Санкт-Петербургский Политехнический Университет имени Петра Великого
Институт Прикладной Математики и Механики
Кафедра "Прикладная Математика"

Отчет по лабораторным работам №6

по дисциплине
"Математическая Статистика"

Выполнил студент:
Тырыкин Я. А.
группа 5030102/80401
Проверил:
к.ф.-м.н., доцент
Баженов А. Н.

Санкт-Петербург
2021

Список иллюстраций

1 Постановка задачи

Найти оценки коэффициентов линейной регрессии $y_i = a + bx_i + e_i$, используя 20 точек на отрезке $[-1.8, 2]$ с равномерным шагом равным 0.2. Ошибку e_i считается нормально распределённой с параметрами (0, 1). В качестве эталонной зависимости берется $y_i = 2 + 2x_i + e_i$. При построении оценок коэффициентов используются два критерия: критерий наименьших квадратов и критерий наименьших модулей. То же самое проделывается для выборки, у которой в значения y_1 и y_{20} вносятся возмущения 10 и -10.

2 Теория

2.1 Простая линейная регрессия

2.1.1 Модель простой линейной регрессии

Регрессионную модель описания данных называют *простой линейной регрессией*, если

$$y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, ..., n \tag{1}$$

где $x_i, ..., x_n$ - заданные числа (значения фактора); $y_i, ..., y_n$ - наблюдаемые значения отклика; $\epsilon_0, ..., \epsilon_n$ - независимые, нормально распределенные $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); $\beta_0, ..., \beta_n$ - неизвестные параметры, подлежащие оцениванию.

В модели (1) отклик y зависит зависит от одного фактора x, и весь разброс экспериментальных точек объясняется только погрешностями наблюдений (результатов измерений) отклика x. Погрешности результатов измерения x в этой модели полагают существенно меньшими погрешностей результатов измерений y, так что ими можно пренебречь.

2.1.2 Метод наименьших квадратов

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распрстранённых подходов заключается в следующем: вводится мера (критерий) рассогласования отклика регрессионной функции, и оценки параметров регрессии определяются так, чтобы сделать это рассогласование наименьшим. Достаточно простые расчётные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции (сумма квадратов остатков):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2 \to \min_{\beta_0, \beta_1} \tag{2}$$

Задача минимизации квадратичного критерия (2) носит название задачи *метода наименьших квадратов* (МНК), а оценки $\widehat{\beta_0}$, $\widehat{\beta_1}$ параметров β_0 и β_1, реализующие минимум критерия (2) называют *МНК-оценками*. Данный метод, вообще говоря, может применяться для аппроксимации заданного набора экспериментальных данных линейной комбинацией линейно независимых функций, размер которой не превосходит мощности множества данных (в случае равенства получаем интерполяцию).

Оптимальность

Критерием оптимальности подобранной аппроксимации является 2 l^2-норма, точнее, для простоты вычисления, её квадрат:

$$\sum_{i=1}^{n} ||\lambda_i f_i(\{x_n\}) - \{y_n\}||_{l^2}^2 \to_{\lambda_i} min \tag{3}$$

Минимум ищется по коэффициентам линейной комбинации, исходя из критерия равенства нулю градиента и положительной определённости Якобиана.

2.1.3 Расчетный формулы для МНК-оценок

МНК-оценки параметров β_0 и β_1 находятся из условия обращения функции $Q(\beta_0, \beta_1)$ в минимум.

Для нахождения МНК-оценок $\widehat{\beta_0}$ и $\widehat{\beta_1}$ выпишем необходимые условия экстремума:

$$\begin{cases} \frac{\delta Q}{\delta \beta_0} = -2 \sum\limits_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i = 0 \\ \frac{\delta Q}{\delta \beta_1} = -2 \sum\limits_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases} \tag{4}$$

Далее для упрощения записи сумм будем опускать индекс суммирования. Из системы 4 получим:

$$\begin{cases} n\widehat{\beta_0} + \widehat{\beta_1} \sum x_i = \sum y_i \\ \widehat{\beta_0} \sum x_i + \widehat{\beta_1} \sum (x_i)^2 = \sum x_i y_i \end{cases} \tag{5}$$

Разделим оба уравнения на n:

$$\begin{cases} \widehat{\beta_0} + \widehat{\beta_1} (\frac{1}{n} \sum x_i) = \frac{1}{n} \sum y_i \\ \widehat{\beta_0} (\frac{1}{n} \sum x_i) + \widehat{\beta_1} (\frac{1}{n} \sum (x_i)^2) = \frac{1}{n} \sum x_i y_i \end{cases} \tag{6}$$

и, используя известные статистические обозначения для выборочных первых и вторых начальных моментов:

$$\overline{x} = \frac{1}{n}\sum x_i \ , \ \overline{y} = \frac{1}{n}\sum y_i \ , \ \overline{x^2} = \frac{1}{n}\sum x_i^2 \ , \ \overline{xy} = \frac{1}{n}\sum x_i y_i \tag{7}$$

По итогу получим:

$$\begin{cases} \widehat{\beta_0} + \widehat{\beta_1}\overline{x} = \overline{y} \\ \widehat{\beta_0}\overline{x} + \widehat{\beta_1}\overline{x^2} = \overline{xy} \end{cases} \tag{8}$$

откуда МНК-оценку $\widehat{\beta_1}$ наклона прямой регрессии находим по формуле Крамера:

$$\widehat{\beta_1} = \frac{\overline{xy} - \overline{x} * \overline{y}}{\overline{x^2} - \overline{y^2}} \tag{9}$$

а МНК-оценку $\widehat{\beta_0}$ определяем непосредственно из первого уравнения системы 8:

$$\widehat{\beta_0} = \overline{y} - \overline{x} * \widehat{\beta_1} \tag{10}$$

2.2 Робастные оценки коэффициентов линейной регрессии

Робастность оценок коэффициентов линейной регрессии (т.е. их устойчивость по отношению к наличию в данных редких, но больших по величине выбросов) может быть обеспечена различными способами. Одним из них является использование метода наименьших модулей вместо метода наименьших квадратов:

Метод наименьших модулей:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_1| \to \min_{\beta_0, \beta_1} \tag{11}$$

Напомним, что использование метода наименьших модулей в задаче оценивания параметра сдвига распределений приводит к оценке в виде выборочной медианы, обладающей робастными свойствами. В отличие от этого случая и от задач метода наименьших квадратов, на практике задача 11 решается численно. Соответствующие процедуры представлены в некоторых современных пакетах программ по статистическому анализу.

Здесь мы рассмотрим простейшую в вычистлительном отношении робастную альтернативу оценкам коэффициентов линейной регрессии по МНК. Для этого сначала запишем выражения для оценок 10 и 9 в другом виде:

$$\begin{cases} \widehat{\beta_1} = \frac{\overline{xy} - \overline{x} * \overline{y}}{\overline{x^2} - \overline{y^2}} = \frac{k_{xy}}{s_x^2} = \frac{k_{xy}}{s_x s_y} * \frac{s_y}{s_x} = r_{xy} * \frac{s_y}{s_x} \\ \widehat{\beta_0} = \overline{y} - \overline{x}\widehat{\beta_1} \end{cases} \tag{12}$$

В формулах 12 заменим выборочные средние $\overline{x}$ и $\overline{y}$ соответственно на робастные выборочные медианы $medx$ и $medy$, среднеквадратические отклонения s_x и s_y на робастные нормированные интерквартильные широты q_x^* и q_y^*, выборочный коэффициент корреляции r_{xy} - на знаковый коэффициент корреляции r_Q:

$$\widehat{\beta_{1R}} = r_Q * \frac{q_y^*}{q_x^*} \tag{13}$$

$$\widehat{\beta_{0R}} = \text{med y} - (\text{med x}) * \widehat{\beta_{1R}} \tag{14}$$

$$r_Q = \frac{1}{n} \sum_{i=1}^{n} sgn(x_i - \text{med x}) sgn(y_i - \text{med y}) \tag{15}$$

$$q_y^* = \frac{y_{(j)} - y_{(l)}}{k_q(n)} \ , \ q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)} \tag{16}$$

$$l = \begin{cases} [\frac{n}{4}] + 1 \text{ при } \frac{n}{4} \text{ дробном} \\ [\frac{n}{4}] \text{ при } \frac{n}{4} \text{ целом} \end{cases} \ , \ j = n - l + 1 \tag{17}$$

$$\text{sgn z} = \begin{cases} 1 \text{ при } z > 0 \\ 0 \text{ при } z = 0 \\ -1 \text{ при } z < 0 \end{cases} \ , \ j = n - l + 1 \tag{18}$$

Уравнение регрессии здесь имеет вид:

$$y = \widehat{\beta_{0R}} + \widehat{\beta_{1R}} x \tag{19}$$

Статистики выборочной медианы и интерквартильной широты обладают робастными свойствами в силу того, что основаны на центральных порядковых статистиках, малочувствительных к большим по величине выбросам в данных. Статистика выборочного знакового коэффициента корреляции робастна, так как знаковая функция $sgnz$ чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии 19 обладает очевидными робастными свойствами устойчивости

к выбросам по координате но она довольно груба.

Оптимальность
Данный метод основан на минимизации l^1 - нормы разности последовательностей полученных кспериментальных данных и значений аппроксимирующей функции.

2.3 Количественная мера оценки качества регрессии

Как уже было сказано метод наименьших квадратов (МНК) минимизирует норму в l^2, а метод наименьших модулей (МНМ) норму в l^1.

Допустим, что мы будем сравнивать между собой полученные оценки для коэффициентов β_0, β_1 как модули разностей полученных значений. Тогда в случае, если для какой-то выборки окажется, что по данному критерию оба метода покажут схожие результаты, то может оказаться, что невязка по l^2-критерию все равно окажется значительно меньше для результатов, полученных с помощью МНК.

Это как раз и является следствием того, что рассмотренные методы минимизирует различные нормы. Обратная ситуация, но уже для l^1-метрики также может иметь место в некоторых случаях.

Соответственно, можно сказать, что l-метрики (в данном случае речь о близости прямых) не позволяют делать однозначных выводов о качестве линейной регрессии в смысле близости искомых коэффициентов аппроксимирующих функций.

3 Модульная структура программы

Лабораторная работа выполнена с применением средств языка Python версии 3.7в среде разработки PyCharm IDE (в частности, с применением встроенных методовбиблиотеки SciPy и MatPlotLib). Исходной код лабораторной работы находится поссылке в приложении к отчёту.

4 Результаты

4.1 Оценки коэффициентов линейной регрессии

4.1.1 Выборка без возмущений

- Критерий наименьших квадратов:

$$\widehat{a} \approx 1.73 \ , \ \widehat{b} \approx 1.74 \tag{20}$$

- Критерий наименьших модулей:

$$\widehat{a} \approx 2.31 \ , \ \widehat{b} \approx 2.27 \tag{21}$$

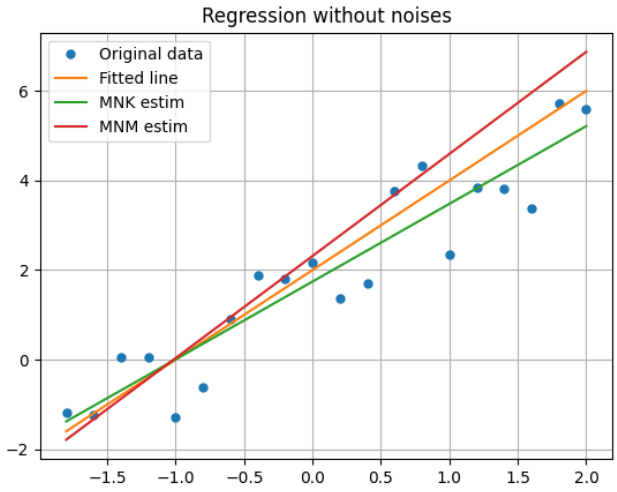

Рис. 1: Выборка без возмущений

4.1.2 Выборка с возмущениями

- Критерий наименьших квадратов:

$$\widehat{a} \approx 2.04 \ , \ \widehat{b} \approx 0.73 \tag{22}$$

- Критерий наименьших модулей:

$$\widehat{a} \approx 2.31 \ , \ \widehat{b} \approx 2.21 \tag{23}$$

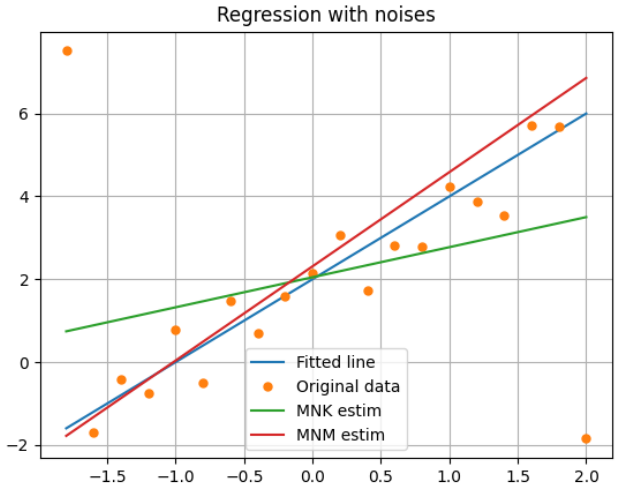

Рис. 2: Выборка с возмущениями

5 Обсуждение

5.1 Оценки коэффициентов линейной регрессии

По полученным результатам (и, в частности, по их графической иллюстрации) можно сказать, что в целом оба критерия (МНК и МНМ) с примерно равноценной точностью оценивают коэффициенты линейной регрессии при отсутствии погрешностей в множестве откликов исследуемой системы. Однако при возникновении последних (особенно при сильной зашумленности данных), критерий МНК начинает работать значительно менее точно, что отчетливо видно по значению $\widehat{b}$ и углу наклона соответствующей кривойю Это говорит о его высокой чувствительности к помехам. Критерий МНМ же в условиях второго эксперимента не теряет своей работоспособности.

6 Ресурсы

Код программы находится на сайте Github по следующей ссылке:

Ветка с исходным кодом и кодом отчета