Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Математическая статистика

Отчёт по курсовой работе

Работу выполнил:
П. П. Филиппов,
А. М. Бирюков
Группа:
5030102/10101
Преподаватель:
А. Н. Баженов

Санкт-Петербург
2024

Содержание

1. Постановка задачи **3**

2. Теория **3**

2.1. Простая линейная регрессия . . . 3

2.1.1. Модель простой линейной регрессии . . . 3

2.1.2. Метод наименьших квадратов . . . 3

2.1.3. Расчётные формулы для МНК-оценок . . . 4

2.2. Робастные оценки коэффициентов линейной регрессии . . . 5

3. Бокс-плот Тьюки **6**

4. Ход работы. **6**

5. Результаты **6**

5.1. Анализ исходных данных . . . 6

5.2. Сравнение МНК и МНМ . . . 9

6. Обсуждение **11**

1. Постановка задачи

Требуется изучить характер оценок при вариации данных или методов (параметров) оценивания.

Для этого оценим коэффициенты линейной регрессии и провести исследования для следующих методова оценивания:

1. Метод наименьших квадратов (МНК).
2. Метод наименьших модулей (МНМ).

Данные для исследования предоставлены преподавателем.

2. Теория

2.1. Простая линейная регрессия

2.1.1. Модель простой линейной регрессии

Регрессионную модель описания данных называют простой линейной регрессией, если

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

где $x_1, ..., x_n$– заданные числа (значения фактора); $y_1, ...y_n$– наблюдаемые значения отклика; $\varepsilon_1, ..., \varepsilon_n$– независимые, нормально распределенные $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); β_0, β_1– 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 y. Погрешности результатов измерений x в этой модели полагают существенно меньшими погрешностей результатов измерений y, так что ими можно пренебречь [1, с. 507].

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 и 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):

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

Задача минимизации квадратичного критерия $Q(\beta_0, \beta_1)$ носит название задачи метода наименьших квадратов (МНК), а оценки $\hat{\beta}_0, \hat{\beta}_1$ параметров β_0, β_1, реализующие минимум критерия $Q(\beta_0, \beta_1)$, называют МНК-оценками [1, с. 508].

2.1.3. Расчётные формулы для МНК-оценок

МНК-оценки параметров $\hat{\beta}_0, \hat{\beta}_1$ находятся из условия обращения функции $Q(\beta_0, \beta_1)$ в минимум.
Для нахождения МНК-оценок $\hat{\beta}_0, \hat{\beta}_1$ выпишем необходимые условия экстремума

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \end{cases}$$

Далее для упрощения записи сумм будем опускать индекс суммирования. Из системы (2.1.3) получим:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

Разделим оба уравнения на n:

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1(\frac{1}{n}\sum x_i) = \frac{1}{n}\sum y_i \\ \hat{\beta}_0(\frac{1}{n}\sum x_i) + \hat{\beta}_1(\frac{1}{n}\sum x_i^2) = \frac{1}{n}\sum x_i y_i \end{cases}$$

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

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

получим

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \\ \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \bar{x^2} = \bar{xy}, \end{cases}$$

откуда МНК-оценку $\hat{\beta}_1$ наклона прямой регрессии находим по формуле Крамера

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

а МНК-оценку $\hat{\beta}_0$ определяем непосредственно из первого уравнения системы (2.1.3):

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

Заметим, что определитель системы (2.1.3):

$$\bar{x^2} - (\bar{x})^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = s_x^2 > 0,$$

если среди значений $x_1, ..., x_n$ есть различные, что и будем предполагать.
Доказательство минимальности функции $Q(\beta_0, \beta_1)$ в стационарной точке проведём с помощью известного достаточного признака экстремума функции двух переменных. Имеем:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2 = 2n\bar{x^2}, \frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} = 2\sum x_i = 2n\bar{x}$$

$$\triangle = \frac{\partial^2 Q}{\partial \beta_0^2}\cdot\frac{\partial^2 Q}{\partial \beta_1^2} - (\frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0})^2 = 4n^2\bar{x^2} - 4n^2(\bar{x})^2 = 4n^2\left[\bar{x^2} - (\bar{x})^2\right] = 4n^2\left[\frac{1}{n}\sum (x_i - \bar{x})\right] = 4n^2 s_x^2 > 0.$$

Этот результат вместе с условием $\frac{\partial^2 Q}{\partial \beta_0^2} = 2n > 0$ означает, что в стационарной точке функция Q имеет минимум [1, с. 508-511].

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:

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

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 (2.2) 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 (2.1.3) и (2.1.3) в другом виде:

$$\begin{cases} \hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{k_{xy}}{s_x^2} = \frac{k_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = r_{xy}\frac{s_y}{s_x} \\ \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \end{cases}$$

В формулах (2.2) заменим выборочные средние $\bar{x}$ и $\bar{y}$ 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:

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

$$\hat{\beta}_{0R} = medy - \hat{\beta}_{1R} medx,$$

$$r_Q = \frac{1}{n}\sum_{i=1}^{n} sgn(x_i - medx) sgn(y_i - medy),$$

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

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

$$j = n - l + 1$$

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

Уравнение регрессии здесь имеет вид

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

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 (2.2) обладает очевидными робастными свойствами устойчивости к выбросам по координате y, но она довольно груба [1, с. 518-519].

3. 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 выбросов). Длину «усов» определяют разность первого квартиля и полутора межквартильных расстояний и сумма третьего квартиля и полутора межквартильных 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 x, такие что:

$$\left[\begin{array}{l} x < X_1^{\text{т}} \\ x > X_2^{\text{т}} \end{array}\right.$$

4. Ход работы.

Курсовая работа выполнена на языке программирования Python.
В ходе работы были использованы следующие библиотеки:

- numpy
- mathplotlib
- pandas
- scipy

GitHub репозиторий:github

5. Результаты

5.1. Анализ исходных данных

Из графика бокс-плотов Тьюки для различных ячеек остановки наблюдаются следующие характеристики:

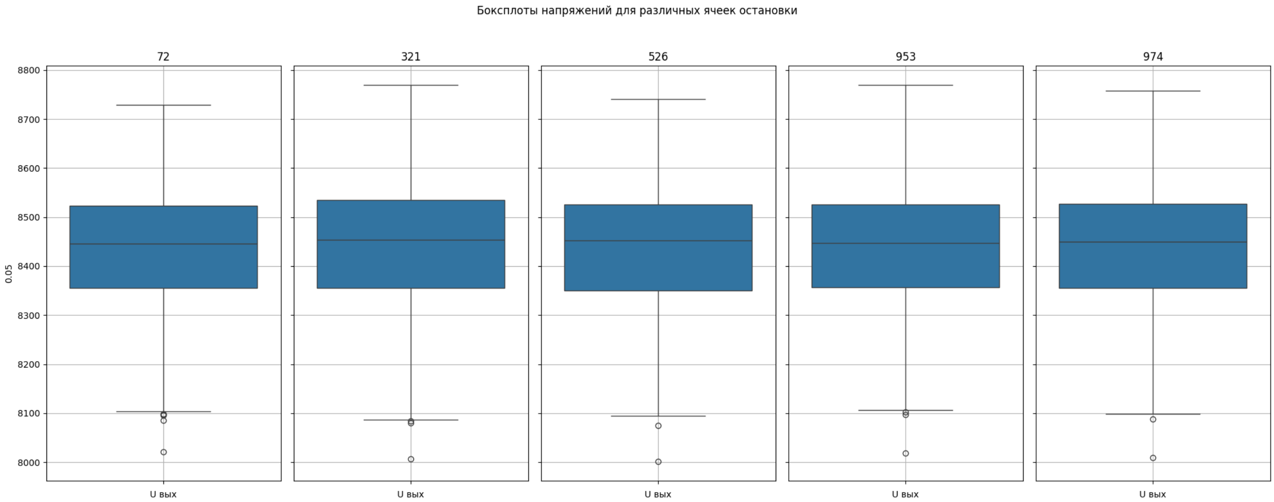

Рисунок 5.1. Бокс-плот Тьюки для исходных данных при значении $U_{in} = 0.05$ при различных ячейках остановки

- Схожие значения медиан и квартилей (за исключением значения ячейки остановки $sp = 72$)
- Все группы данных имеют выбросы снизу (т.е. данные имеют схожий характер)
- Усы расположены симметрично относительно ящиков, однако их длины различны для разных значений sp.

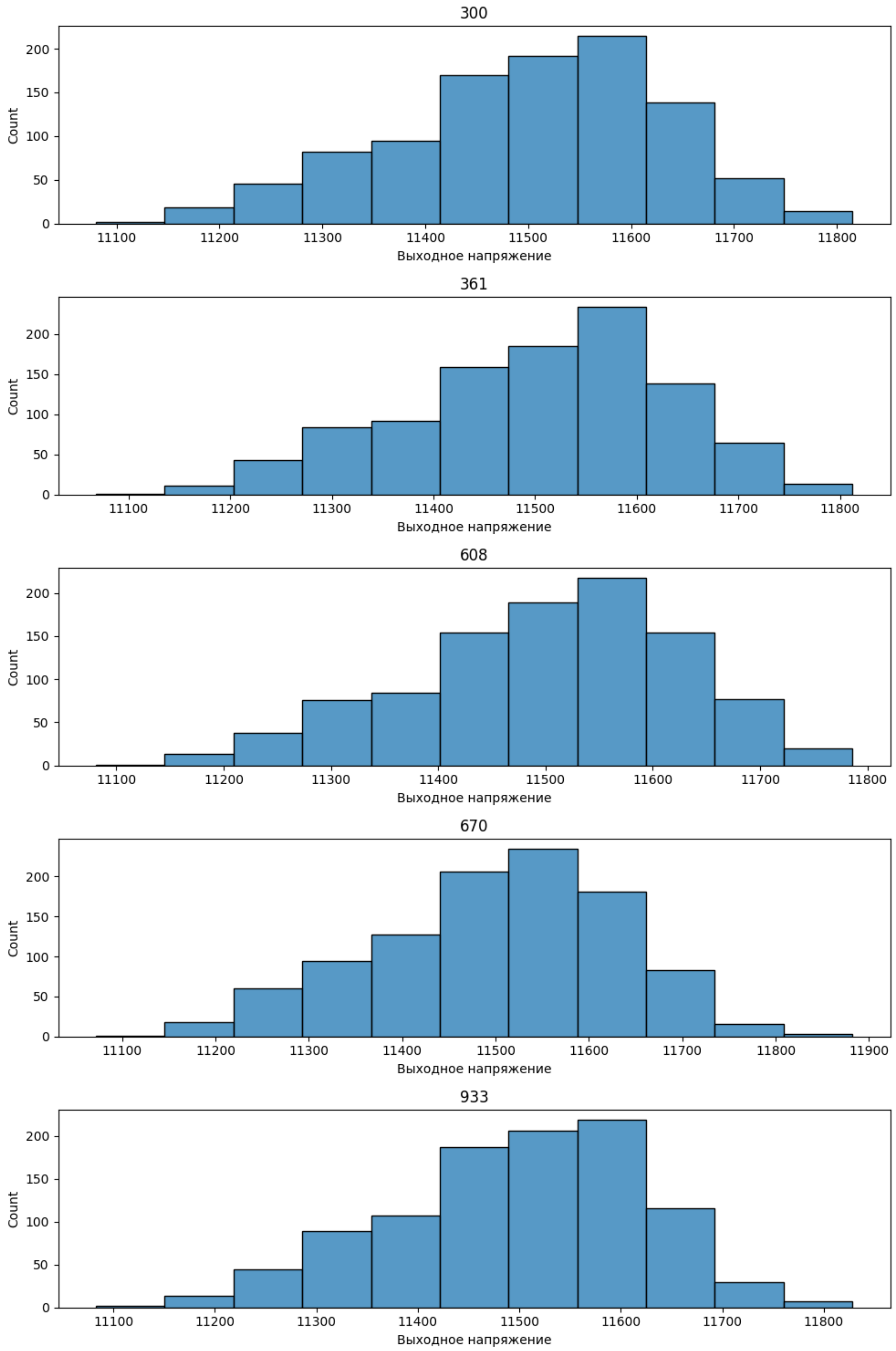

Рисунок 5.2. Гистограммы исходных данных при значении $U_{in} = 0.05$

Из гистограмм видно, что данные имеют нормальное распределение со смещением в сторону положительной оси. Это означает, что большинство значений сосредоточены вокруг среднего, но есть незначительное количество значений, которые вытягивают распределение вправо.

5.2. Сравнение МНК и МНМ

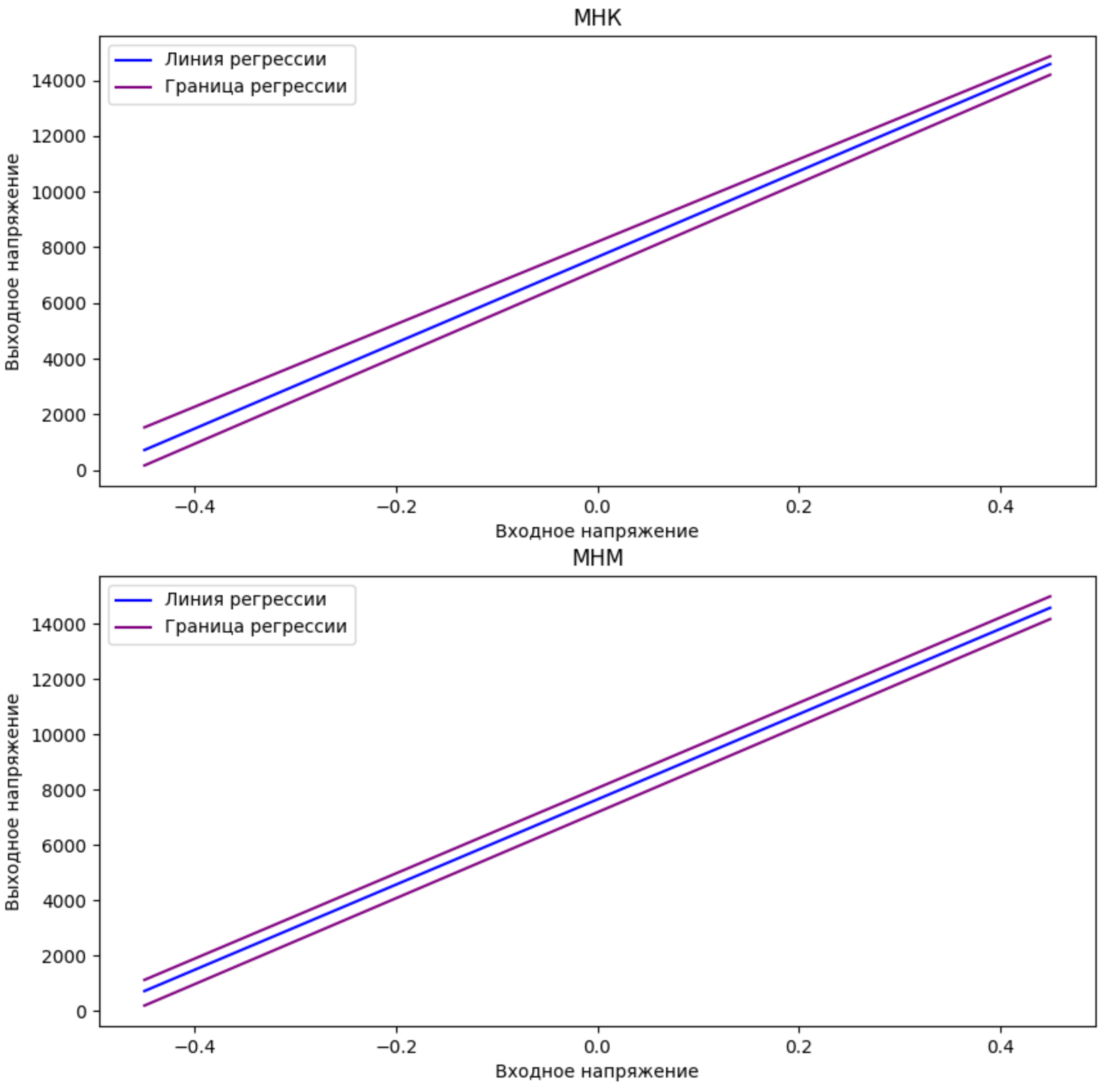

Рисунок 5.3. График линейной регрессии для МНК и МНМ

Таблица 5.1

МНК	$\hat{\beta}_0$	$\hat{\beta}_1$
	7734.484	15522.993
МНМ	$\hat{\alpha}_0$	$\hat{\alpha}_1$
	7735.452	15515.655

Оценка параметров a и b методами МНК и МНМ

	$\hat{\beta}_0$	$\hat{\beta}_1$
Верхняя граница	8274.0	14837.939
Нижняя граница	7253.364	15698.545

Таблица 5.2

Коридор ошибок для МНК

	$\hat{\alpha}_0$	$\hat{\alpha}_1$
Верхняя граница	8115.286	15462.857
Нижняя граница	7266.5	15550.0

Таблица 5.3

Коридор ошибок для МНМ

6. Обсуждение

1. Из графика бокс-плотов Тьюки для различных ячеек остановки наблюдаются следующие характеристики: схожие значения медиан и квартилей (за исключением значения ячейки остановки $sp = 72$), все группы данных имеют выбросы снизу (т.е. данные имеют схожий характер), а также усы расположены симметрично относительно ящиков, однако их длины различны для разных значений sp.

2. Из гистограмм можно сделать вывод о том, что имеют нормальное распределение со смещением в сторону положительной оси.

3. Как исследования методом наименьших квадратов (МНК), так и методом наименьших модулей (МНМ) выявили результаты, которые статистически нельзя различить между собой. Это свидетельствует о том, что оба подхода равноценны и одинаково эффективны в решении данной проблемы. Ни один из методов не выделяется как более предпочтительный при использоавнии этих данных.

4. Кроме того, построенные коридоры ошибок также указывает на их эквивалентность в контексте данной задачи.