Московский авиационный институт
(Национальный исследовательский университет)

Институт: «Информационные технологии и прикладная математика»
Кафедра: 804 «Теория вероятностей и компьютерное моделирование»
Дисциплина: «Эконометрика»

Курсовая работа

Регрессионный анализ

Студент: Батова Е.Д.
Группа: М8О-401Б-19
Преподаватель: Платонов Е.Н.
Оценка:
Вариант №1

Москва, 2022

Содержание

1. Текст задания с вариантом

1. Теоретическая часть

Написать эссе по теме Разрывный дизайн

2. Практическая часть

	Функция f(h)	Носитель для h	Дисперсия шума σ^2
1	0.2*h+1+sin(2*h)	-5 < h < 5	0.25

1. Модельная часть

Смоделировать данные самостоятельно в соответствии с вариантом

$$X_k = f(h_k) + e_k, k = 1, , 60$$

где ek — независимый случайные величины с распределением $N(0, \sigma^2)$.

Точки внутри носителя для h выбирать равномерно.

Смоделировать тестовую выборку объема 40, половина значений правее наблюдаемых значений, половина левее.

2. Метод наименьших квадратов

Для регрессии вида $X_k = \theta_0 + \theta_1 h_k + e_k, k = 1, , 60$ (1)

1. Найти МНК-оценки неизвестных параметров.
2. Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.
3. Вычислить коэффициент детерминации и найти оценку ковариационной матрицы МНК-оценки.
4. Найти значения информационных критериев
5. С помощью критерия Фишера проверить гипотезу $\theta_0 = 0, \theta_1 = 0$
6. Построить доверительный интервал надежности 0.95 и 0.8 для полезного сигнала $X = \theta_1 + theta_2 h$ при h из исходного носителя ±50%.

7. Построить оценку метода наименьших модулей, отобразить ее на графике
8. Оценить качество построенных регрессий на тестовой выборке

   Для остатков $\hat{e}_k = X_k - \hat{X}_k$:

1. Построить гистограмму, ядерную оценку плотности распределения
2. По остаткам проверить гипотезу, что *e* имеет гауссовское распределение с помощью одного из критериев:
   - критерий Шапиро-Уилка (Shapiro–Wilk);
   - критерий D'Agostino K2;
   - критерий Харке-Бера (Jarque–Bera).

3. Проверить наличие автокорреляции с помощью критерия Дарбина-Уотсона.
4. Проверить наличие гетероскедастичности с помощью одного из критериев.

2.3. Полиномиальная регрессия

Построить регрессию с помощью МНК

$X = \theta_0 + \theta_1 h + \theta_2 h^2 + \ + \theta_p h^p$

1. Порядок полинома p подбирать несколькими способами:
   по значению среднеквадратической погрешности МНК-оценки (на обучающей и/или тествой)
   по значению статистики критерия Фишера для гипотезы ${}_p = 0$
   по MSE на тестовой выборке
   ваш способ
   Выбираем единственное значение p.
2. Провести анализ остатков по схеме из пункта 2.2.
3. Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.
4. Проверить для подобранной модели является ли матрица

2.4. Регрессия для наблюдений с выбросами

1. Смоделировать ошибки для модели регрессии (1) с помощью распределения Тьюки, приняв долю выбросов $\sigma = 0.08$, номинальную дисперсию $\sigma_0^2 = \sigma^2$, дисперсию аномальных наблюдений $\sigma_1^2 = 100\sigma^2$.
2. Построить МНК-оценку неизвестных параметров для модели (1) и оценить ее качество.
3. Провести анализ остатков по схеме из пункта 2.2.
4. Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.
5. Провести отбраковку выбросов и пересчитать МНК-оценку и оценить качество оценки.
6. Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.
7. Провести анализ остатков по схеме из пункта 2.2.
8. Построить оценку метода наименьших модулей.
9. Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.
10. Провести анализ остатков по схеме из пункта 2.2.
11. Построить робастную оценку Хубера (дополнительное задание)

2.5. Квантильная регрессия

1. Смоделировать несимметричные ошибки для исходных данных, заменив у 90% отрицательных ошибок знак с минуса на плюс.
2. Построить МНК и МНМ оценки для получившихся наблюдений и регрессии (1).
3. Построить несколько квантильных регрессий (для разных значений параметра α) и оценить их качество.
4. Построить график, на котором отобразить наблюдения, исходную функцию и линии регрессии.

2. Теоретическая часть.

Впервые разрывный дизайн был предложен Дональдом Тистлтуэйтом и Дональдом Кэмпбеллом в 1960 году. Статья показывает связь между получением стипендии и общественным признанием в дальнейшем . В исследовании стипендию получали все, кто на специальном экзамене получил оценку выше пороговой. Данное исследование показало, что в дальнейшем студенту, который получил стипендию, легче поддерживать успеваемость на высоком уровне, чем тому, кто не получил стипендию. Но разрывный дизайн свою практическую значимость в экономических и других исследованиях обрел не так давно. На данный момент можно найти достаточно относительно много статей, исследующих разрывный дизайн в разных социально-экономических моделях. Одной из причин его популярности среди квази-экспериментов является то, что данный дизайн требует менее строгих допущений.

Разрывный и дизайн, или “метод разрывной регрессии”, используется для исследования воздействия. В этом случае должен быть известен критерий отбора в определенную группу, получающую воздействия. Чаще всего используются две группы. Отбор в группу определяется с помощью критериальной переменной относительно пороговой переменной. В зависимости от заданных правил, для попадания в группу критериальная переменная должна быть больше (меньше), чем пороговая переменная. При рассмотрении связи поступления в институт и дальнейшей благополучной жизни сумма баллов за экзамены при поступлении является критериальной переменной, а минимальный балл является пороговой переменной. Иначе говоря,

$$Y_i = (1 - W_i)Y_i(0) + W_iY_i(1) = \begin{cases} Y_i(0) & W_i = 0 \\ Y_i(1) & W_i = 1 \end{cases}$$

где $W_i = \{0, 1\}$ - типы воздействия, Y_i - исход, $i = 1, ..N$ - индекс объекта. Основная идея разрывного дизайна состоит в том, что воздействие определяется полностью или частично значением переменной X_i, которая определяет право на участие в программе, также называется переменной отбора. Можно выделить два подхода разрывного дизайна: параметрический и непараметрический. Параметрический подход обычно представляется полиномиальной регрессией. Он представляется в следующем виде: $Y = \alpha + \tau W + \beta X + \epsilon$

Смотря насколько строго соблюдаются правила отбора, можно разделить на две категории: строгий и нестрогий (четкий и нечеткий). При строгом дизайне вероятность получения воздействия в пороговой точке изменяется с 0 до 1, в нестрогом же не изменяется с 0 до 1, то есть воздействие может определяться не полностью.

Нужно учитывать условия, которые накладываются на критериальную переменную. Не должно быть связи между критериальной переменной и результирующей переменной. Но должна быть сильная связь между категориальной переменной и вероятностью получения воздействия.

Также надо учитывать случаи, когда параметры могут быть изменены намеренно. Такой случай не будет являться разрывным дизайном. Самый простой случай возникновения такой ситуации: списывание теста студентом. Тогда группы изначально будут распределены неправильно.

В качестве примера можно взять статистику новых заболевших COVID-19 в городе Квебек. Выделив день начала эпидемии COVID-19, можно выделить зависимость смерти от COVID-19. Получившиеся данные представляют собой разрывный дизайн.

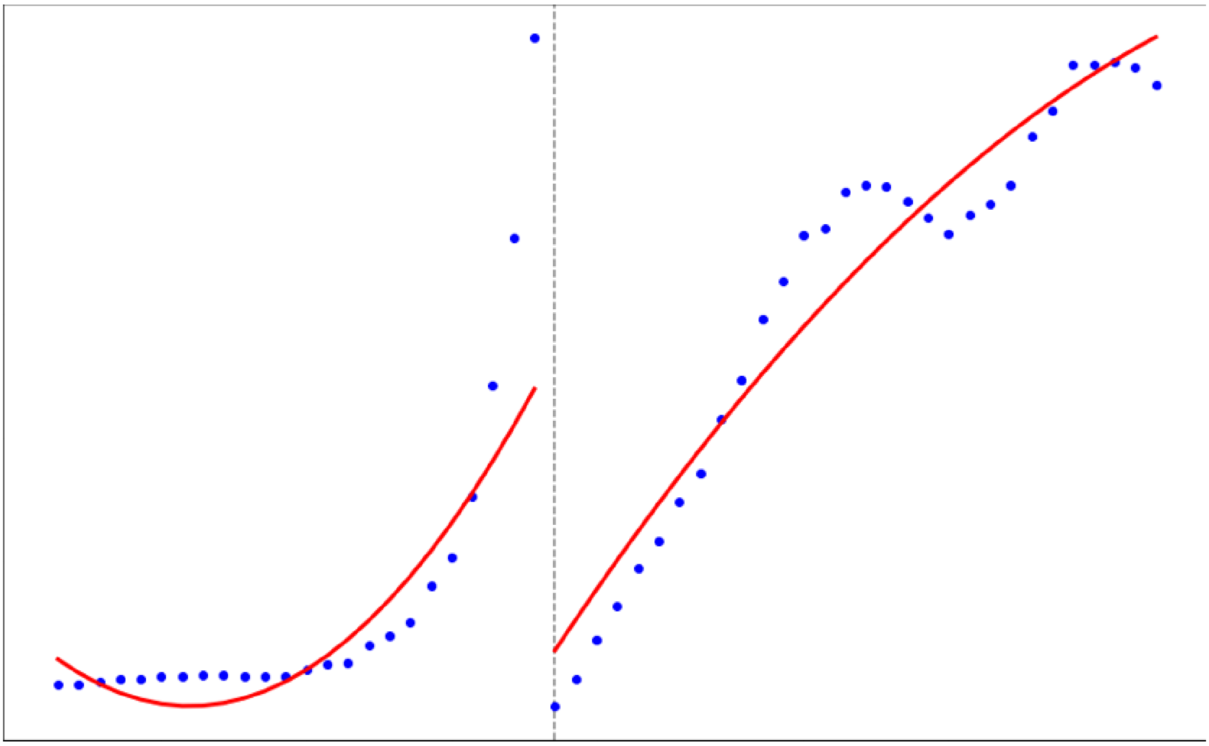

Рис 1. График разрывного дизайна

3. Практическая часть.

3.1. Модельная часть

$f(h) = 0.2 * h + 1 + sin(2 * h)$

$-5 < h < 5$

$\sigma^2 = 0.25$

Смоделируем обучающую и тестовую выборку. Тестовая выборка будет состоять из 40 значений, которые наблюдаются в слева и справа от обучающей выборки. Распределение ошибок ϵ является нормальным с параметрами $\mu = 0, \sigma^2 = 0.5$. Для задания ϵ в программной реализации используется функция $np.random.normal$

3.2. Метод наименьших квадратов.

Построим для выше заданной модели МНК и МНМ оценки.

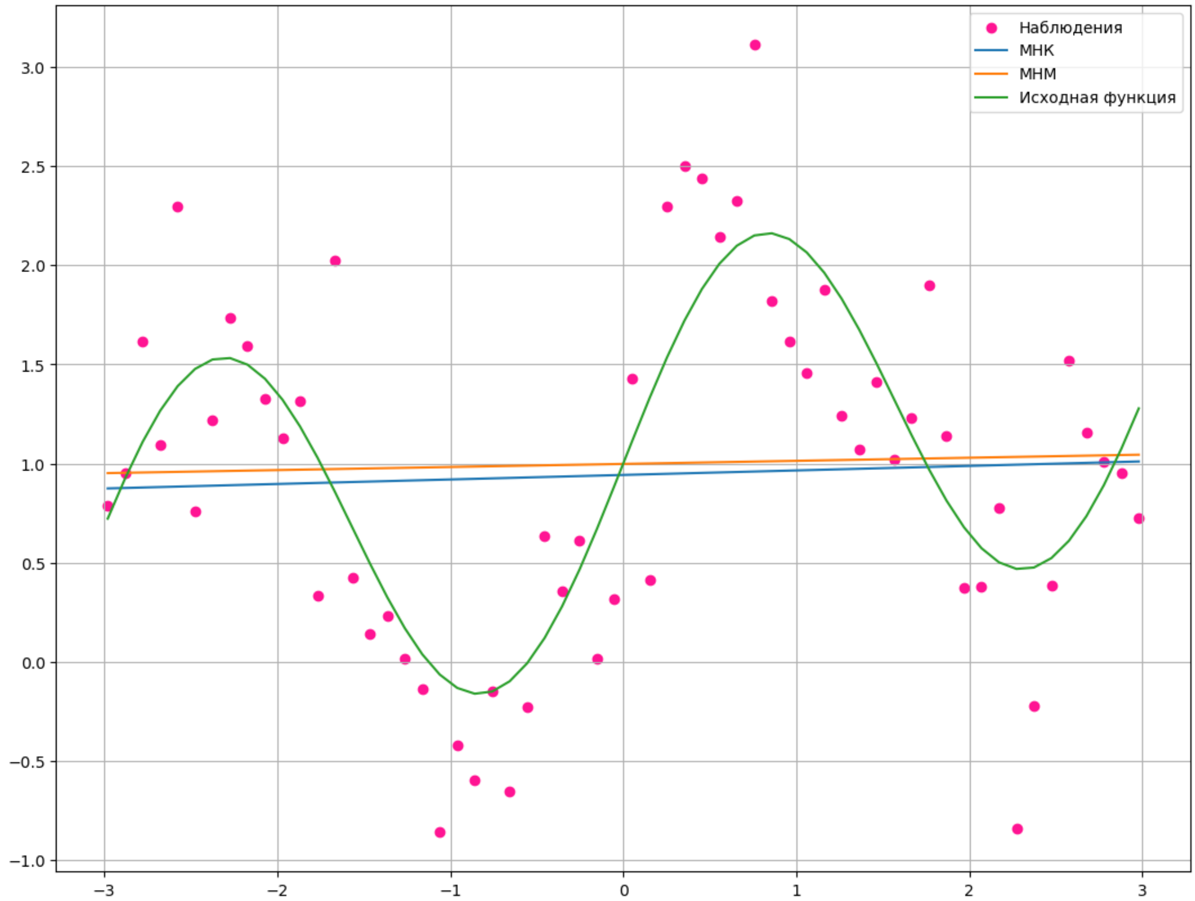

Рис 1. МНК и МНМ оценки

Для МНК и МНМ оценок были получены параметры
$\theta_0 = 0.9422628 \quad \theta_1 = 10.02278359$ и $\theta_0 = 0.49895908$
$\theta_1 = 0.01555645$

Были найдены коэффициент детерминации, ковариационная матрица, информационный критерий:

$R^2 = 0.0019890799251850444$

$K = [[1.37413612e-021.26299639e-18][1.26299639e-184.49055007e-03]]$

$AIC = -79.3463009585262$

С помощью критерия Фишера проверяем гипотезу
$\theta_0 = 0, \theta_1 = 0$

Получим статистику Фишера и значение p-value для данной статистики:
$F = 32.36388500561973, p-value = 3.6313604921141753e-10$
, то есть гипотеза не принимается на уровне значимости
$\alpha = 0.05$

Значение суммы квадратов отклонений на тестовой выборке для МНК и МНМ оценок равны 80.40911007515433 и 100.85761368515347 соответственно.

Также выведем доверительные интервалы:

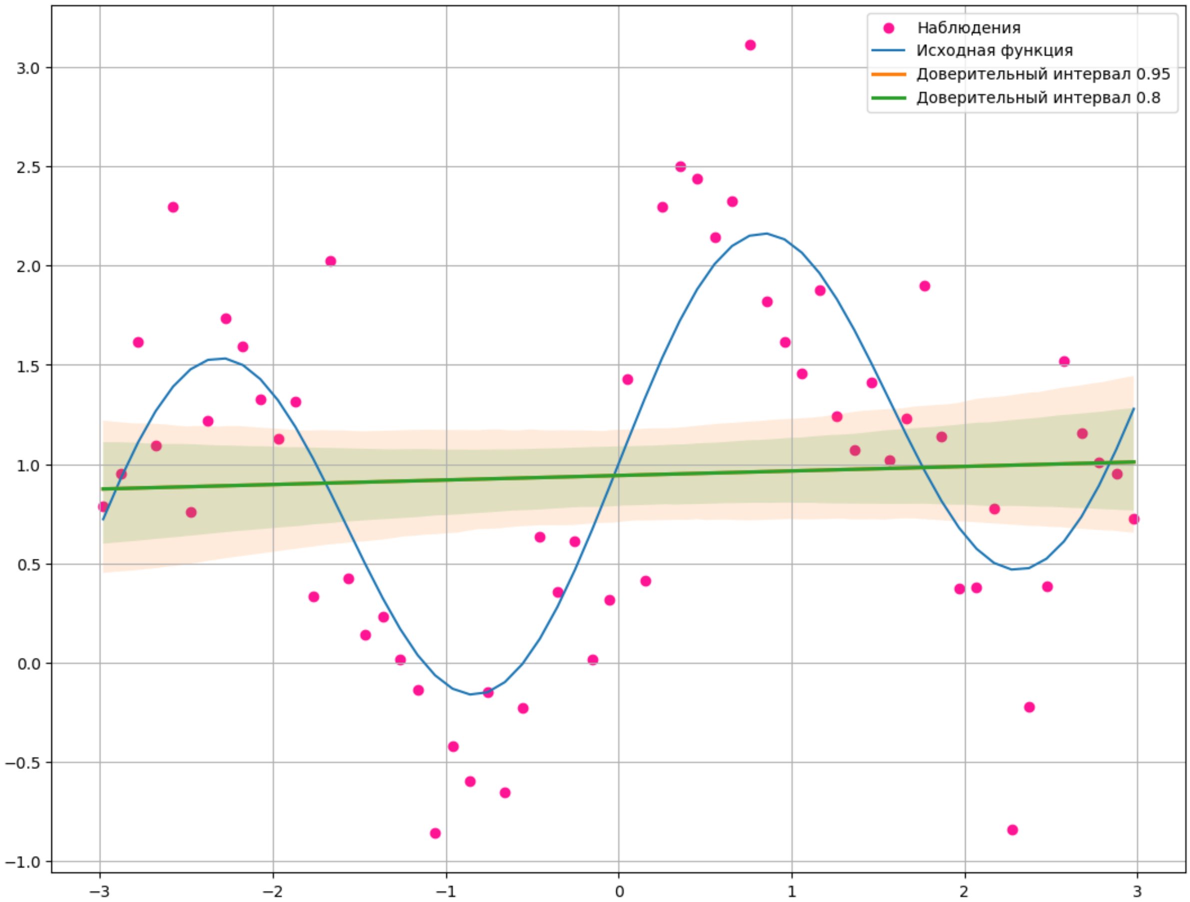

Рис 2. Доверительные интервалы

Для остатков $\hat{e}_k = X_k - \hat{X}_k$ построим гистограмму, ядерную оценку плотности распределения и проверим гипотезы:

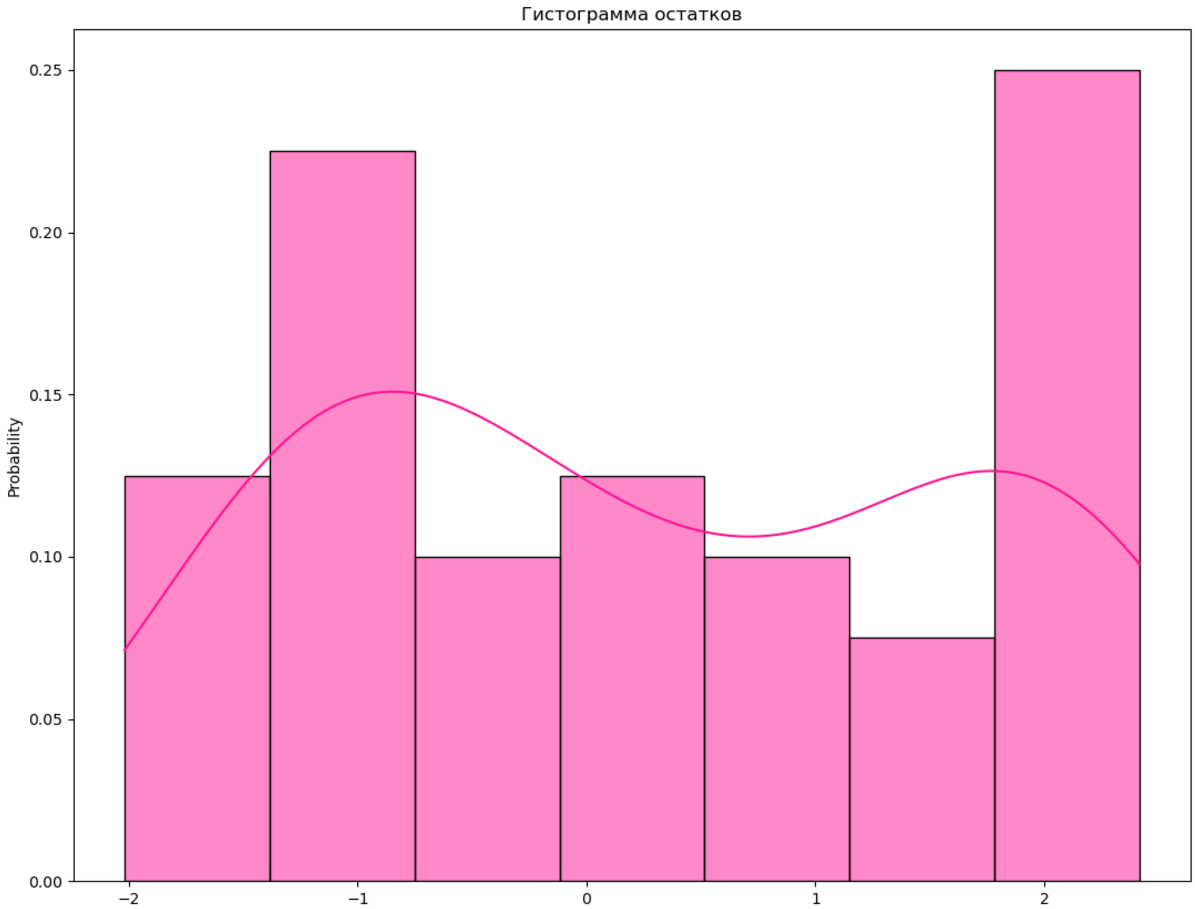

Рис 3. Гистограмма и ядерная оценка плотности распределения

Для проверки гипотезы о нормальном распределении используем критерий D'Agostino K2, который равен 20.004688290840647, (p-value=4.529363036434918e-05).

Для проверки гипотезы о наличии автокорреляции используем критерий Дарбина-Уотсона, который равен 0.2399556923915355

Для проверки гипотезы о наличии гетероскедастичности используем тест Уайта с LM-статистикой, который равен 7.522592483618449 (p-value 0.023253578569867567)

В данной части были построены две оценки: МНК и МНМ. Так как коэффициент детерминации мал, можно сказать, что регрессионная модель не соответствует заданной модели. Гипотеза о нормальном распределении не принимается, т.к p-value < 0.05, критерий Дарбина-Уотсона показывает положительную последовательность корреляции, тест Уайта показывает, что гипотеза не отвергается.

3.3. Полиномиальная регрессия.

Для полинома $X = \theta_0 + \theta_1 h + \theta_2 h^2 + \; + \theta_p h^p$ подберем параметр p несколькими способами. При подборе с помощью МНК оценки наилучшие результаты показывает p=5:

theta = [0.91155294 1.28191638 0.06068857 -0.49569262 -0.00476648 0.04131495]

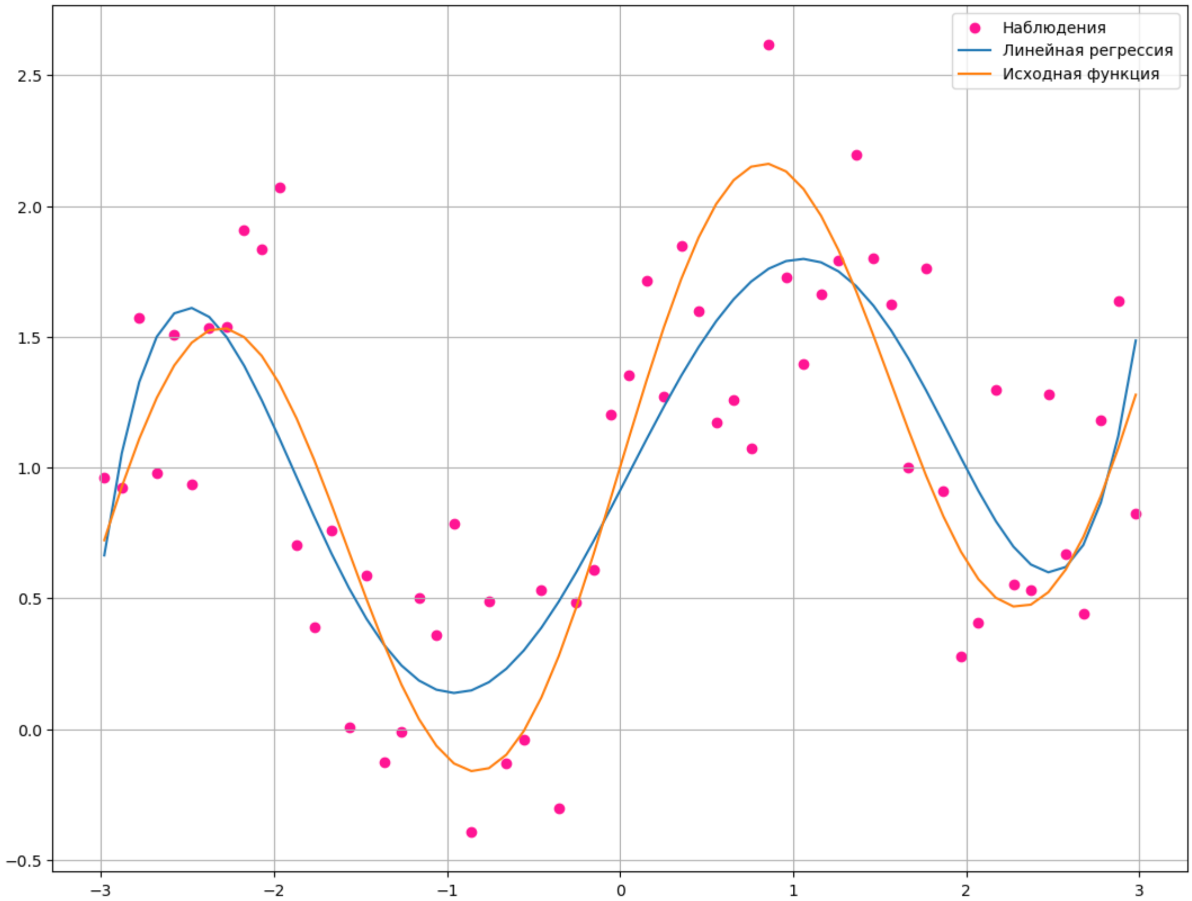

Рис 4. МНК оценка

При подборе с помощью оценки на тестовой выборке получим, что p=4 показывает лучше результат (15.216040994358057). Выберем p=5.

Для остатков $\hat{e}_k = X_k - \hat{X}_k$ построим гистограмму, ядерную оценку плотности распределения и проверим гипотезы:

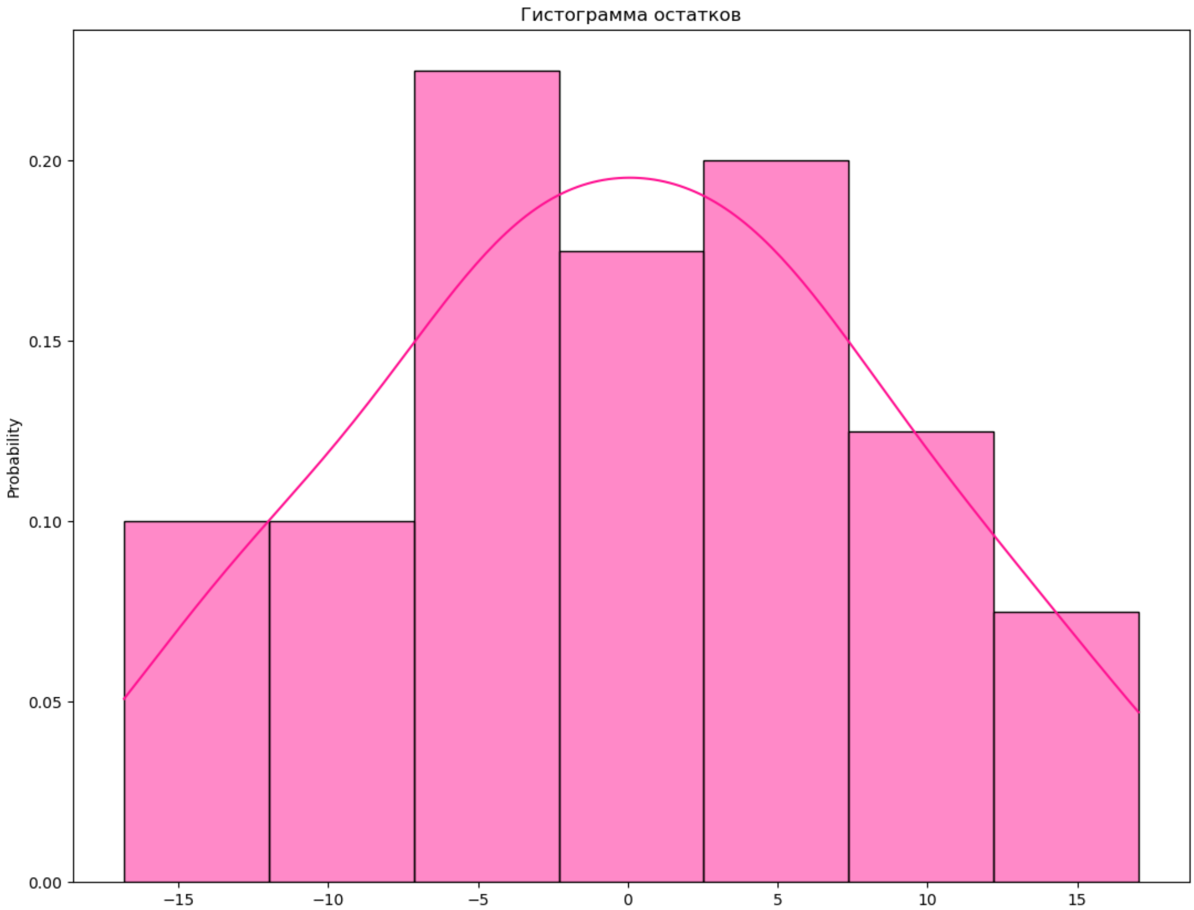

Рис 5. Гистограмма и ядерная оценка плотности распределения

Для проверки гипотезы о нормальном распределении используем критерий D'Agostino K2, который равен 0.9848925471305847, (p-value=0.8610321879386902).

Для проверки гипотезы о наличии автокорреляции используем критерий Дарбина-Уотсона, который равен 0.02005632827478433

Для проверки гипотезы о наличии гетероскедастичности используем тест Уайта с LM-статистикой, который равен 33.734936057579866 (p-value 4.726631135154426e-08)

Также матрица является мультиколлинеарной (43841.87284570927), поэтому строим ридж-оценку:

Коэффициенты Ридж оценки[0.   1.20891118 0.05986126 -0.46625808 -0.00466972 0.03875507]

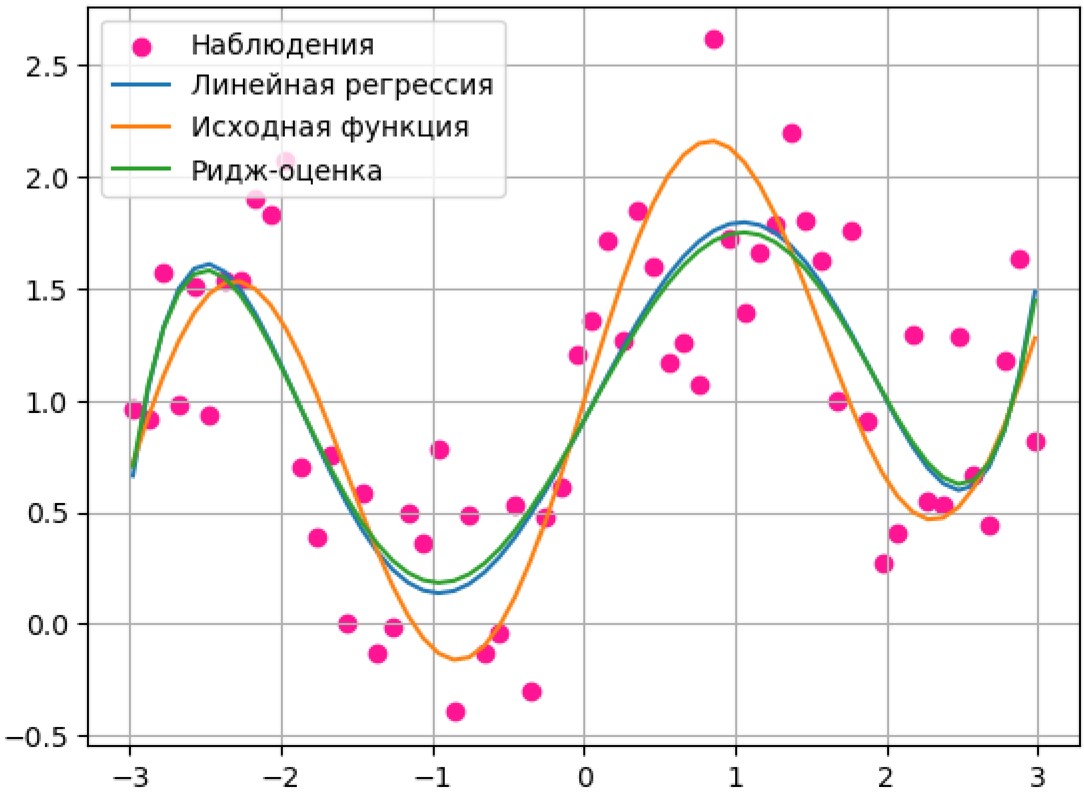

Рис 6. Ридж-оценка

В данной части была рассмотрена разновидность линейной регрессии, а именно полиномиальная при p=5.

Гипотеза о нормальном распределении не отвергается, т.к p-value > 0.05, критерий Дарбина-Уотсона показывает положительную последовательность корреляции, тест Уайта показывает, что гипотеза не отвергается. Матрица

3.4. Регрессия для наблюдений с выбросами.

Построим распределение Тьюки с долей выбросов $\sigma = 0.08$, номинальной дисперсией $\sigma_0^2 = \sigma^2$, дисперсией аномальных наблюдений $\sigma_1^2 = 100\sigma^2$.

Построим МНК оценку, получим $\theta_0 = 0.954602251 \theta_1 = -0.03302672$ с качеством 94.32952197000358

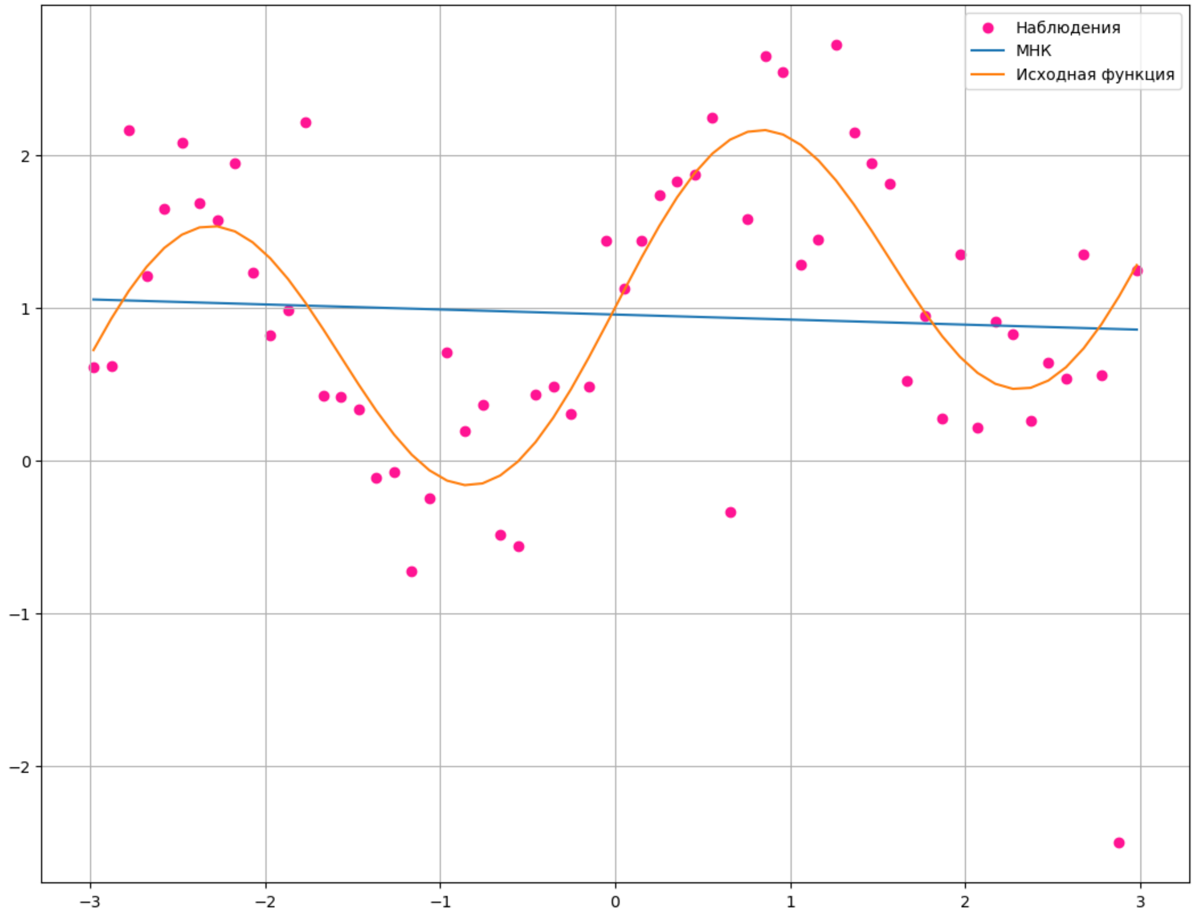

Рис 7. МНК оценка

Для остатков $\hat{e}_k = X_k - \hat{X}_k$ построим гистограмму, ядерную оценку плотности распределения и проверим гипотезы:

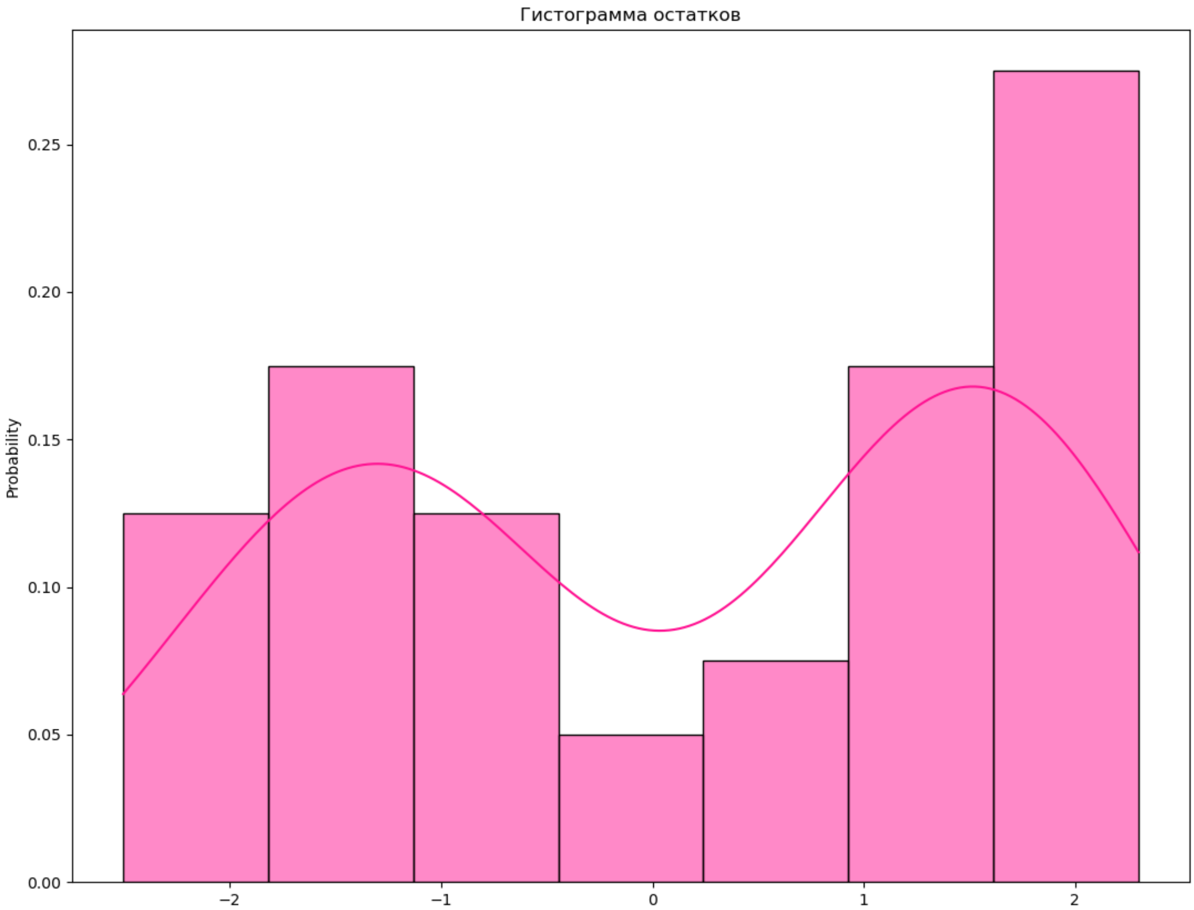

Рис 8. Гистограмма и ядерная оценка плотности распределения

Для проверки гипотезы о нормальном распределении используем критерий D'Agostino K2, который равен 38.360389980543616, (p-value=4.678936528815365e-09).

Для проверки гипотезы о наличии автокорреляции используем критерий Дарбина-Уотсона, который равен 0.3013970072611362

Для проверки гипотезы о наличии гетероскедастичности используем тест Уайта с LM-статистикой, который равен 0.7376954076126463 (p-value 0.6915307199274442)

Проведем отбраковку выбросов и пересчитаем МНК-оценку.

МНК оценку, получим $\theta_0 = 0.954602251 \theta_1 = -0.03302672$

с качеством 94.32952197000358

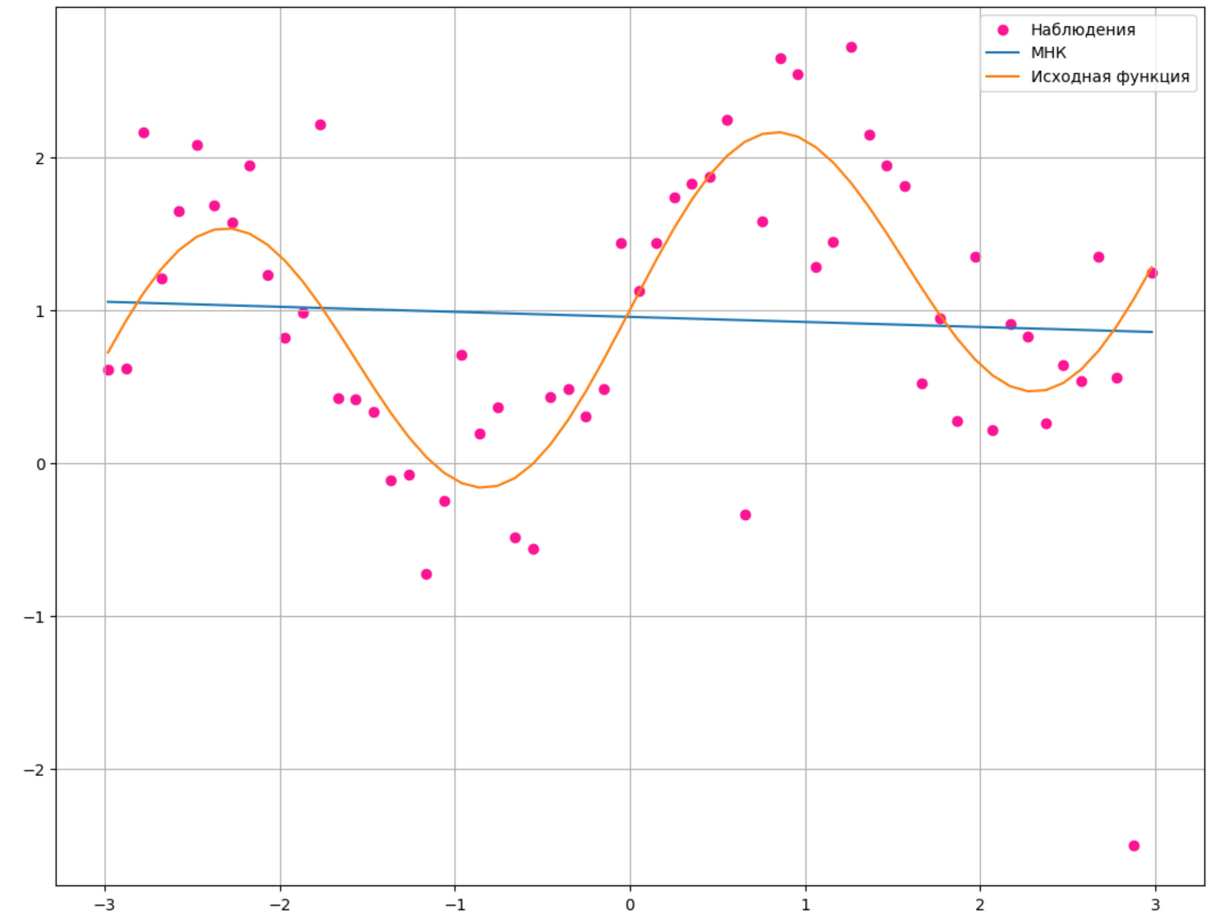

Рис 9. МНК оценка

Для остатков $\hat{e}_k = X_k - \hat{X}_k$ построим гистограмму, ядерную оценку плотности распределения и проверим гипотезы:

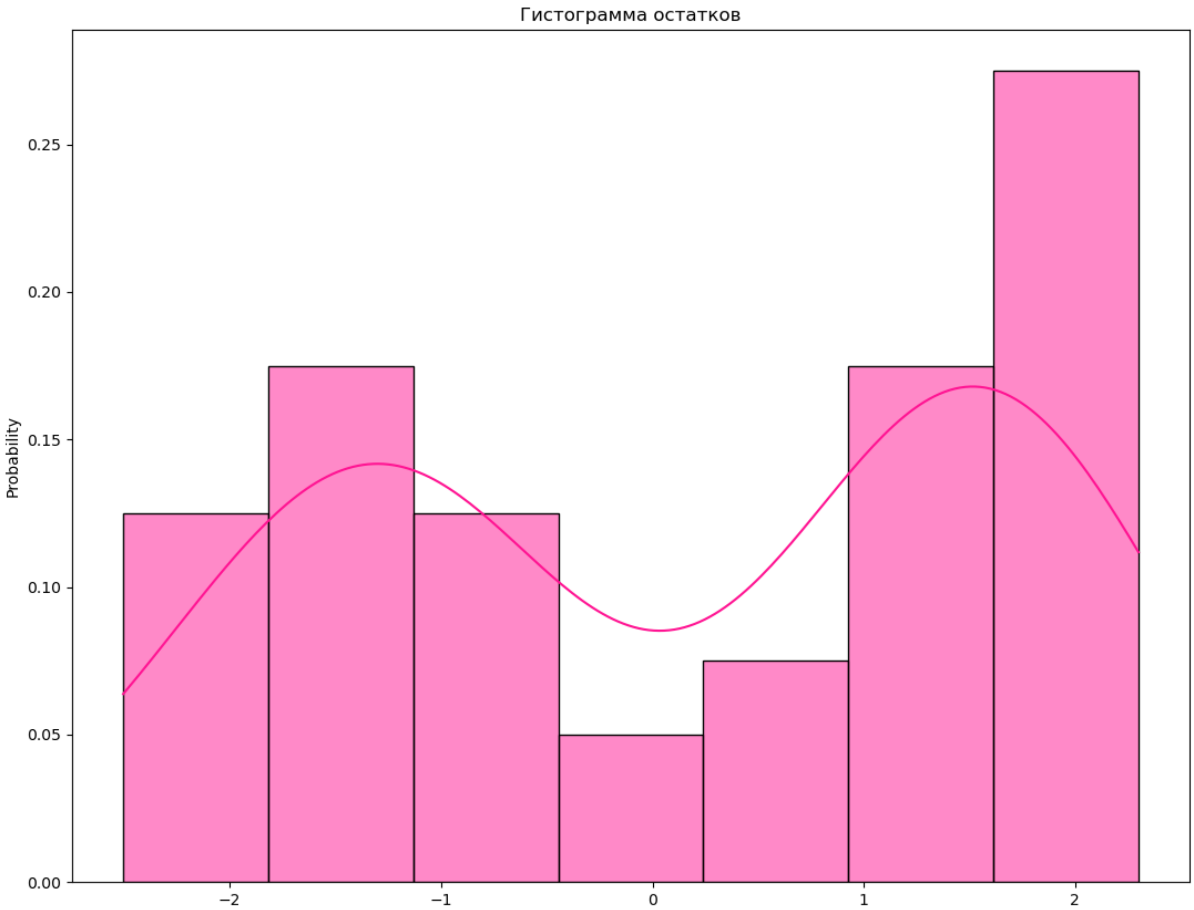

Рис 10. Гистограмма и ядерная оценка плотности распределения

Для проверки гипотезы о нормальном распределении используем критерий D'Agostino K2, который равен 38.360389980543616, (p-value=4.678936528815365e-09).

Для проверки гипотезы о наличии автокорреляции используем критерий Дарбина-Уотсона, который равен 0.3013970072611362

Для проверки гипотезы о наличии гетероскедастичности используем тест Уайта с LM-статистикой, который равен 0.7376954076126463 (p-value 0.6915307199274442)

Построим оценку метода наименьших модулей $\theta_0 = 0.47440008, \theta_1 - 0.0199769$, качество - 102.37339725362997

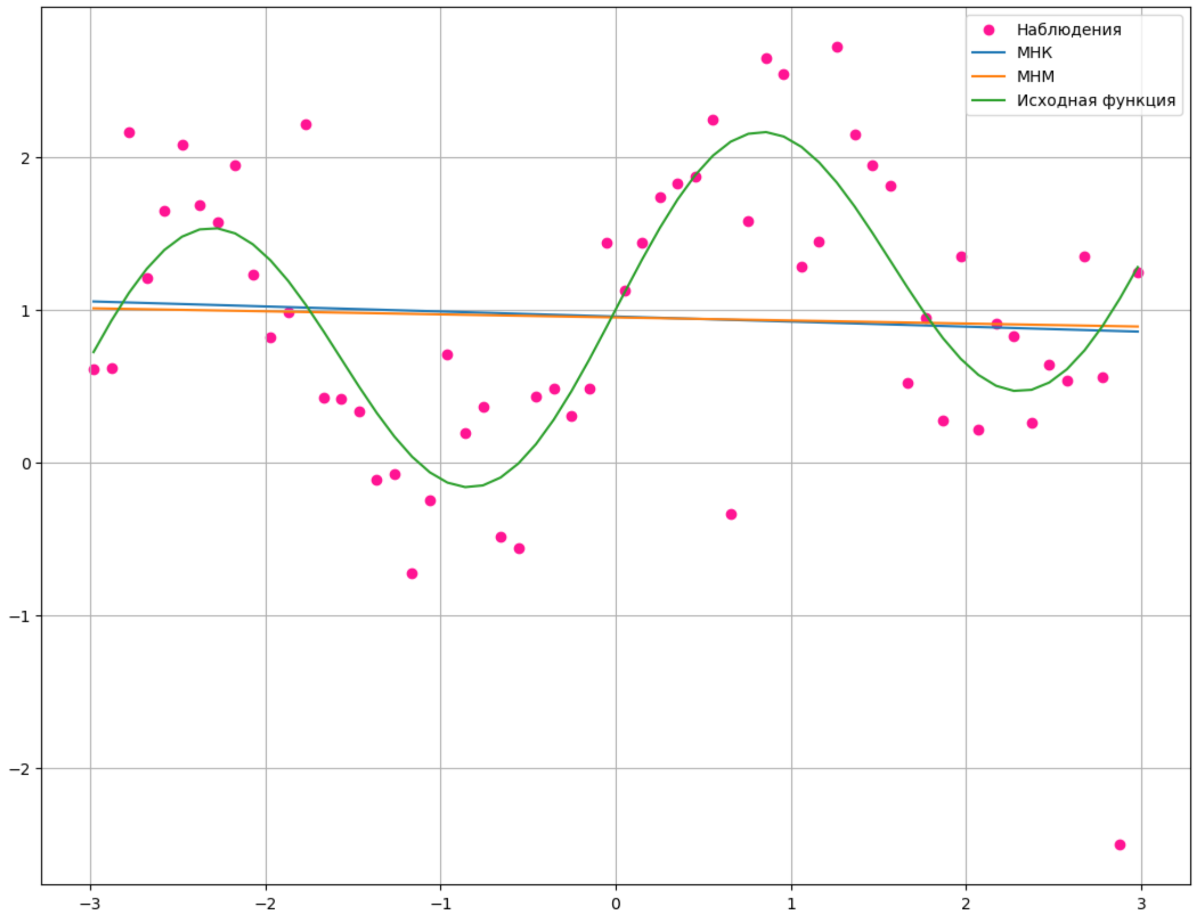

Рис 11. МНК и МНМ оценки

Для проверки гипотезы о нормальном распределении используем критерий D'Agostino K2, который равен =34.34914151802766, (p-value=3.476791844891835e-08).

Для проверки гипотезы о наличии автокорреляции используем критерий Дарбина-Уотсона, который равен 0.27158427305221455

Для проверки гипотезы о наличии гетероскедастичности используем тест Уайта с LM-статистикой, который равен 0.18.870695140797224 (p-value 7.985104707230058e-05)

Построим оценку Хуберта $\theta_0 = 0.49415107$ $\theta_1 = -0.01340154$ с качеством: 98.6977285575859

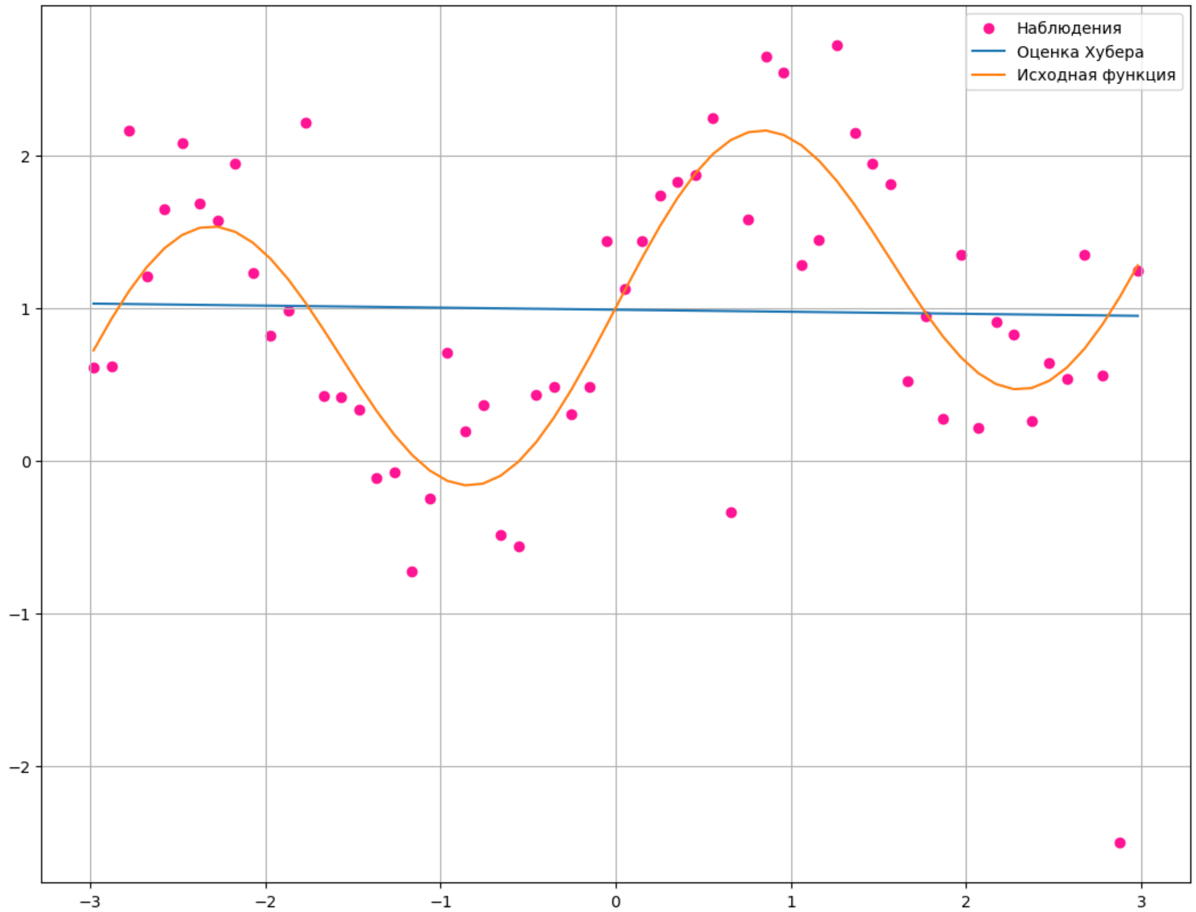

Рис 12. Оценка Хуберта.

В данном разделе были смоделированы данные с использованием распределения Тьюки в качестве ошибки. При отбраковке выбросов данные не были изменены, что может быть остатки меньше 2σ. Оценка Хьюберта получилась немного лучше, чем МНК оценка.

3.5. Квантильная регрессия.

Смоделируем несимметричные ошибки для исходных данных, заменив у 90% отрицательных ошибок знак с минуса на плюс.

Построим МНК и МНМ оценки. Получим

$\theta_0 = 1.01695361 theta_1 = 0.04316177$ и

$\theta_0 = 0.5263189 theta_1 = 0.04552599$ с качеством

64.25808451495975 и 75.18778030539156.

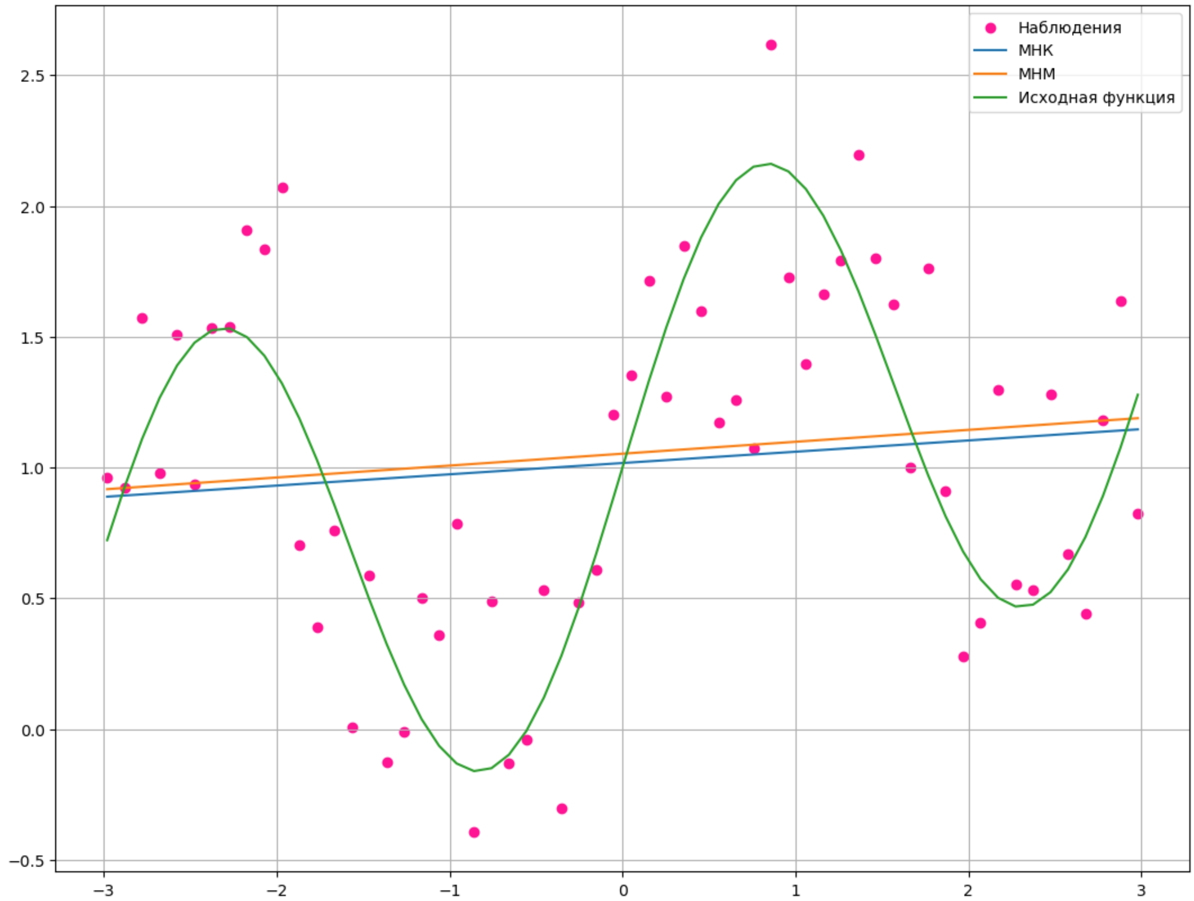

Рис 11. МНК и МНМ оценки

Построим несколько квантильных регрессий. Получим:

для q= 0.2 $\theta_0 = -7.21208721e - 17$

$\theta_1 = -1.76985257e - 15$ с качеством

125.81912736562352;

для q= 0.4 $\theta_0 = -2.74849524e - 13$

$\theta_1 = 9.50021848e - 14$ с качеством 125.81912736560872;

для q= 0.6 $\theta_0 = 1.03581162e - 13$

$\theta_1 = 3.31659166e - 14$ с качеством 125.81912736560078;

для q= 0.7 $\theta_0 = 2.07155723e - 10$

$\theta_1 = 2.40255536e - 11$ с качеством 125.81912733837189

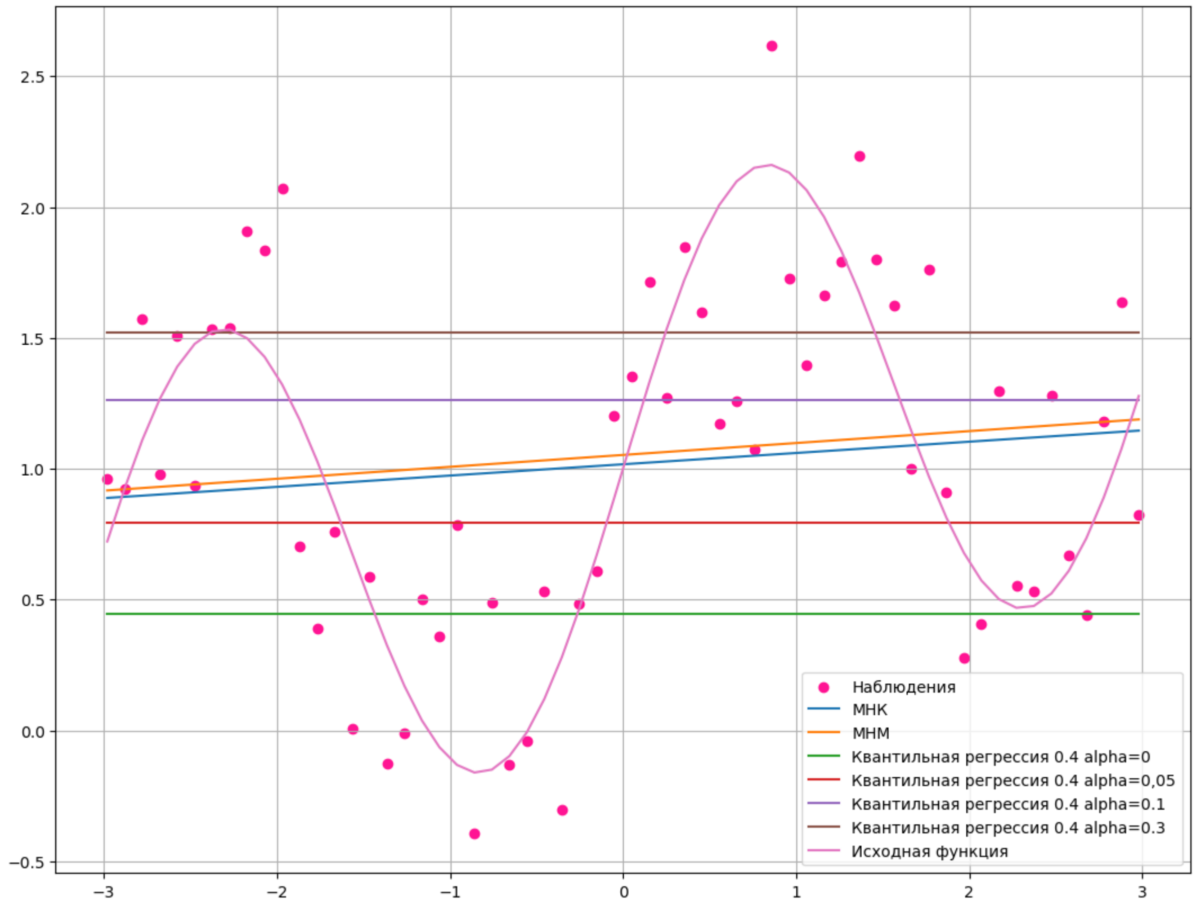

Рис 11. Квантильная регрессия

При изменении ошибки оценка модели улучшилась. Но квантильная регрессия показала плохие результаты.

4. Список использованной литературы.

1. Прикладные методы анализа статистических данных [Текст] : учеб. пособие / Е. Р. Горяинова, А. Р. Панков, Е. Н. Платонов ; Нац. исслед. ун-т «Высшая школа экономики». — М.: Изд. дом Высшей школы экономики, 2012. — 310, [2] с.
2. Ниворожкин, Антон (2009). «Разрывный дизайн», Квантиль, №7, стр. 1–8. Citation: Nivorozhkin, Anton (2009). “Regression discontinuity design,” Quantile, No.7, pp. 1–8
3. Метод разрывной регрессии и метод отбора подобного по вероятности для оценки эффекта одного года обучения: опыт применения на примере данных PISA 2009, “Социология: методология, методы, математическое моделирование.” 2014. № 38. С. 7-37. Кузьмина Ю. В.
4. Ссылка на текст программы https://github.com/e-k-a/econometrics