Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики

Кафедра «Телематика (при ЦНИИ РТК)»

Отчет по лабораторной работе

Простая линейная регрессия

По дисциплине «Теория вероятностей и Математическая статистика»

Выполнил
Студент гр. 3630201/80101 _____ М. Д. Маляренко

Руководитель
к.ф.-м.н., доцент _____ А. Н. Баженов

«___»_____ 2020г.

Санкт-Петербург
2020

Содержание

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. Ошибка e_i распределена по стандартному нормальному закону $N(0, 1)$. В качестве эталонной зависимости взять $y_i = 2 + 2x_i + e_i$. При построении оценок коэффициентов использовать два критерия:

1. Критерий наименьших квадратов
2. Критерий наименьших модулей

Проделать то же самое 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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

где $x_1, \ldots, x_n$ — заданные числа (значения фактора); $y_1, \ldots, y_n$ — наблюдаемые значения отклика; $\varepsilon_1, \ldots, \varepsilon_n$ — независимые, нормально распределённые $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); β_0, β_1 — неизвестные параметры, подлежащие оцениванию.

2.1.2 Метод наименьших квадратов

Метод наименьших квадратов (МНК) [1]:

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

2.1.3 Расчётные формулы для МНК-оценок

МНК-оценки параметров β_0 и β_1 [1]:

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \tag{3}$$

$$\hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1 \tag{4}$$

2.2 Робастные оценки коэффициентов линейной регрессии

Метод наименьших модулей [1]:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0, \beta_1}. \tag{5}$$

$$\hat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*}, \tag{6}$$

$$\hat{\beta}_{0R} = med\ y - \hat{\beta}_{1R} med\ x, \tag{7}$$

$$r_Q = \frac{1}{n} \sum_{i=1}^{n} sgn(x_i - med\ x) sgn(y_i - med\ y), \tag{8}$$

$$q_y^* = \frac{y_j - y_l}{k_q(n)}, \qquad q_x^* = \frac{x_j - x_l}{k_q(n)} \tag{9}$$

$$l = \begin{cases} [n/4] + 1 \ \text{ при } \ n/4 \ \text{ дробном,} \\ n/4 \ \text{ при } \ n/4 \ \text{ целом.} \end{cases}$$

$$j = n - l + 1.$$

$$sgn\ z = \begin{cases} 1 \text{ при } z > 0, \\ 0 \text{ при } z = 0, \\ -1 \text{ при } z < 0. \end{cases}$$

Уравнение регрессии здесь имеет вид

$$y = \hat{\beta}_{0R} + \hat{\beta}_{1R} x. \tag{10}$$

3 Реализация

Расчёты и построение графиков производились в среде аналитических вычислений Maxima с графической оболочкой wxMaxima. Для нахождения параметров $\beta_0, \beta_1, \hat{\beta}_0, \hat{\beta}_1$ по формулам (3), (4), (6), (7) были написаны функции `LSM` для МНК и `LMM` для МНМ. Исходный код скрипта для Maxima представлен в репозитории на GitHub. Графики были построены с помощью интегрированной утилиты `gnuplot`.

4 Результаты

4.1 Выборка без возмущений

В результате оценки параметров линейной регрессии для выборки без возмущений были получены следующие значения коэффициентов:

- МНК: $\hat{\beta}_0 \approx 2.05 \quad \hat{\beta}_1 \approx 2.06$
- МНН: $\hat{\beta}_{0R} \approx 1.65, \quad \hat{\beta}_{1R} \approx 1.54$

На Рис. 1 представлен график модели, точки выборки, а также графики линейной регрессии с коэффициентами вычисленными по МНК и МНН.

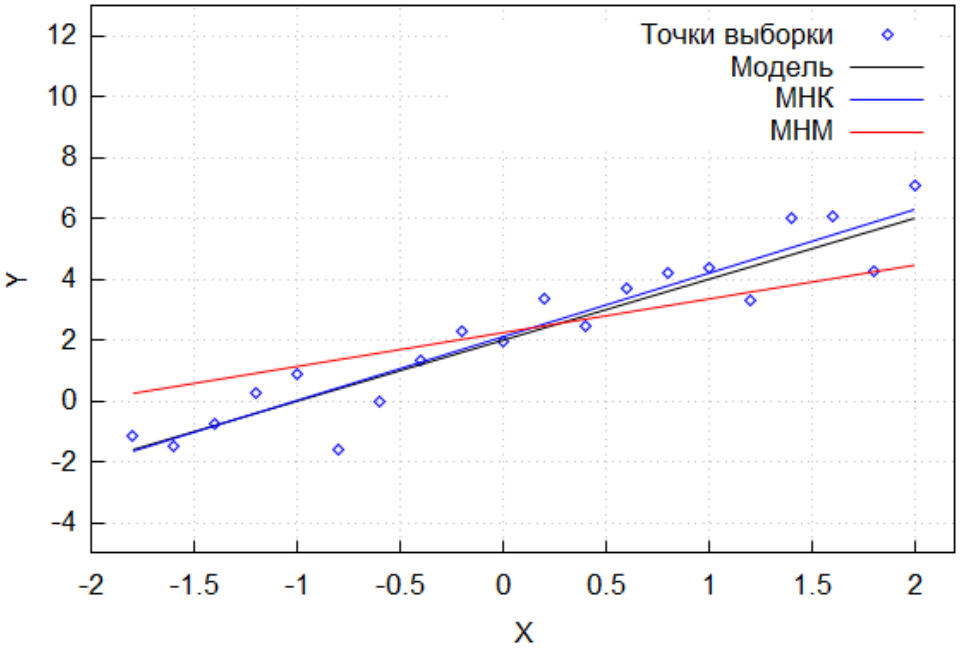

Рис. 1: Выборка без возмущений

4.2 Выборка с возмущениями

В результате оценки параметров линейной регрессии для выборки с возмущениями в крайних элементах были получены следующие значения коэффициентов:

- МНК: $\hat{\beta}_0 \approx 1.97, \quad \hat{\beta}_1 \approx 0.73$
- МНН: $\hat{\beta}_{0R} \approx 2.04, \quad \hat{\beta}_{1R} \approx 1.68$

На Рис. 2 представлен график модели, точки выборки, а также графики линейной регрессии с коэффициентами вычисленными по МНК и МНН.

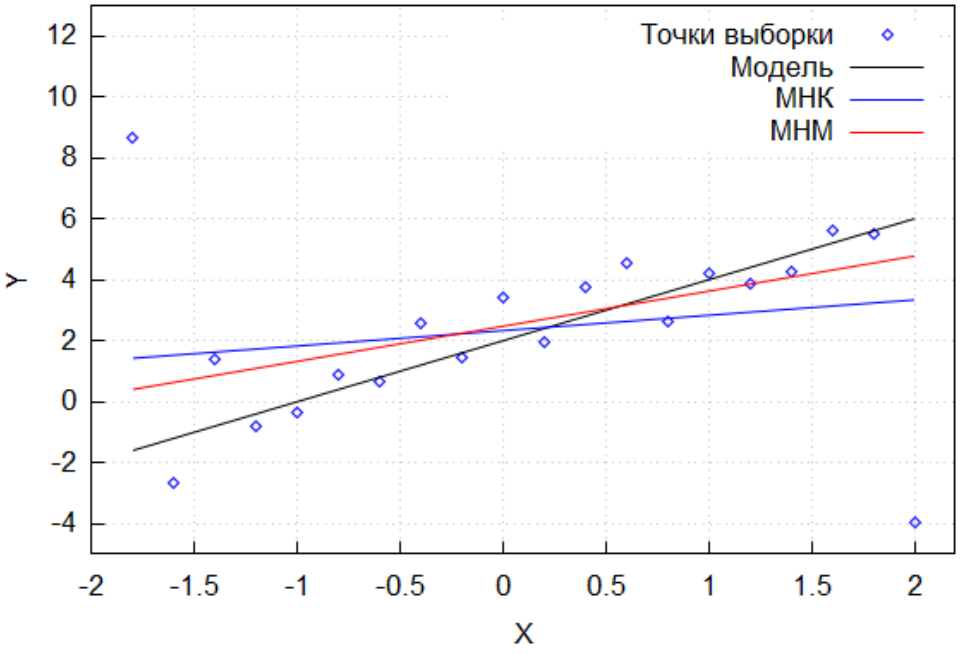

Рис. 2: Выборка с возмущениями

Заключение

В результате лабораторной работы были вычислены коэффициенты линейной регрессии по методам наименьших квадратов и наименьших модулей. Как видно из оценки коэффициентов регрессии а также по графикам, метод наименьших квадратов более точен на выборке без возмущений, но уступает по точности методу наименьших модулей на выборке с редкими, но значительными выбросами.

Можно сделать вывод, что для выборок с небольшими выбросами для нахождения коэффициентов линейной регрессии предпочтительнее использовать МНК, а для выборок с большими выбросами следует использовать МНМ как менее точный, но более устойчивый.

Список литературы

[1] Теоретическое приложение к лабораторным работам №5-8 по дисциплине «Математическая статистика». – СПб.: СПбПУ, 2020. – 22 с

Приложение А. Репозиторий с исходным кодом

Исходный код скрипта для среды аналитических вычислений Maxima находится в репозитории GitHub – URL `https://github.com/malyarenko-md/TeorVer`