Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

Отчет
по лабораторным работам №5-8
по дисциплине
«Математическая статистика»

Выполнил студент:
Иванова А.С.
группа: 5030102/00101

Проверил:
к.ф.-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2023 г.

Содержание

Список иллюстраций

Список таблиц

1 Постановка задачи

1. Сгенерировать двумерные выборки размерами 20, 60, 100 для нормального двумерного распределения $N(x, y, 0, 0, 1, 1, \rho)$. Коэффициент корреляции ρ взять равным 0, 0.5, 0.9. Каждая выборка генерируется 1000 раз и для неё вычисляются: среднее значение, среднее значение квадрата и дисперсия коэффициентов корреляции Пирсона, Спирмена и квадратного коэффициента корреляции. Повторить все вычисления для смеси нормальных распределений:

$$f(x, y) = 0.9N(x, y, 0, 0, 1, 1, 0.9) + 0.1N(x, y, 0, 0, 10, 10, -0.9) \tag{1}$$

   Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

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

3. Сгенерировать выборку объемом 100 элементов для номрального распределения $N(x, 0, 1)$. По сгенерированной выборке оценить параметры μ и σ нормального закона методом максимального правдоподобия. В качестве основной гипотезы H_0 будем считать, что сгенерированное распределение имеет вид $N(x, \hat{\mu}, \hat{\sigma})$. Проверить основную гипотезу, использу критерий согласия χ^2. В качестве уровня значимости взять $\alpha = 0.05$. Привести таблицу вычислений χ^2.

   Исследовать точность (чувствительность) критерия χ^2- сгенерировать выборки равномерного распределения и распределения Лапласа малого объема (например, 20 элементов). Проверить их на нормальность.

4. Для двух выборок размерами 20 и 100 элементов, сгенерированных согласно нормальному закону $N(x, 0, 1)$, для параметров положения и масштаба построить асимптотически нормальные интервальные оценки на основе точечных оценок метода максимального правдоподобия и классические интервальные оценки на основе статистик χ^2 и Стьюдента. В качестве параметра надёжности взять $\gamma = 0.95$.

2 Теория

2.1 Двумерное нормальное распределение

Двумерная случайная величина (X, Y) называет распределенной нормально (нормальной), если ее плотность вероятности определена формулой

$$N(x, y, \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times exp\Big\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\bar{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2}\right]\Big\} \tag{2}$$

Компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями $\overline{x}, \overline{y}$ и средними квадратическими отклонениями σ_x, σ_y соответственно. Параметр ρ называется коэффициентом корреляции.

2.2 Корреляционный момент (ковариация) и коэффициент корреляции

Корреляционным моментом, иначе ковариацией, двух случайных величин X и Y называется математическое ожидание произведения отклонений этих случайных величин от их математических ожиданий.

$$K = cov(X, Y) = M[(X - \bar{x})(Y - \bar{y})] \tag{3}$$

Коэффициентом корреляции ρ двух случайных величин X и Y называется отношение их корреляционного момента к произведению их средних квадратических отклонений:

$$\rho = \frac{K}{\sigma_x\sigma_y} \tag{4}$$

Коэффициент корреляции — это нормированная числовая характеристика, являющаяся мерой близости зависимости между случайными величинами к линейной.

2.3 Выборочные коэффициенты корреляции

2.3.1 Выборочный коэффициент корреляции Пирсона

Пусть по выборке значений $\{x_i, y_i\}_{i=1}^n$ двумерной с.в. (X,Y) требуется оценить коэффициент корреляции $\rho = \dfrac{cov(X, Y)}{\sqrt{DXDY}}$. Естественной оценкой для ρ служит его статистический аналог в виде выборочного коэффициента корреляции, предложенного К.Пирсоном, —

$$r = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2\frac{1}{n}\sum(y_i - \bar{y})^2}} = \frac{K}{s_X s_Y}, \tag{5}$$

где K, s_X^2, s_Y^2 — выборочные ковариация и дисперсии с.в. X и Y.

2.3.2 Выборочный квадрантный коэффициент корреляции

Кроме выборочного коэффициента корреляции Пирсона, существуют и другие оценки степени взаимосвязи между случайными величинами. К ним относится выборочный квадрантный коэффициент корреляции

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \tag{6}$$

где n_1, n_2, n_3, n_4– количества точек с координатами x_i, y_i, попавшими соответственно в I, II, III, IV квадранты декартовой системы с осями $x' = x - medx$, $y' = y - medy$ и с центром в точке с координатами $(medx,\ medy)$

2.3.3 Выборочный коэффициент ранговой корреляции Спирмена

На практике нередко требуется оценить степень взаимодействия между качественными признаками изучаемого объекта. Качественным называется признак, который нельзя измерить точно, но который позволяет сравнивать изучаемые объекты между собой и располагать их в порядке убывания или возрастания их качества. Для этого объекты выстраиваются в определённом порядке в соответствии с рассматриваемым признаком. Процесс упорядочения называется ранжированием, и каждому члену упорядоченной последовательности объектов присваивается ранг, или порядковый номер. Например, объекту с наименьшим значением признака присваивается ранг 1, следующему за ним объекту — ранг 2, и т.д. Таким образом, происходит сравнение каждого объекта со всеми объектами изучаемой выборки.
Если объект обладает не одним, а двумя качественными признаками — переменными X и Y , то для исследования их взаимосвязи используют выборочный коэффициент корреляции между двумя последовательностями рангов этих признаков.
Обозначим ранги, соотвествующие значениям переменной X, через u, а ранги, соотвествующие значениям переменной Y, — через v.
Выборочный коэффициент ранговой корреляции Спирмена определяется как выборочный коэффициент корреляции Пирсона между рангами u,v переменных X,Y :

$$r_S = \frac{\frac{1}{n}\sum (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n}\sum (u_i - \bar{u})^2 \frac{1}{n}\sum (v_i - \bar{v})^2}}, \tag{7}$$

где $\bar{u} = \bar{v} = \frac{1+2+...+n}{n} = \frac{n+1}{2}$ — среднее значение рангов.

2.4 Эллипсы рассеивания

Рассмотрим поверхность распределения, изображающую функцию (1). Она имеет вид холма, вершина которого находится над точкой $(\bar{x}, \bar{y})$.
В сечении поверхности распределения плоскостями, параллельными оси $N(x, y, \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho)$, получаются кривые, подобные нормальным кривым распределения. В сечении поверхности распределения плоскостями, параллельными плоскости xOy, получаются эллипсы. Напишем уравнение проекции такого эллипса на плоскость xOy:

$$\frac{(x-\bar{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2} = const \tag{8}$$

Уравнение эллипса 8 можно проанализировать обычными методами аналитической геометрии. Применяя их, убеждаемся, что центр эллипса 8 находится в точке с координатами $(\bar{x}, \bar{y})$; что касается направления осей симметрии эллипса, то они составляют с осью Ox углы, определяемые уравнением

$$tg(2\alpha) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} \tag{9}$$

Это уравнение дает два значения углов: α и α_1, различающиеся на $\frac{\pi}{2}$.
Таким образом, ориентация эллипса 8 относительно координатных осей находится в прямой зависимости

от коэффициента корреляции ρ системы (X, Y); если величины не коррелированны (т.е. в данном случае и независимы), то оси симметрии эллипса параллельны координатным осям; в противном случае они составляют с координатными осями некоторый угол.
Пересекая поверхность распределения плоскостями, параллельными плоскости xOy, и проектируя сечения на плоскость xOy мы получим целое семейство подобных и одинаково расположенных эллипсов с общим центром $(\bar{x}, \bar{y})$. Во всех точках каждого из таких эллипсов плотность распределения $N(x, y, \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho)$ постоянна. Поэтому такие эллипсы называются эллипсами равной плотности или, короче эллипсами рассеивания. Общие оси всех эллипсов рассеивания называются главными осями рассеивания.

2.5 Простая линейная регрессия

2.5.1 Модель простой линейной регрессии

Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1..n \tag{10}$$

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 (10) 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.

2.5.2 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):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to \min_{\beta_0, \beta_1} \tag{11}$$

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.

2.5.3 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

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases} \tag{12}$$

Далее для упрощения записи сумм будем опускать индекс суммирования. Из системы (12) получим:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i \end{cases} \tag{13}$$

Разделим оба уравнения на n:

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 (\frac{1}{n} \sum x_i) = \frac{1}{n} \sum y_i \\ \hat{\beta}_0 (\frac{1}{n} \sum x_i) + \hat{\beta}_1 (\frac{1}{n} \sum x_i^2) = \frac{1}{n} \sum x_i y_i \end{cases} \tag{14}$$

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

$$\bar{x} = \frac{1}{n} \sum x_i, \bar{y} = \frac{1}{n} \sum y_i, \bar{x^2} = \frac{1}{n} \sum x_i^2, \bar{xy} = \frac{1}{n} \sum x_i y_i, \tag{15}$$

получим

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \\ \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \bar{x^2} = \bar{xy}, \end{cases} \tag{16}$$

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

$$\hat{\beta}_1 = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - (\bar{x})^2} \tag{17}$$

а МНК-оценку $\hat{\beta}_0$ определяем непосредственно из первого уравнения системы (16):

$$\hat{\beta}_0 = \bar{y} - \bar{x} \hat{\beta}_1 \tag{18}$$

Заметим, что определитель системы (16):

$$\bar{x^2} - (\bar{x})^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = s_x^2 > 0, \tag{19}$$

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:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \frac{\partial^2 Q}{\partial \beta_1^2} = 2 \sum x_i^2 = 2n\bar{x^2}, \frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} = 2 \sum x_i = 2n\bar{x} \tag{20}$$

$$\triangle = \frac{\partial^2 Q}{\partial \beta_0^2} \cdot \frac{\partial^2 Q}{\partial \beta_1^2} - (\frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0})^2 = 4n^2 \bar{x^2} - 4n^2 (\bar{x})^2 = 4n^2 \left[\bar{x^2} - (\bar{x})^2\right] = 4n^2 \left[\frac{1}{n} \sum (x_i - \bar{x})\right] = 4n^2 s_x^2 > 0. \tag{21}$$

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

2.6 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:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0, \beta_1} \tag{22}$$

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 (22) 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 (18) и (17) в другом виде:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - (\bar{x})^2} = \dfrac{k_{xy}}{s_x^2} = \dfrac{k_{xy}}{s_x s_y} \cdot \dfrac{s_y}{s_x} = r_{xy} \dfrac{s_y}{s_x} \\ \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \end{cases} \tag{23}$$

В формулах (23) 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:

$$\hat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*}, \tag{24}$$

$$\hat{\beta}_{0R} = medy - \hat{\beta}_{1R} medx, \tag{25}$$

$$r_Q = \frac{1}{n} \sum_{i=1}^{n} sgn(x_i - medx) sgn(y_i - medy), \tag{26}$$

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

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

$$j = n - l + 1$$

$$sgn(z) = \begin{cases} 1 \text{ при } z > 0 \\ 0 \text{ при } z = 0 \\ -1 \text{ при } z < 0 \end{cases} \tag{27}$$

Уравнение регрессии здесь имеет вид

$$y = \hat{\beta}_{0R} + \hat{\beta}_{1R}x \tag{28}$$

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 (28) обладает очевидными робастными свойствами устойчивости к выбросам по координате y, но она довольно груба.

2.7 Метод максимального правдоподобия

$L(x_1, ..., x_n, \theta)$ — функция правдоподобия (ФП), представляющая собой совместную плотность вероятности независимых с.в. $x_1, ..., x_n$ и рассматриваемая как функция неизвестного параметра θ:

$$L(x_1, ..., x_n, \theta) = f(x_1, \theta)f(x_2, \theta)...f(x_n, \theta) \tag{29}$$

Определение. Оценкой максимального правдоподобия (о.м.п) будем называть такое значение $\hat{\theta}$ из множества допустимых значений параметра θ, для которого ФП принимает наибольшее значение при заданных $x_1, ..., x_n$:

$$\hat{\theta} = \arg\max_{\theta} L(x_1, ..., x_n, \theta) \tag{30}$$

Система уравнений правдоподобия (в случае дифференцируемости функции правдоподобия):

$$\frac{\partial L}{\partial \theta_k} = 0 \text{ или } \frac{\partial \ln L}{\partial \theta_k} = 0, k = 1, ..m \tag{31}$$

2.8 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Исчерпывающей характеристикой изучаемой случайной величины является её закон распределения. Поэтому естественно стремление исследователей построить этот закон приближённо на основе статистических данных.
Сначала выдвигается гипотеза о виде закона распределения.
После того как выбран вид закона, возникает задача оценивания его параметров и проверки (тестирования) закона в целом.
Для проверки гипотезы о законе распределения применяются критерии согласия. Таких критериев существует много. Мы рассмотрим наиболее обоснованный и наиболее часто используемый в практике — критерий χ^2 (хи-квадрат), введённый К.Пирсоном (1900 г.) для случая, когда параметры распределения известны. Этот критерий был существенно уточнён Р.Фишером (1924 г.), когда параметры распределения оцениваются по выборке, используемой для проверки.
Ограничимся рассмотрением случая одномерного распределения.
Итак, выдвинута гипотеза H_0 о генеральном законе распределения с функцией распределения $F(x)$.
Рассматриваем случай, когда гипотетическая функция распределения $F(x)$ не содержит неизвестных параметров.
Разобьём генеральную совокупность, т.е. множество значений изучаемой случайной величины X на k непересекающихся подмножеств $\Delta_1, \Delta_2, ..., \Delta_k$.
Пусть $p_i = P(X \in \Delta_i), i = 1, ..., k$.

Если генеральная совокупность – вся вещественная ось, то подмножества $\Delta_i = (a_{i-1}, a_i]$ – полуоткрытые промежутки $(i = 2, ..., k-1)$. Крайние промежутки будут полубесконечными: $\Delta_1 = (-\infty, a_1], \Delta_k = (a_{k-1}, +\infty)$. В этом случае $p_i = F(a_i) - F(a_{i-1}); a_0 = -\infty, a_k = +\infty (i = 1, ..., k)$.
Отметим, что $\sum_{i=1}^{k} p_i = 1$. Будем предполагать, что все $p_i > 0$ (i = 1, ... ,k).
Пусть, далее, $n_1, n_2, ..., n_k$ — частоты попадания выборочных элементов в подмножества $\Delta_1, \Delta_2, ..., \Delta_k$ соответственно.
В случае справедливости гипотезы H_0 относительные частоты n_i/n при большом n должны быть близки к вероятностям p_i $(i = 1, ..., k)$, поэтому за меру отклонения выборочного распределения от гипотетического с функцией $F(x)$ естественно выбрать величину

$$Z = \sum_{i=1}^{k} c_i (\frac{n_i}{n} - p_i)^2, \tag{32}$$

где c_i — какие-нибудь положительные числа (веса). К.Пирсоном в качестве весов выбраны числа $c_i = n/p_i$ (i = 1, ... ,k). Тогда получается статистика критерия хи-квадрат К.Пирсона

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i} (\frac{n_i}{n} - p_i)^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \tag{33}$$

которая обозначена тем же символом, что и закон распределения хи-квадрат.
К.Пирсоном доказана теорема об асимптотическом поведении статистики χ^2, указывающая путь её применения.
Теорема К.Пирсона. Статистика критерия χ^2 асимптотически распределена по закону χ^2 с $k-1$ степенями свободы.
Это означает, что независимо от вида проверяемого распределения, т.е. функции $F(x)$, выборочная функция распределения статистики χ^2 при $n \to \infty$ стремится к функции распределения случайной величины с плотностью вероятности

$$f_{k-1}(x) = \begin{cases} 0\ ,\ x \le 0 \\ \frac{1}{2^{\frac{k-1}{2}} \Gamma(\frac{k-1}{2})} x^{\frac{k-3}{2}} e^{-\frac{x}{2}}\ ,\ x > 0 \end{cases} \tag{34}$$

Для прояснения сущности метода χ^2 сделаем ряд замечаний.
Замечание 1. Выбор подмножеств $\Delta_1, \Delta_2, ..., \Delta_k$ и их числа k в принципе ничем не регламентируется, так как $n \to \infty$. Но так как число n хотя и очень большое, но конечное, то k должно быть с ним согласовано. Обычно его берут таким же, как и для построения гистограммы, т.е. можно руководствоваться формулой

$$k \approx 1.72 \sqrt[3]{n} \tag{35}$$

или формулой Старджесса

$$k \approx 1 + 3.3 lgn \tag{36}$$

При этом, если $\Delta_1, \Delta_2, ..., \Delta_k$ — промежутки, то их длины удобно сделать равными, за исключением крайних — полубесконечных.
Замечание 2. (о числе степеней свободы). Числом степеней свободы функции (по старой терминологии) называется число её независимых аргументов. Аргументами статистики χ^2 являются частоты $n_1, n_2, ..., n_k$. Эти частоты связаны одним равенством $n_1 + n_2 + ... + n_k = n$, а в остальном независимы в силу независимости элементов выборки. Таким образом, функция χ^2 имеет $k-1$ независимых аргументов: число

частот минус одна связь. В силу теоремы Пирсона число степеней свободы статистики χ^2 отражается на виде асимптотической плотности $f_{k-1}(x)$.
На основе общей схемы проверки статистических гипотез сформулируем следующее правило.
Правило проверки гипотезы о законе распределения по методу χ^2.
1. Выбираем уровень значимости α.
2. По таблице [3, с. 358] находим квантиль $\chi^2_{1-\alpha}(k-1)$ распределения хи-квадрат с k−1 степенями свободы порядка $1-\alpha$.
3. С помощью гипотетической функции распределения $F(x)$ вычисляем вероятности $p_i = P(X \in \Delta_i)$, $i = 1, ..., k$.
4. Находим частоты n_i попадания элементов выборки в подмножества Δ_i, $i = 1, ..., k$.
5. Вычисляем выборочное значение статистики критерия χ^2:

$$\chi^2_B = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}. \tag{37}$$

6. Сравниваем χ^2_B и квантиль $\chi^2_{1-\alpha}(k-1)$.
а) Если $\chi^2_B < \chi^2_{1-\alpha}(\mathrm{k}-1)$, то гипотеза H_0 на данном этапе проверки принимается.
б) Если $\chi^2_B >= \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 отвергается, выбирается одно из альтернативных распределений, и процедура проверки повторяется.
Замечание 3. Из формулы (33) видим, что веса $c_i = n/p_i$ пропорциональны n, т.е. с ростом n увеличиваются. Отсюда следует, что если выдвинутая гипотеза неверна, то относительные частоты n_i/n не будут близки к вероятностям p_i, и с ростом n величина χ^2_B будет увеличиваться. При фиксированном уровне значимости α будет фиксировано пороговое число - квантиль $\chi^2_{1-\alpha}(k-1)$, поэтому, увеличивая n, мы придём к неравенству $\chi^2_B > \chi^2_{1-\alpha}(k-1)$, т.е. с увеличением объёма выборки неверная гипотеза будет отвергнута.
Отсюда следует, что при сомнительной ситуации, когда $\chi^2_B \approx \chi^2_{1-\alpha}(k-1)$, можно попытаться увеличить объём выборки (например, в 2 раза), чтобы требуемое неравенство было более чётким.
Замечание 4. Теория и практика применения критерия χ^2 указывают, что если для каких-либо подмножеств Δ_i $(i = 1, ..., k)$ условие $np_i \geq 5$ не выполняется, то следует объединить соседние подмножества (промежутки).
Это условие выдвигается требованием близости величин $\frac{(n_i - np_i)}{\sqrt{np_i}}$, квадраты которых являются слагаемыми χ^2 к нормальным $N(0,1)$. Тогда случайная величина в формуле (33) будет распределена по закону, близкому к хи-квадрат. Такая близость обеспечивается достаточной численностью элементов в подмножествах Δ_i.

2.9 Доверительные интервалы для параметров нормального распределения

2.9.1 Доверительный интервал для математического ожидания m нормального распределения

Дана выборка $(x_1, x_2, ..., x_n)$ объёма n из нормальной генеральной совокупности. На её основе строим выборочное среднее $\bar{x}$ и выборочное среднее квадратическое отклонение s. Параметры m и σ нормального распределения неизвестны.
Доказано, что случайная величина

$$T = \sqrt{n-1}\frac{\bar{x} - m}{s} \tag{38}$$

называемая *статистикой Стьюдента*, распределена по закону Стьюдента с $n-1$ степенями свободы. Пусть $f_T(x)$ — плотность вероятности этого распределения. Тогда

$$P\left(-x < \sqrt{n-1}\frac{\bar{x}-m}{s} < x\right) = P\left(-x < \sqrt{n-1}\frac{m-\bar{x}}{s} < x\right) = \\ = \int_{-x}^{x} f_T(t)dt = 2\int_0^x f_T(t)dt = 2\left(\int_{-\infty}^{x} f_T(t)dt - \frac{1}{2}\right) = 2F_T(x) - 1 \quad (39)$$

Здесь $F_T(x)$ — *функция распределения Стьюдента с* $n-1$ *степенями свободы.*
Полагаем $2F_T(x) - 1 = 1 - \alpha$, где α — выбранный уровень значимости. Тогда $F_T(x) = 1 - \alpha/2$. Пусть $t_{1-\alpha/2}(n-1)$ — квантиль распределения Стьюдента с $n-1$ степенями свободы и порядка $1-\alpha/2$. Из предыдущих равенств мы получаем

$$P\left(\bar{x} - \frac{sx}{\sqrt{n-1}} < m < \bar{x} + \frac{sx}{\sqrt{n-1}}\right) = 2F_T(x) - 1 = 1 - \alpha, \\ P\left(\bar{x} - \frac{st_{1-\alpha/2}(n?1)}{\sqrt{n-1}} < m < \bar{x} + \frac{st_{1-\alpha/2}(n?1)}{\sqrt{n-1}}\right) = 1 - \alpha, \quad (40)$$

что и даёт доверительный интервал для m с доверительной вероятностью $\gamma = 1 - \alpha$ [1, с. 457-458].

2.9.2 Доверительный интервал для среднего квадратического отклонения σ нормального распределения

Дана выборка $(x_1, x_2, ..., x_n)$ объёма n из нормальной генеральной совокупности. На её основе строим выборочную дисперсию s^2. Параметры m и σ нормального распределения неизвестны. Доказано, что случайная величина ns^2/σ^2 распределена по закону χ^2 с $n-1$ степенями свободы.
Задаёмся уровнем значимости α и находим квантили $\chi^2_{\alpha/2}(n-1)$ и $\chi^2_{1-\alpha/2}(n-1)$.
Это значит, что

$$P\left(\chi^2(n-1) < \chi^2_{\alpha/2}(n-1)\right) = \alpha/2, \\ P\left(\chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)\right) = 1 - \alpha/2 \quad (41)$$

Тогда

$$P\left(\chi^2_{\alpha/2}(n-1) < \chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)\right) = \\ = P\left(\chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)\right) - P\left(\chi^2(n-1) < \chi^2_{\alpha/2}(n-1)\right) = \\ = 1 - \alpha/2 - \alpha/2 = 1 - \alpha \quad (42)$$

Отсюда

$$P\left(\chi^2_{\alpha/2}(n-1) < \frac{ns^2}{\sigma^2} < \chi^2_{1-\alpha/2}(n-1)\right) = P\left(\frac{1}{\chi^2_{1-\alpha/2}(n-1)} < \frac{\sigma^2}{ns^2} < \frac{1}{\chi^2_{\alpha/2}(n-1)}\right) = \\ = P\left(\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right) = 1 - \alpha \quad (43)$$

Окончательно

$$P\left(\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right) = 1-\alpha, \tag{44}$$

что и даёт доверительный интервал для σ с доверительной вероятностью $\gamma = 1-\alpha$ [1, с. 458-459].

2.10 Доверительные интервалы для математического ожидания m и среднего квадратического отклонения σ произвольного распределения при большом объёме выборки. Асимптотический подход

При большом объёме выборки для построения доверительных интервалов может быть использован асимптотический метод на основе центральной предельной теоремы.

2.10.1 Доверительный интервал для математического ожидания m произвольной генеральной совокупности при большом объёме выборки

Выборочное среднее $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ при большом объёме выборки является суммой большого числа взаимно независимых одинаково распределённых случайных величин. Предполагаем, что исследуемое генеральное распределение имеет конечные математическое ожидание m и дисперсию σ^2. Тогда в силу центральной предельной теоремы центрированная и нормированная случайная величина $(\bar{x} - M\bar{x})/\sqrt{D\bar{x}} = \sqrt{n}(\bar{x}-m)/\sigma$ распределена приблизительно нормально с параметрами 0 и 1. Пусть

$$\Phi(x) = \frac{1}{2\pi}\int_{-\infty}^{x} e^{-t^2/2}dt \tag{45}$$

- *функция Лапласа.* Тогда

$$P\left(-x < \sqrt{n}\frac{\bar{x}-m}{\sigma} < x\right) = P\left(-x < \sqrt{n}\frac{m-\bar{x}}{\sigma} < x\right) \approx$$

$$\approx \Phi(x) - \Phi(-x) = \Phi(x) - [1-\Phi(x)] = 2\Phi(x) - 1 \tag{46}$$

Отсюда

$$P\left(\bar{x} - \frac{\sigma x}{\sqrt{n}} < m < \bar{x} - \frac{\sigma x}{\sqrt{n}}\right) \approx 2\Phi(x) - 1 \tag{47}$$

Полагаем $2\Phi(x) - 1 = \gamma = 1-\alpha$; тогда $\Phi(x) = 1-\alpha/2$. Пусть $u_{1-\alpha/2}$ — квантиль нормального распределения N(0,1) порядка $1-\alpha/2$. Заменяя в равенстве (47) σ на s, запишем его в виде

$$P\left(\bar{x} - \frac{su_{1-\alpha/2}}{\sqrt{n}} < m < \bar{x} - \frac{su_{1-\alpha/2}}{\sqrt{n}}\right) \approx \gamma, \tag{48}$$

что и даёт доверительный интервал для m с доверительной вероятностью $\gamma = 1-\alpha$ [1, с. 460].

2.10.2 Доверительный интервал для среднего квадратического отклонения σ произвольной генеральной совокупности при большом объёме выборки

Выборочная дисперсия $s^2 = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n}$ при большом объёме выборки является суммой большого числа практически взаимно независимых случайных величин (имеется одна связь $\sum_{i=1}^{n} x_i = n\bar{x}$, которой при

большом n можно пренебречь). Предполагаем, что исследуемая генеральная совокупность имеет конечные первые четыре момента.
В силу центральной предельной теоремы центрированная и нормированная случайная величина $(s^2 - Ms^2)/\sqrt{Ds^2}$ при большом объёме выборки n распределена приблизительно нормально с параметрами 0 и 1. Пусть $\Phi(x)$ — функция Лапласа (45). Тогда

$$P\left(-x < \frac{s^2 - Ms^2}{\sqrt{Ds^2}} < x\right) \approx \Phi(x) - \Phi(-x) = \Phi(x) - [1 - \Phi(x)] = 2\Phi(x) - 1 \tag{49}$$

Положим $2\Phi(x) - 1 = \gamma = 1 - \alpha$. Тогда $\Phi(x) = 1 - \alpha/2$. Пусть $u_{1-\alpha/2}$ — корень этого уравнения — квантиль нормального распределения N(0,1) порядка $1 - \alpha/2$. Известно, что $Ms^2 = \sigma^2 - \frac{\sigma^2}{n} \approx \sigma^2$ и $Ds^2 = \frac{\mu_4 - \mu_2^2}{n} + o(\frac{1}{n}) \approx \frac{\mu_4 - \mu_2^2}{n}$. Здесь μ_k — центральный момент k-го порядка генерального распределения; $\mu_2 = \sigma^2$; $\mu_4 = M[(x - Mx)^4]$; $o(\frac{1}{n})$ — бесконечно малая высшего порядка, чем $1/\mathrm{n}$, при $n \to \infty$. Итак, $Ds^2 \approx \frac{\mu_4 - \mu_2^2}{n}$. Отсюда

$$Ds^2 \approx \frac{\sigma^4}{n}(\frac{\mu_4}{\sigma^4} - 1) = \frac{\sigma^4}{n}((\frac{\mu_4}{\sigma^4} - 3) + 2) = \frac{\sigma^4}{n}(E + 2) \approx \frac{\sigma^4}{n}(e + 2), \tag{50}$$

где $\mathrm{E} = \frac{\mu_4}{\sigma^4} - 3$ — эксцесс генерального распределения, $\mathrm{e} = \frac{m_4}{s^4} - 3$ — выборочный эксцесс; $m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$ - четвёртый выборочный центральный момент. Далее,

$$\sqrt{Ds^2} \approx \frac{\sigma^2}{\sqrt{n}}\sqrt{e + 2} \tag{51}$$

Преобразуем неравенства, стоящие под знаком вероятности в формуле
$P\left(-x < \frac{s^2 - Ms^2}{\sqrt{Ds^2}} < x\right) = \gamma$:

$$\begin{aligned} -\sigma^2 U < s^2 - \sigma^2 < \sigma^2 U; \\ \sigma^2(1 - U) < s^2 < \sigma^2(1 + U); \\ 1/[\sigma^2(1 + U)] < 1/s^2 < 1/[\sigma^2(1 - U)]; \\ s^2/(1 + U) < \sigma^2 < s^2/(1 - U); \\ s(1 + U)^{-1/2} < \sigma < s(1 - U)^{-1/2}, \end{aligned} \tag{52}$$

где $U = u_{1-\alpha/2}\sqrt{(e + 2)/n}$ или
$s(1 + u_{1-\alpha/2}\sqrt{(e + 2)/n})^{-1/2} < \sigma < s(1 - u_{1-\alpha/2}\sqrt{(e + 2)/n})^{-1/2}$
Разлагая функции в биномиальный ряд и оставляя первые два члена, получим

$$s(1 - 0.5U) < \sigma < s(1 + 0.5U) \tag{53}$$

или

$$s(1 - 0.5u_{1-\alpha/2}\sqrt{(e + 2)/n}) < \sigma < s(1 + 0.5u_{1-\alpha/2}\sqrt{(e + 2)/n}) \tag{54}$$

Формулы (52) или (54) дают доверительный интервал для σ с доверительной вероятностью $\gamma = 1 - \alpha$.
Замечание. Вычисления по формуле (52) дают более надёжный результат, так как в ней меньше грубых приближений.

3 Реализация

Лабораторная работа была выполнена с помощью встроенных средств языка программирования Python (библиотеки: NumPy, SciPy, Matplotlib, Seaborn) в среде разработки Visual Studio Code. Исходный код работы приведен в приложении.

Ссылка на репозиторий с исходным кодом: `https://github.com/anivse/MathematicalStatistics`

4 Результаты

4.1 Выборочные коэффициенты корреляции

$size = 20$	r	r_s	r_q
$E(z)$	0.00559	0.00604	0.00180
$E(z^2)$	0.05722	0.05508	0.05388
$D(z)$	0.05719	0.05504	0.05388
$size = 60$	r	r_s	r_q
$E(z)$	-0.00476	-0.00386	-0.00260
$E(z^2)$	0.01807	0.01819	0.01728
$D(z)$	0.01805	0.01817	0.01727
$size = 100$	r	r_s	r_q
$E(z)$	-0.00309	-0.00270	-0.00348
$E(z^2)$	0.00980	0.00990	0.01003
$D(z)$	0.00979	0.00989	0.01002

Таблица 1: Двумерное нормальное распределение $\rho = 0$

$size = 20$	r	r_s	r_q
$E(z)$	0.48728	0.46564	0.33940
$E(z^2)$	0.27003	0.25306	0.16420
$D(z)$	0.03259	0.03624	0.04901
$size = 60$	r	r_s	r_q
$E(z)$	0.49830	0.47663	0.33273
$E(z^2)$	0.25762	0.23768	0.12535
$D(z)$	0.00932	0.01050	0.01464
$size = 100$	r	r_s	r_q
$E(z)$	0.49535	0.47482	0.32672
$E(z^2)$	0.25123	0.23225	0.11668
$D(z)$	0.00586	0.00679	0.00993

Таблица 2: Двумерное нормальное распределение $\rho = 0.5$

$size = 20$	r	r_s	r_q
$E(z)$	0.89710	0.86851	0.69720
$E(z^2)$	0.80705	0.75889	0.51384
$D(z)$	0.00225	0.00458	0.02775
$size = 60$	r	r_s	r_q
$E(z)$	0.89786	0.88201	0.70300
$E(z^2)$	0.80679	0.77903	0.50279
$D(z)$	0.00064	0.00110	0.00858
$size = 100$	r	r_s	r_q
$E(z)$	0.89863	0.88605	0.70752
$E(z^2)$	0.80794	0.78571	0.50582
$D(z)$	0.00039	0.00062	0.00524

Таблица 3: Двумерное нормальное распределение $\rho = 0.9$

$size = 20$	r	r_s	r_q
$E(z)$	0.78243	0.74746	0.56700
$E(z^2)$	0.62071	0.57078	0.35628
$D(z)$	0.00851	0.01208	0.03479
$size = 60$	r	r_s	r_q
$E(z)$	0.79042	0.77002	0.58047
$E(z^2)$	0.62747	0.59652	0.34752
$D(z)$	0.00271	0.00359	0.01058
$size = 100$	r	r_s	r_q
$E(z)$	0.78787	0.76874	0.57448
$E(z^2)$	0.62225	0.59306	0.33664
$D(z)$	0.00151	0.00210	0.00662

Таблица 4: Смесь нормальных распределений

4.2 Эллипсы рассеивания

Для уравнения эллипса выбиралась константа равная $const = 3\sigma$

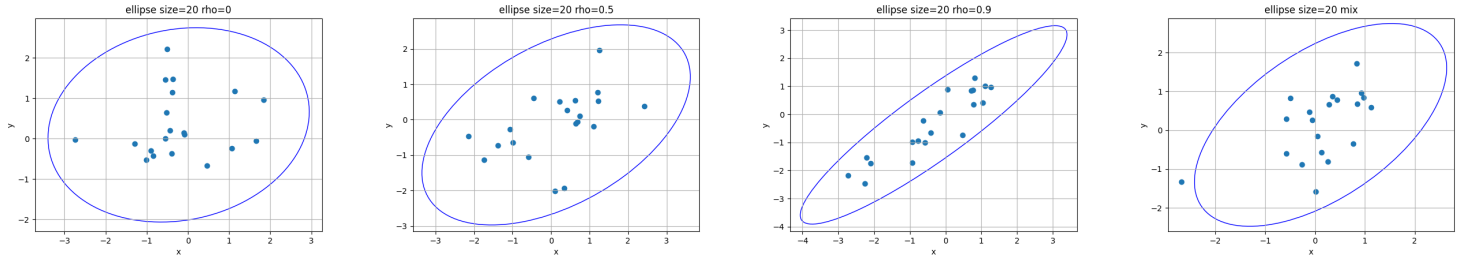

Рис. 1: Эллипсы рассеивания для двумерного нормального распределения и смеси нормальных распределений, n = 20

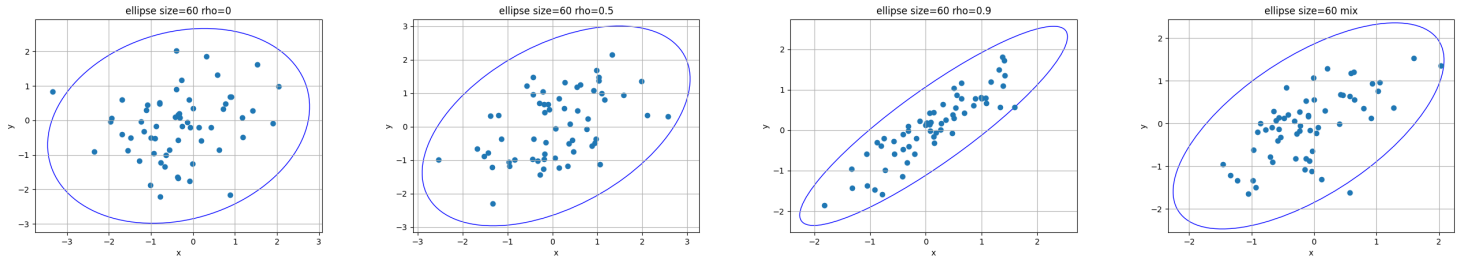

Рис. 2: Эллипсы рассеивания для двумерного нормального распределения и смеси нормальных распределений, n = 60

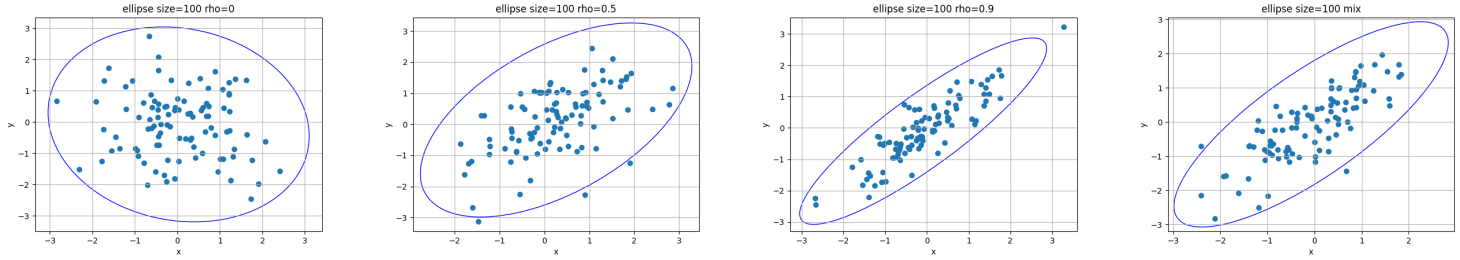

Рис. 3: Эллипсы рассеивания для двумерного нормального распределения и смеси нормальных распределений, n = 100

4.3 Оценки коэффициентов линейной регрессии

Метрика удаленности: $distance = \sum_{i=0}^{n} (y_{model}[i] - y_{regr}[i])^2$

4.3.1 Выборка без возмущений

Критерий наименьших квадратов: $\hat{a} \approx 2.33932, \quad \hat{b} \approx 2.03879$

Критерий наименьших модулей: $\hat{a} \approx 2.07868, \quad \hat{b} \approx 2.02210$

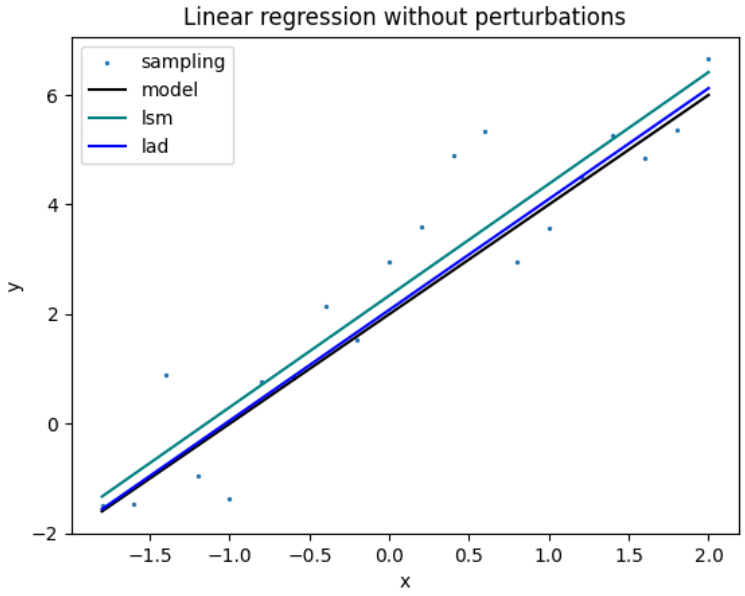

Рис. 4: Линейная регрессия для выборки без возмущений

4.3.2 Выборка с возмущениями

Были внесены возмущения 10 и -10 в y_1 и y_{20} соответственно.

Критерий наименьших квадратов: $\hat{a} \approx 2.48218, \quad \hat{b} \approx 0.61022$

Критерий наименьших модулей: $\hat{a} \approx 2.17987, \quad \hat{b} \approx 1.77162$

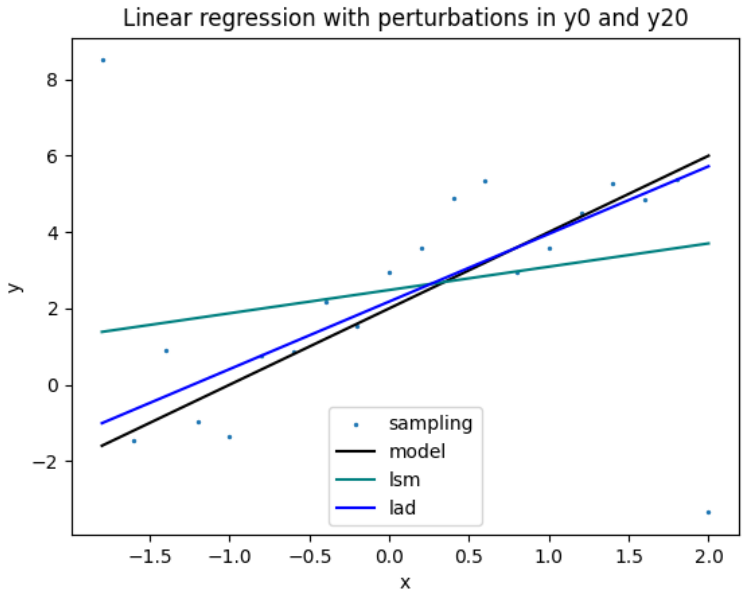

Рис. 5: Линейная регрессия для выборки с возмущениями

5 Обсуждение

Литература

[1] Histogram. URL: https://en.wikipedia.org/wiki/Histogram

[2] Вероятностные разделы математики. Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д. — Спб.: «Иван Федоров», 2001. — 592 с., илл.

[3] Box plot. URL: https://en.wikipedia.org/wiki/Box_plot

[4] Анатольев, Станислав (2009) «Непараметрическая регрессия», Квантиль, №7, стр. 37-52.