Санкт-Петербургский Политехнический университет Петра Великого

Отчет по лабораторной работе №4 по дисциплине "Интервальный анализ"

Линейная регрессия

Выполнил студент: Кромачев Максим
группа: 5030102/10201
Проверил: доцент Баженов Александр Николаевич

Содержание

1 Постановка задачи

Дан измеритель, на вход которого поступает калибровочный сигнал - набор постоянных напряжений

$$X = \{x_i\}_{i=1}^{100}$$

а данные на выходе - набор интервальных данных

$$Y = \{y_k\}_{k=1}^{100}, \text{ rad } y = \frac{1}{2^N}B,\ N = 14$$

Файлы данных хранятся в бинарном формате и считывается в соответствии со следующим преобразованием:

$$V = Code/16384 - 0.5$$

Необходимо оценить значения β_0 и β_1 - параметров линейной регрессии

$$y = \beta_0 + \beta_1 * x$$

Оценки значений Y:

- внутренняя (in) - интервал между первым и третьим квартилем
- внешняя (ex) - границы бокс-плота

Требуется:

- Решить ИСЛАУ (1) для внутренних и внешних оценок y
- Построить множество решений β_0, β_1
- Построить коридор совместных зависимостей, используя пример — https://github.com/szhilin/octave-interval-examples/ blob/master/SteamGenerator.ipynb

2 Теоретическое обоснование

2.1 Бокс-плот Тьюки

Боксплот (англ. box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей. Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили и выбросы. Границами ящика служат первый и третий квартили, линия в середине ящика — медиана. Концы усов - края статистически значимой выборки (без выброса). Длину «усов» определяют разность первого квартиля и полутора межквартальных расстояний и сумма третьего квартиля и полутора межквартальных расстояний. Формула имеет вид
$X_1 = Q_1 - \frac{3}{2}(Q_3 - Q_1)$, $X_2 = Q_3 + \frac{3}{2}(Q_3 - Q_1)$
где X_1 - нижняя граница уса, X_2 — верхняя граница уса, Q_1 — первый квартиль, Q_3 - третий квартиль. Данные, выходящие за границы усов (выбросы), отображаются на графике в виде маленьких кружков. Выбросами считаются величины , такие что:
$\begin{bmatrix} x < X_1^T \\ x > X_2^T \end{bmatrix}$

2.2 Интервальная мода

Пусть имеется интервальная выборка
$X = \{x_i\}$ Сформируем массив интервалов z из концов интервалов X.
Для каждого интервала z_i подсчитываем число μ_i интервалов из выборки X_i, включающих z_i. Максимальные $\mu_i = \max \mu$ достигаются для индексного множества K. Тогда можно найти интервальную моду как мультиинтервал
mode X = $\bigcup_{k \in K} z_k$

2.3 Интервальная медиана Крейновича

Пусть дана выборка X = x_i. Пусть $\underline{c} = \underline{x_i}$, $\overline{c} = \overline{x_i}$ — конфигурация точек, составленные, соответственно, из левых и правых концов интервалов из X. Медиана Крейновича $med_K X$ интервальной выборки X — это интервал
$med_K = [med_{\underline{c}}, med_{\overline{c}}]$.

2.4 Интервальная медиана Пролубникова

Зададим отношения порядка на алгебре R. Говорят, что неравенство $a \leq b$ выполняется

1. в сильном смысле, если $\forall a \in R\ \forall b \in R : \overline{a} \leq \underline{b}$
2. в слабом смысле, если $\exists a \in R\ \exists b \in R : \underline{a} \leq \overline{b}$
3. в $\forall\exists$-смысле, если $\forall a \in R\ \exists b \in R : \underline{a} \leq \underline{b}$
4. в $\exists\forall$-смысле, если $\exists a \in R\ \forall b \in R : \overline{a} \leq \overline{b}$
5. в центральном смысле, если $\frac{\overline{a}+\underline{a}}{2} \leq \frac{\overline{b}+\underline{b}}{2}$

Для элементов выборки X можно определить линейный порядок, используя любое из пяти вышеуказанных отношений порядка на R. То есть, если $i \neq j$, то либо $x_i \leq x_j$, либо $x_i \geq x_j$ для любого из этих отношений порядка.
Медиана Пролубникова $med_P X$ выборки X — это интервал x_m, для которого половина интервалов из X лежит слева, а половина — справа.
В ситуации, когда имеются два элемента подинтервала x_m и x_{m+1}, расположенных посередине вариационного ряда, $x_m \neq x_{m+1}$ медиана может быть определена естественным обобщением взятия полусуммы точечных значений, расположенных посередине ряда из точечных значений, в случае интервальной выборки взятие полусуммы интервалов x_m и x_{m+1}:
$med_P X = \frac{x_m + x_{m+1}}{2}$.

3 Описание работы

Лабораторная работа выполнена на языке программирования Python в среде разработки VSCode. В ходе работы были использованы следующие библиотеки: numpy, scipy, intervalpy и matplotlib. GitHub репозиторий: `https://github.com/kromachmax/Intervalka.git`

3.1 Описание алгоритма

Каждый из 8 файлов содержит 100 фреймов, каждый из которых включает 1024 массива, состоящих из 8 двухбайтовых значений. В результате обработки этих данных было сформировано $1024 \times 8 = 8192$ ИСЛАУ, представленных в следующем виде:

$$\begin{pmatrix} [\mathrm{x}_1,\, x_1] & [1,1] \\ \vdots & \vdots \\ [\mathrm{x}_8,\, x_8] & [1,1] \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix} = \begin{pmatrix} \widehat{y_{1i}} \\ \vdots \\ \widehat{y_{8i}} \end{pmatrix}$$

- j — порядковый номер файла, $j \in \overline{1,8}$
- i — номер пикселя внутри файла, $i \in \overline{1,8192}$
- x_j — вольтаж, определяемый по первой цифре названия файла
- $\widehat{y_{ji}}$ — оценка значения, соответствующее каждому пикселю, по всем 100 фреймам
- β_0 и β_1 — искомые параметры линейной регрессии.

Каждая система линейных алгебраических уравнений была решена с использованием метода Дж. Рона [1], реализованном в библиотеке intervalpy. В результате были получены два множества интервалов оценок: $B_0 = \{\beta_0\}_{i=1}^{8192}$ и $B_1 = \{\beta_1\}_{i=1}^{8192}$

Оценка каждого из параметров линейной регрессии будем производить следующим образом:

1. $\widehat{\beta_0} = med_K B_0$, $\widehat{\beta_1} = med_K B_1$
2. $\widehat{\beta_0} = med_P B_0$, $\widehat{\beta_1} = med_P B_1$
3. $\widehat{\beta_0} = \mathrm{mode} B_0$, $\widehat{\beta_1} = \mathrm{mode} B_1$

Таким образом, конечные значения $\widehat{\beta_0}$ и $\widehat{\beta_1}$ служат наиболее вероятными оценками параметров регрессии, что позволяет более точно анализировать зависимость между переменными в исследуемых данных.

4 Результаты

4.1 Оценки

В ходе лабораторной работы для внутренней оценки были получены следующие результаты:

- $\bigcap_{i=1}^{8192}$, $\beta_{0,i} = \emptyset$
- $\bigcap_{i=1}^{8192} \beta_{1,i} = \emptyset$
- $med_K B_0 = [8029.18,\ 8190.85]$ и $med_K B_1 = [12857.0,\ 13289.5]$ для внутренней оценки медианой Крейновича
- $med_P B_0 = [8023.65,\ 8195.88]$ и $med_P B_1 = [12879.9,\ 13280.5]$ для внутренней оценки медианой Пролубникова

- modeB_0 = [8083.32, 8083.33], [8086.78, 8086.80] и modeB_1 = [13070.5, 13072.5] для внутренней оценки модой.

Для внешней оценки были получены следующие результаты:

- $\bigcap_{i=1}^{8192} \beta_{0,i} = \emptyset$
- $\bigcap_{i=1}^{8192} \beta_{1,i} = \emptyset$
- $med_K B_0$ = [7780.52, 8430.21] и $med_K B_1$ = [12193.0, 13950.3] для внутренней оценки медианой Крейновича
- $med_P B_0$ = [7765.31, 8454.22] и $med_P B_1$ = [12279.1, 13881.3] для внутренней оценки медианой Пролубникова
- modeB_0 = [7927.51, 8224.58] и modeB_1 = [13097.9, 13573.8] для внутренней оценки модой.

4.2 Графики

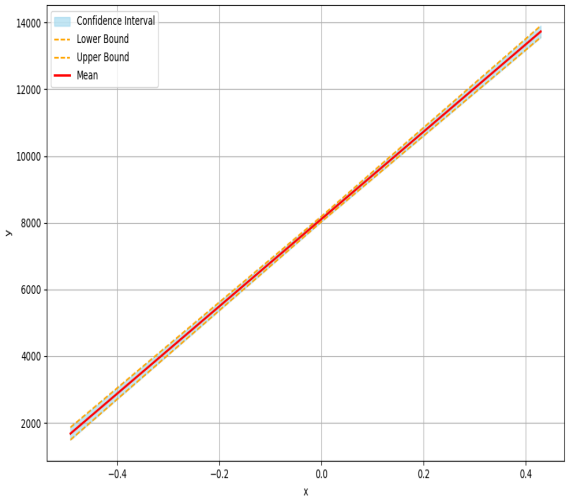

Рис. 1: Коридор совместных зависимостей для внутренней оценки медианой Крейновича

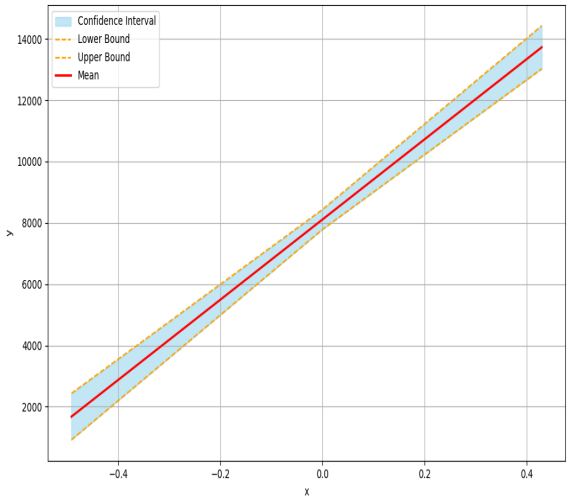

Рис. 2: Коридор совместных зависимостей для внешней оценки медианой Крейновича.

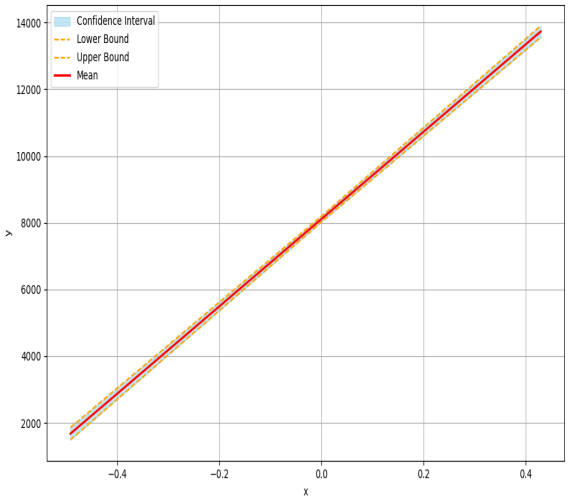

Рис. 3: Коридор совместных зависимостей для внутренней оценки медианой Пролубникова

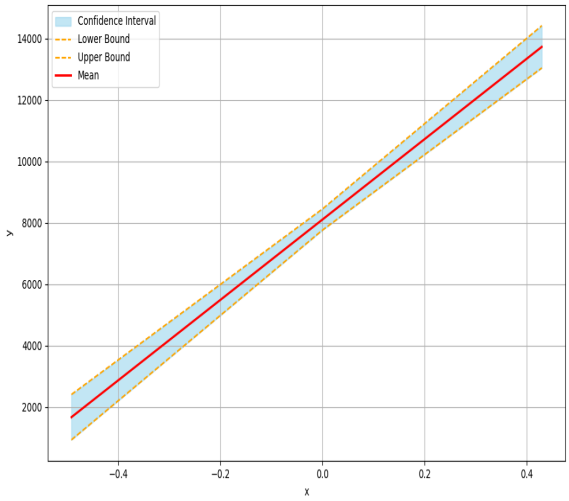

Рис. 4: Коридор совместных зависимостей для внешней оценки медианой Пролубникова

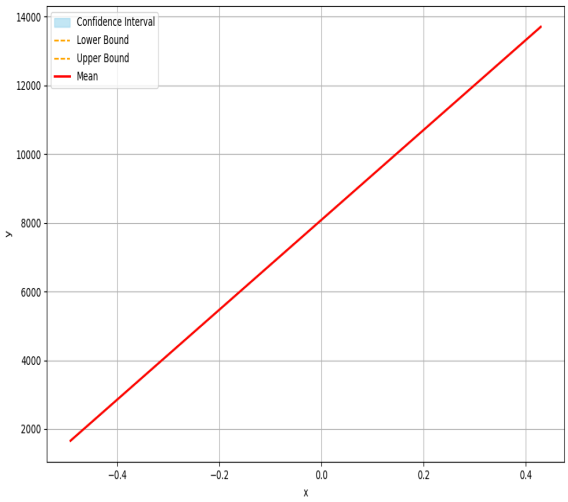

Рис. 5: Коридор совместных зависимостей для внутренней оценки модой

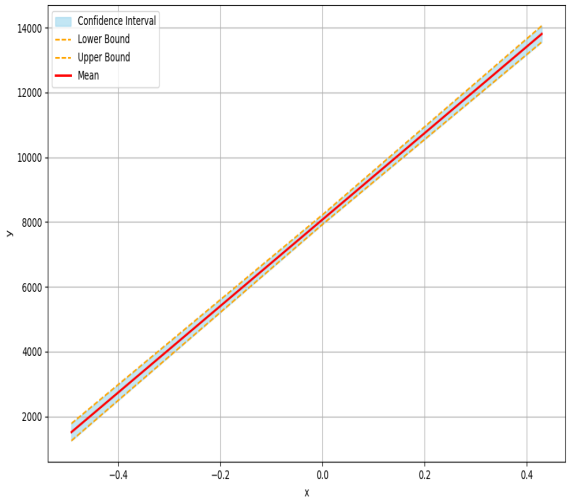

Рис. 6: Коридор совместных зависимостей для внешней оценки модой

5 Заключение

В процессе выполнения лабораторной работы была разработана методика для оценки параметров линейной регрессии на основе интервальных данных. Основные достижения включают:

- Создан алгоритм для вычисления внутренних и внешних оценок параметров линейной регрессии, что позволяет учитывать неопределённость в исходных данных
- Получены интервальные оценки параметров β_0 и β_1, которые демонстрируют диапазон возможных значений параметров регрессии.
- Построены коридоры совместных зависимостей, визуализирующие интервальные решения и способствующие анализу устойчивости модели

Полученные результаты свидетельствуют о том, что предложенный подход обеспечивает более точное моделирование зависимостей в данных, принимая во внимание возможные вариации и ошибки. Это особенно актуально в тех областях, где точность измерений может изменяться, и требуется надежная оценка параметров модели.

6 Литература

1. А.Н.Баженов. Интервальный анализ. Основы теории и учебные примеры. СПБПУ. 2020.

2. J. Rohn — «Enclosing solutions of overdetermined systems of linear interval equations», Reliable Computing 2 (1996), 167-171.