами рассматривается модификация индекса Жаккара для интервальных данных:

$$JK(x) = \frac{wid(\wedge x_i)}{wid(\vee x_i)} \tag{13}$$

В качестве меры рассматривается ширина интервала, а вместо операций пересечения и объединения – операции взятия минимума и максимума по включению двух величин в интервальной арифметике (Каухера). Заметим что минимум по включению может быть неправильным интервалом, а значит данный коэффициент будет нормирован в отрезке $[-1, 1]$.

2.5 Процедура оптимизации

Чтобы найти оптимальный параметр калибровки R_{21} (2) необходимо поставить и решить задачу максимизации коэффициента Жаккара, зависящего от параметра калибровки:

$$JK(x_{all}(R)) \to max, \tag{14}$$

где JK – это коэффициент Жаккара (13), x_{all} – это выборка, полученная как

$$x_{all}(R) = I_1^c \cdot R \cup I_2^c, \tag{15}$$

т.е. конкатенация полученных спрямлённых выборок I_1^c и I_2^c, последняя из которых домножена на R – параметр калибровки. Поиск оптимального R будем проводить методом дихотомии в отрезке $[1; 3]$. Найденный таким образом R и будет искомым оптимальным R_{21} в силу наибольшего совпадения, оцененного коэффицентом Жаккара.

3 Результаты

Рис. 2: Исходные данные из экспериментов

Рис. 3: Интервальное представление исходных данных

Рис. 4: Линейная модель дрейфа данных

Рис. 5: Гистограммы значений множителей коррекции w

Рис. 6: Скорректированные модели данных

Рис. 7: Гистограммы скорректированных данных

Рис. 8: Значение коэффициента Жаккара от калибровочного множителя от R_{21}

Рис. 9: Гистограмма объединённых данных при оптимальном значении R_{21}

4 Обсуждение

4.1 Множители коррекции w

По преобладанию множителя 1 на гистограммах значений множителей коррекции (рис. 3), можно сказать что примерно половина данных не требует коррекции. Этот факт свидетельствует о том, что линейная модель дрейфа данных является разумным приближением.

4.2 Коэффициент Жаккара

На рис. 3 видно, что оптимальным множителем R_{21} является число, равное 1.091266955. Однако видно, что коэффициент Жаккара при оптимальном значении едва-едва превышает 0, а интервал, при котором JK $\geq 0, (10^{-9} - 10^{-10})$. Это показывает на то, что исходные данные имеют ряд неточностей, которые сложно устранить. Это же можно было и заметить на рис. 3, иллюстрирующий входные данные. Однако, как будет далее видно, подобранный коэффициент R_{21} приблизит данные первого фотоприемника к данным второго фотоприемника.

4.3 Гистограмма объединённых данных при оптимальном значении R

Сравнивая гистограмму объединённых данных при оптимальном значении R_{21} (рис. 3) с гистограммами скорректированных данных (рис. 3), видно, что гистограмма объединённых данных повторяет форму гистограммы входных данных с ФП1, однако пик гистограммы смещен в сторону значения пика на гистограмме ФП2.

5 Реализация

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

- pandas v.1.5.3
- matplotlib v.3.7.1
- numpy v.1.24.2
- scipy v.1.10.1

6 Приложение

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

Литература

[1] `Баженов А.Н., Тельнова А.Ю.` Обобщение коэффициента Жаккара для анализа данных с интервальной неопределённостью // Измерительная техника. – 2022. № 12, С. 12-15.

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