Московский авиационный институт
национальный исследовательский университет

Факультет информационных технологий и прикладной математики
Кафедра компьютерных методов в математическом моделировании сложных систем

Курсовая работа по Эконометрике на тему: регрессионный анализ

Студент: Королев Егор Владимирович
Преподаватель: Платонов Евгений Николаевич
Группа: М8О-401Б-18

Москва, 2021

Содержание

1 Задание

1.1 Теоретическая часть

Написать эссе по Байесовской регрессии.

1.2 Практическая часть

1.2.1 Модельная часть

Смоделировать данные:

$$X_k = f(h_k) + \varepsilon_k, \quad k = \overline{1,60},$$

где $f(h) = 1.5h - 2 - \frac{1}{2h}$, $h \in [0.1; 2]$, ε_k – независимый случайные величины с распределением $\mathcal{N}(0, \sigma^2)$.
Точки внутри носителя для h выбираются равномерно.
Смоделировать тестовую выборку объема 40, половина значений правее наблюдаемых значений, половина левее

1.2.2 Метод наименьших квадратов

Для регрессии вида:

$$X_k = \theta_0 + \theta_1 h_k + \varepsilon_k, k = \overline{1,60}$$

1 Найти МНК-оценки неизвестных параметров;

2 построить график, на котором отобразить наблюдения, исходную функцию и линию регрессии;

3 вычислить коэффициент детерминации и найти оценку ковариационной матрицы МНК-оценки;

4 найти значения информационных критериев;

5 с помощью критерия Фишера проверить гипотезу: $\theta_0 = \theta_1 = 0$;

6 построить доверительный интервал надежности 0.95 и 0.8 для полезного сигнала $X = \theta_0 + \theta_1 h$ при h из исходного носителя $\pm 50\%$;

7 построить оценку метода наименьших модулей, отобразить ее на графике;

8 оценить качество построенных регрессий на тестовой выборке.

Для остатков $\hat{\varepsilon_k} = X_k - \hat{X_k}$:

1 построить гистограмму;

2 на графике изобразить ядерную оценку плотности распределения;

3 по остаткам проверить гипотезу, что $\hat{\varepsilon}$ имеет гауссово распределение с помощью одного из критериев:

- критерий Шапиро-Уилка;
- критерий D'Agostino K^2;
- критерий Зарке-Бера;

4 проверить наличие автокорреляции с помощью критерия Дарбина-Уотсона;

5 проверить наличие гетероскедастичности с помощью одного из критериев;

1.2.3 Полиномиальная регрессия

Построить следующие регрессии с помощью МНК:

$$X = \sum_{i=0}^{p} \theta_i h^i$$

Порядок полинома p подобрать несколькими способами:

1 по значению среднеквадратичной погрешности МНК-оценки (на обучающей и/или тестовой);

2 по значению статистики критерия Фишера для гипотезы $\theta_p = 0$;

3 по MSE на тестовой выборке;

4 другим способом;

Для выбранного значения p:

- провести анализ остатков по схеме из пункта 2.2;
- построить график, на котором отобразить наблюдения, исходную функцию и линию регрессии;
- проверить для подобранной модели является ли матрица $H^T H$ мультиколлинеарной, если да, то построить оценку параметров с помощью метода редукции (ридж-оценка);

1.2.4 Регрессия для наблюдений с выбросами

Смоделировать ошибки для модели регрессии $X_k = \theta_0 + \theta_1 h_k + \varepsilon_k$ с помощью распределения Тьюки, приняв долю выбросов $\delta = 0.08$, номинальную регрессию $\sigma_0^2 = \sigma^2$, дисперсию аномальных наблюдений $\sigma_1^2 = 100\sigma^2$.
Построить МНК-оценку неизвестных параметров модели и оценить ее качество.
Провести анализ остатков по схеме их пункта 2.2.
Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессии.
Провести отбраковку выбросов, пересчитать МНК-оценку и оценить качество оценки.
После отбраковки построить новый график, на котором отобразить наблюдения, исходную функцию и линию регрессии.
Провести анализ остатков по схеме их пункта 2.2.
Построить оценку метода наименьших модулей.
Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессии метода наименьших модулей.
Провести анализ остатков по схеме их пункта 2.2.
Дополнительно: построить робастную оценку Хубера.

1.2.5 Квантильная регрессия

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

2 Байесовский подход в регрессионном анализе

Введение.

Байесовский подход в регрессии в отличие от классической статистики – принципиально другой подход к оцениванию неизвестных коэффициентов. Суть байесовской регрессии заключается в задании неизвестного параметра θ в виде априорного закона распределения.

Из предметной области может быть известно множество определения неизвестного параметра (возможно более широкого, чем истинное множество определения). Например, если смысл параметра p – вероятность какого-либо события, то $p \in [0; 1]$, либо, если p – спрос на некий товар или услугу, то параметр должен быть неотрицательным $p \in [0; \infty)$

Априорная плотность распределения неизвестного параметра определяется экспертно.

После получения новых наблюдений априорные значения и данные позволяют получить обновленную апостериорную модель, учитывающую новые наблюдения.

Следует отметить, что модели регрессии (например, линейная регрессия, полиномиальная регрессия, медианная регрессия, квантильная регрессия, регрессия Тьюки и т.д.) могут быть такими же, как и в стандартной статистике. Меняется лишь способ моделирования неизвестных параметров. В байесовском подходе считается, что неизвестные параметры – случайные величины, которые имеют априорные законы распределения.

Для применения байесовской регрессии на заданном наборе данных достаточно определить:

- модель для данных, например линейная регрессия;
- априорное распределение неизвестных параметров.

После по формуле условной вероятности $f(\theta|y) = \dfrac{f(y|\theta) \cdot f(\theta)}{f(y)}$ получают апостериорное распределение параметров $f(\theta|y)$.

Замечание: поскольку $f(y)$ не зависит от неизвестных параметров, то по этой причине принято записывать формулу условной вероятности в следующем виде: $f(\theta|y) \sim f(y|\theta) \cdot f(\theta)$. Величина $f(y)$ фактически является нормировочным множителем, который можно вычислить по свойству интеграла от плотности вероятности: $\int\limits_{-\infty}^{+\infty} f(y) = 1$

Априорное распределение вероятностей неизвестных параметров носит смысл начального приближения, при этом, если априорное распределение выбрано верно, то при добавлении новых данных модель становится более точной по сравнению с начальным приближением.

Пример 1.

В качестве примера рассмотрим следующую задачу: пусть есть счетчик, расположенный рядом с некоторой оживленной улицей, который умеет классифицировать машины по признаку такси/не такси. Пусть с вероятностью p мы наблюдаем такси, тогда с вероятностью $1-p$ мы будем наблюдать не такси. Данная схема представлет собой схему Бернулли.

Пусть были произведены наблюдения: $Y = (1, 1, 0, 1, 0, 1)^T$, где 1 – наблюдение такси, а 0 – обратное. Возьмем равномерное априорное распределение на отрезке $[0; 1]$.

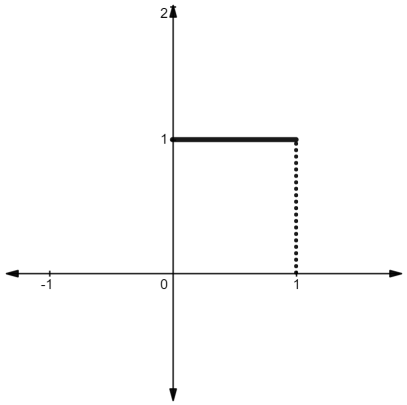


Рис. 1: априорная плотность вероятности параметра

С учетом наблюдений Y вычислим скорректированную плотность вероятности для параметра $f(p|Y)$ – апостериорная плотность распределения параметра p.

По формуле условной плотности вероятности $f(p|Y) = \dfrac{f(p,Y)}{f(Y)} = \dfrac{f(Y|p) \cdot f(p)}{f(Y)} \sim$ $\sim f(Y|p) \cdot f(p)$. Так как y_i независимы и одинаково распределены, то $f(Y|p) = p \cdot p \cdot (1-p) \cdot p \cdot (1-p) \cdot p = p^4 \cdot (1-p)^2$. А $f(p) = 1$ по априорному предположению.

Следовательно, $f(p|Y) \sim p^4 \cdot (1-p)^2$. Воспользуемся свойством интеграла от плотности вероятности $\int\limits_{-\infty}^{+\infty} f(x)dx = 1 \Longrightarrow \int\limits_{0}^{1} C \cdot p^4 \cdot (1-p)^2 dp = 1 \Longrightarrow C \cdot \dfrac{1}{105} = 1 \Longrightarrow C = 105 \Longrightarrow$ $f(p|Y) = 105 \cdot p^4 \cdot (1-p)^2$.

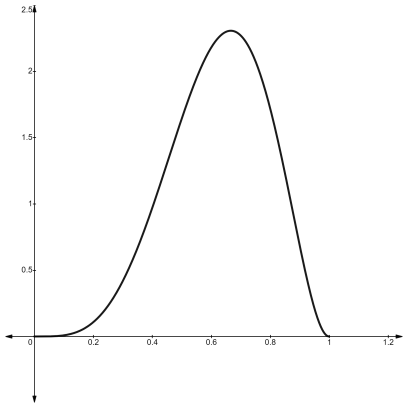


Рис. 2: апостериорная плотность вероятности параметра

Из графиков видно, что математическое ожидание в апостериорной плотности больше, чем в априорной, которая равна 0.5. Данный результат не удивителен, поскольку в исходной выборке оказалось больше наблюдений $'1'$, чем $'0'$.

Пример 2.

В условии предыдущего примера сделаем априорное предположение: в ночное время вероятность обнаружить такси больше, чем обычную машину. Изменим априорную плотность вероятности параметра так, чтобы соблюдалось априорное предположение.

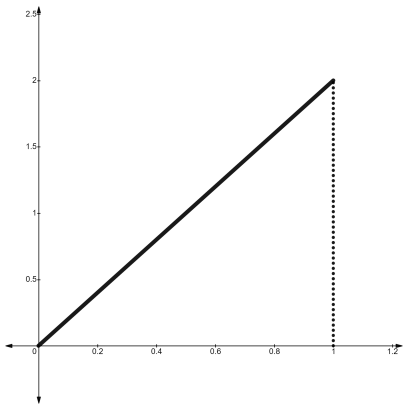


Рис. 3: априорная плотность вероятности параметра

Для того же набора данных вычислим апостериорную функцию плотности вероятности. $f(p|Y) \sim p^4 \cdot (1-p)^2 \cdot 2p$. Воспользовавшись свойством нормированности функции плотности распределения, вычислим точное значение $f(p|Y) = 168 \cdot p^5 \cdot (1-p)^2$.

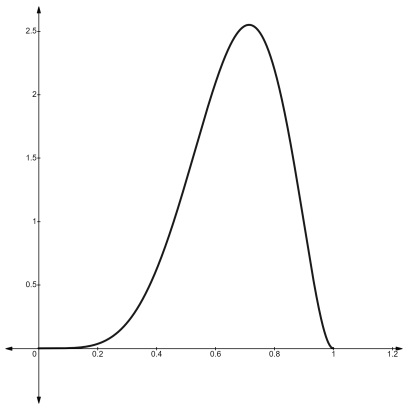


Рис. 4: апостериорная плотность вероятности параметра

Вычислим математические ожидания оцениваемого параметра для двух случаев: $M[p] = \frac{2}{3}$, $M[p|Y] = \frac{2}{3}$. Математические ожидания не изменились для обоих случаев они в точности совпадают, что обуславливается тем, что априорное предположение и наблюдения согласуются.

Вычислим вероятности: $P(p > 0.5) = \int\limits_{0.5}^{1} 2pdp = 0.75$, $P(p > 0.5|Y) = 0.855$. Вероятность в апостериорной модели возросла. Следовательно априорное предположение, что $p > 0.5$ стало 'более вероятным'.

Заключение.

Байесовский подход в оценивании параметров является принципиально новым подходом. Оцениваемые параметры представляют сабой случайные величины. Из положительных сторон Байесовского подхода можно отметить, что он позволяет просто проверять гипотезы, например $P(\theta > 0|Y)$, $M[\beta|Y] = 10$, в классическом подходе данные гипотезы не имеют смысла. Также можно отметить, что апостериорное распределение всегда существует. Байесовский подход работает даже на очень небольших выборках и каждое новое значение будет изменять апостериорную плотность вероятности. Среди минусов такого подхода можно отметить, что он требует большого объема вычислений.

3 Практическая часть

Листинг программы доступен по ссылке: https://github.com/KorolevEgor/Econometrics.

3.1 Модельная часть

Для заданной функции $f(h) = 1.5h - 2 - \frac{1}{2h}$ смоделируем обучающую выборку $f(h_k) + \varepsilon_k$. Точки h_k выбраны равномерно на отрезке $[0.1; 2]$, изменяется от 1 до 60. Аналогично смоделируем тестовую выборку с количеством наблюдений равным 40. В качетсве параметров нормального распределения ошибок ε было выбрано: $\mu = 0, \sigma = 1$. Тестовая выборка была сгенерирована только по одну сторону от обучающей, поскольку заданная функция имеет полюс первого порядка в точке $h = 0$, и, следовательно функция при приближении к этой точке быстро растет, поэтому при генерировании точек вблизи $h = 0$ значения могут быть крайне большими по модулю по сравнению с обучающей выборкой.

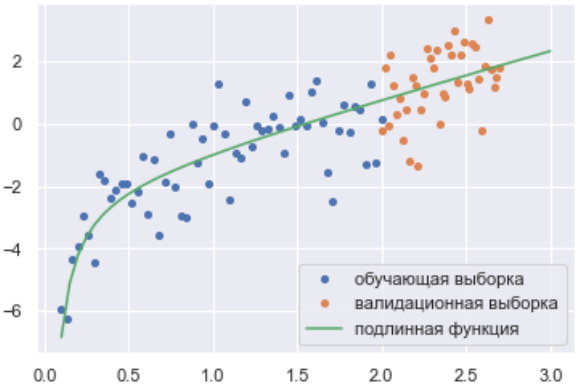


Рис. 5: Обучающая и валидационная выборки

3.2 Метод наименьших квадратов

Для модели простой линейной регрессии $X_k = \theta_0 + h\theta_1$ построим оценки МНК и МНМ, измерим качество, построенных моделей.

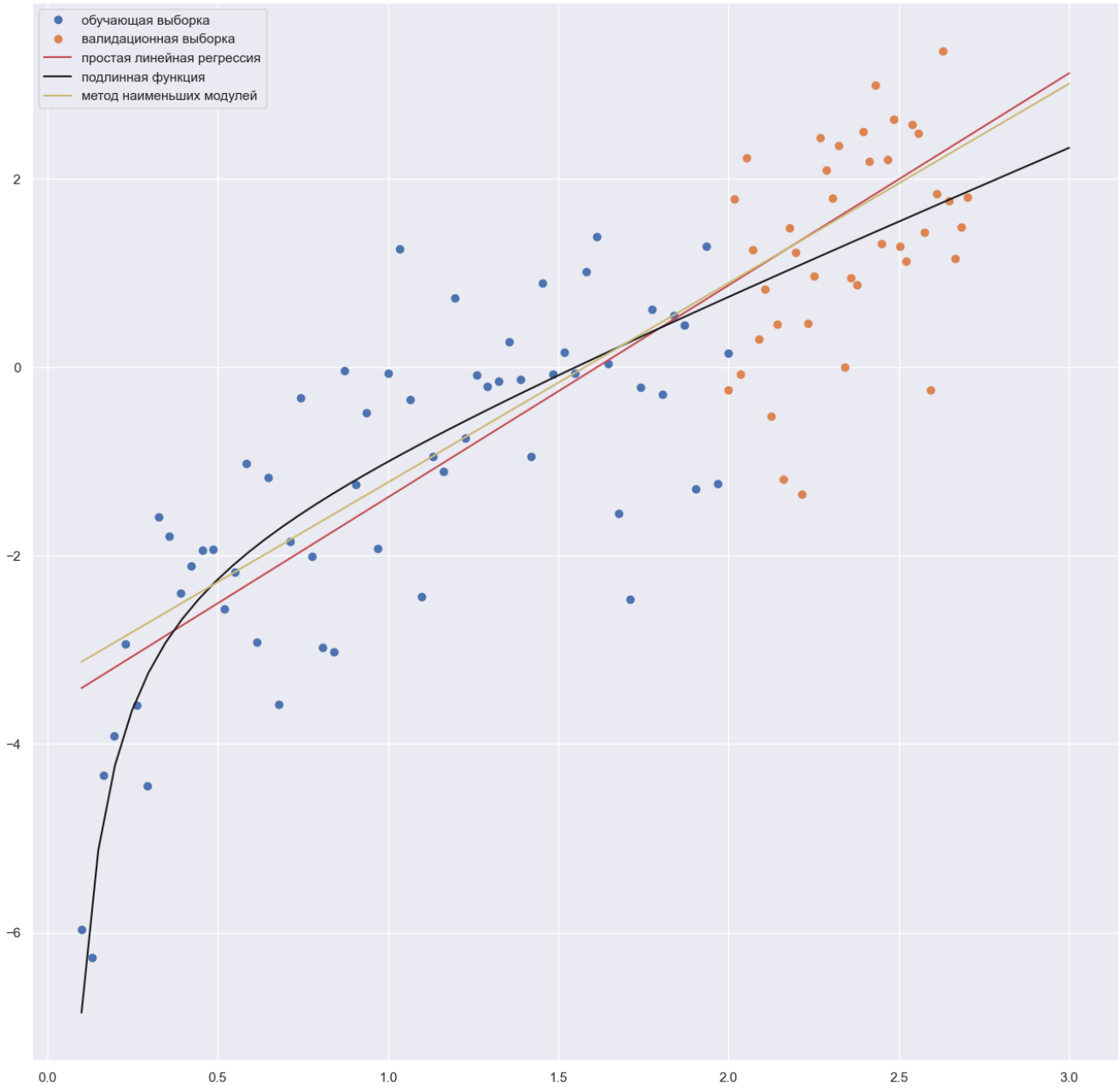


Рис. 6: Линии регрессии МНК и МНМ простой линейной регрессии

	МНК	МНМ
Уравнения прямых	$x = -3.63 + 2.25h$	$x = -3.34 + 2.12h$
R^2	0.55	0.55
RMSE	1.15	1.16
$\sum_i \varepsilon_i^2$ (на обуч. выборке)	78.87	80.58
$\sum_i \varepsilon_i^2$ (на тест. выборке)	44.22	43.63

Таблица 1: Сравнение моделей МНК и МНМ

Некоторые измерения над моделью с МНК.

Оценка ковариационной матрицы:

$$\hat{K} = \begin{pmatrix} 0.10 & -0.08 \\ -0.08 & 0.08 \end{pmatrix}$$

След оценки ковариационной матрицы $tr = 0.178$.

Функция логарифмического правдоподобия $l = -94.88$; информационный критерий Акаике $AIC = 0.39$; скорректированный (для малых выборок) $AIC_c = 0.6$; критерий Шварца $BIC = 197.95$.

Гипотеза, что $\forall i\ \theta_i = 0$ не принялась критерием Фишера на уровне значимости $\alpha = 0.05$. Гипотеза, что $\theta_n = 0$ не принялась критерием Фишера на уровне значимости $\alpha = 0.05$.

Для проверки мультиколлинеарности матрицы $H^T H$ был использован коэффициент инфляции дисперсии (VIF).

Для данной модели $VIF = (4.54, 1.0)$, следовательно проблема мультиколлинеарности методом VIF не обнаружена.

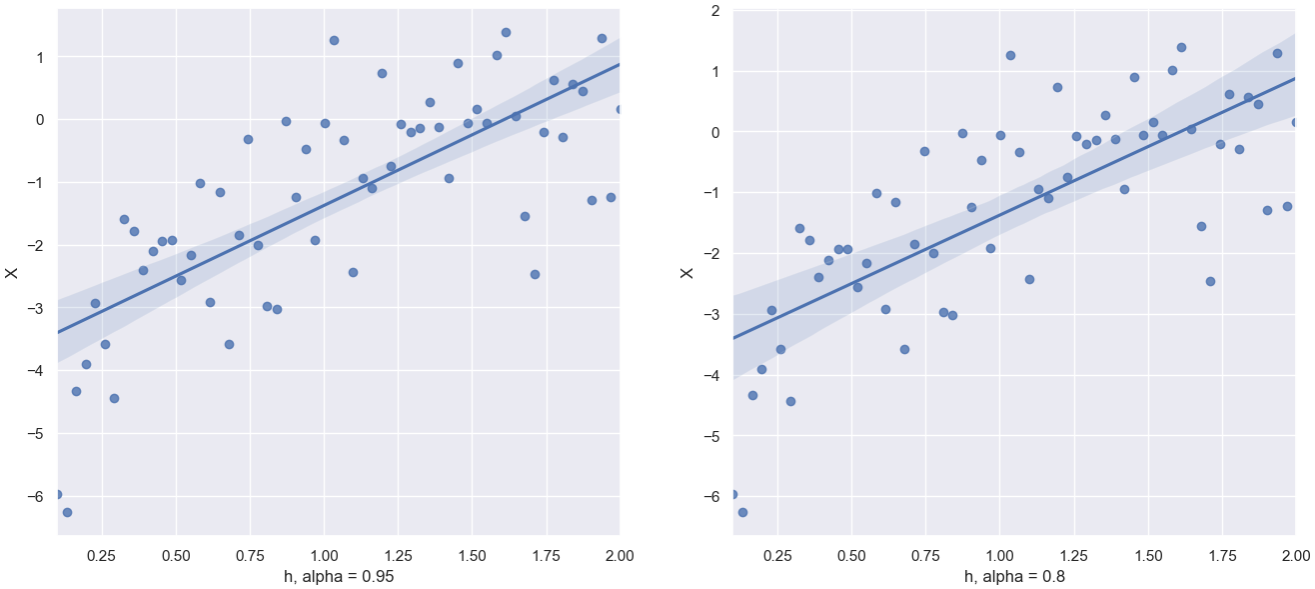


Рис. 7: Доверительные интервалы для X с надежностью 0.8 и 0.95

Анализ остатков.

Критерий Шапиро-Уилка: $T = 0.98, pvalue = 0.29$.
Гипотезу о нормальном распределении ошибок на уровне значимости 0.05 не удается принять.

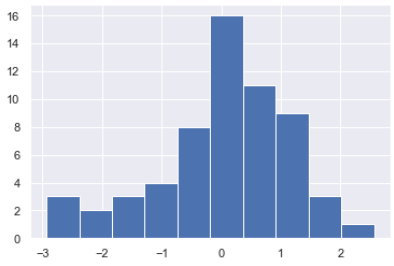


Рис. 8: Гистограмма ошибок

Значение статистики Дарбина-Уотсона = 1.7.
Выборочный коэффициент корреляции $r = 0.15$.
Гопотеза о некоррелированности принимается.

Критерий Бройша-Пагана: $(T_1 = 2.08, pvalue_1 = 0.15), (T_2 = 2.08, pvalue_2 = 0.15)$.
Гипотеза о гетероскедастичности принимается на уровне значимости 0.05.

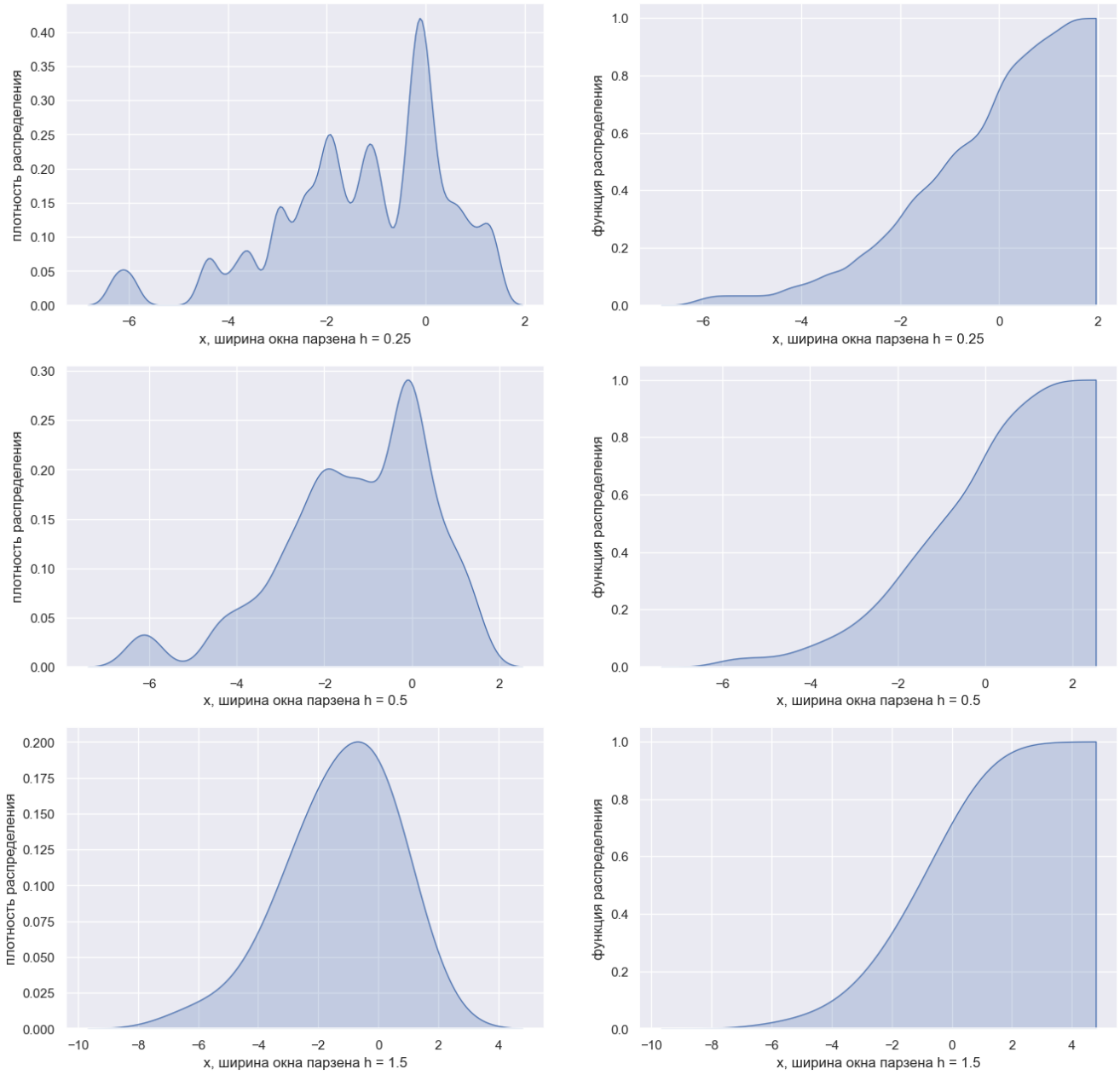


Рис. 9: Ядерные оценки плотности и распределения вероятности

Промежуточные итоги.

В данной части работы было построено 2 модели простой линейной регрессии. В первой модели была использована квадратичная функция потерь, во второй был использован модуль ошибки в качестве функции потерь. Значения коэффициентов оказались близкими. Коэффициенты детерминации отличаются в 7-м знаке после запятой. При этом значение суммы квадратов отклонений на обучающей выборке немного лучше у МНК, чем у МНМ, однако на тестовой выборке наоборот. Поскольку мощность выборки не велика (60 наблюдений), был использован скорректированный критерий Акаике, который используется при малых выборках, когда число наблюдений к числу параметров меньше 40. Также были проверены гипотезы: о равенстве нулю всех параметров, о равенстве нулю последнего параметра (обе гипотезы отверглись критерием Фишера на $\alpha = 0.05$), о нормальном распределении (отвергласъ критерием Шапиро-Уилка на $\alpha = 0.05$), о автокорреляции (отвергалась критерием Дарбина-Уотсона), о гетероскедастичности (принялась критерием Бройша-Пагана на $\alpha = 0.05$). Построил графики ядерных оценок плотности вероятности и распределения вероятностей для разных широт окна Парзена, чем больше ширина окна, тем график более гладкий. В качестве ядра было использовано гауссово ядро.

3.3 Полиномиальная регрессия

Рассмотрим 6 моделей полиномиальной регрессии:

Количество параметров у модели (без учета θ_0):	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$
Алгебраический вид:	$X = \sum_{i=0}^{1} \theta_i h^i$	$X = \sum_{i=-1}^{1} \theta_i h^i$	$X = \sum_{i=-1}^{2} \theta_i h^i$	$X = \sum_{i=-1}^{3} \theta_i h^i$	$X = \sum_{i=-1}^{4} \theta_i h^i$	$X = \sum_{i=-1}^{5} \theta_i h^i$

Таблица 2: Определение моделей

Заметим, что модель, соответствующая $p = 1$, представляет собой модель простой линейной регрессии, рассмотренной на предыдущем шаге. В данной части она включена для наглядности сравнения с полиномиальными моделями с бОльшим числом параметров.

Обучив данные модели, были получены следующие МНК-оценки параметров:
Модель 1: $[-3.63, 2.25]$;
Модель 2: $[-1.51, -0.52, 1.05]$;
Модель 3: $[-2.76, -0.37, 3.13, -0.83]$;
Модель 4: $[-0.2, -0.60, -3.94, 5.62, -1.81]$;
Модель 5: $[1.3, -0.71, -9.77, 14.36, -7.24, 1.18]$;
Модель 6: $[-7.78, -0.14, 36.02, -83.64, 91.56, -45.23, 8.18]$.

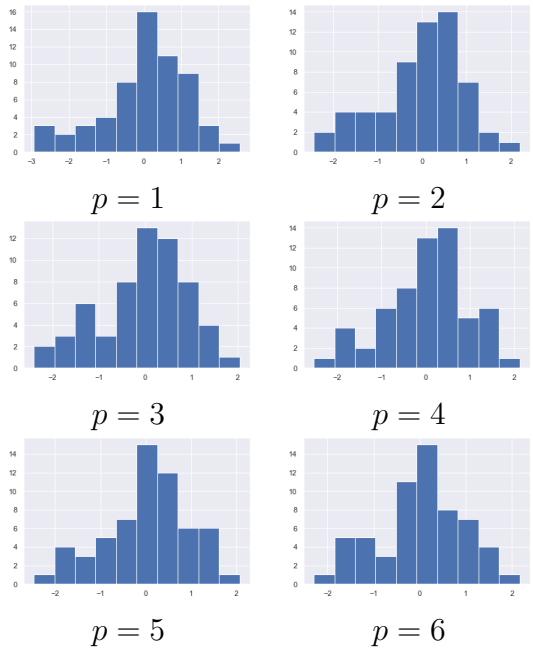
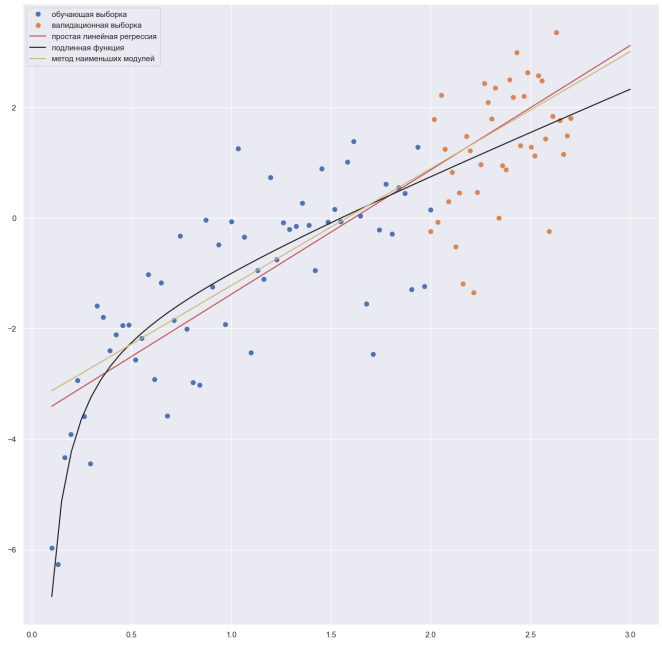


Рис. 10: Гистограммы ошибок моделей полиномиальных регрессий

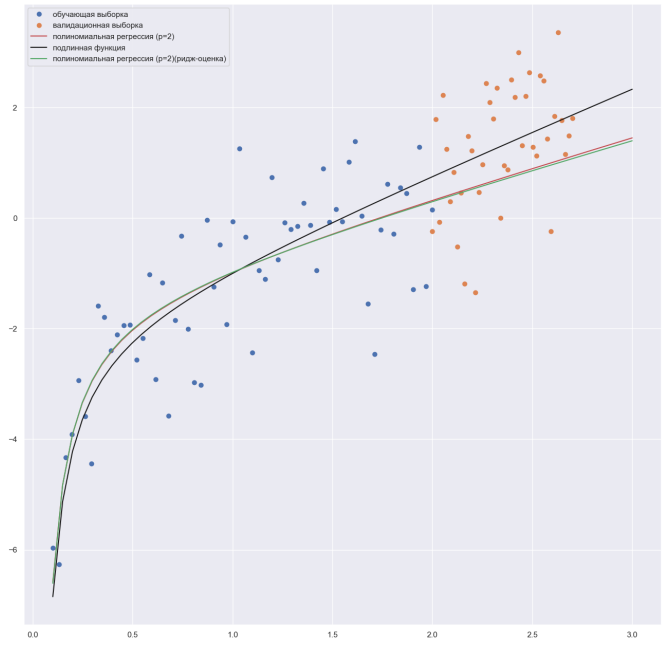
Сводная таблица с гипотезами и метриками качества моделей:

Номер модели	1	2	3	4	5	6
Ср. кв. погрешность МНК-оценки	1.15	0.96	0.95	0.94	0.93	0.92
Гипотеза $\theta_p = 0$ (по Фишеру)	-	-	-	-	-	+
Гипотеза $\forall i\ \theta_i = 0$ (по Фишеру)	-	-	-	-	-	+
MSE (обуч. выборка)	78.87	55.49	54.06	52.47	52.31	52.8
MSE (тест. выборка)	44.22	54.85	145.1	794.14	216.31	16591.33
мультиколлинеарность матрицы $H^T H$	-	+	+	+	+	+
MSE ridge (обуч. выборка)	-	55.5	54.82	53.26	52.74	52.8
MSE ridge (тест. выборка)	-	56.54	76.21	264.88	793.6	1195.54
$tr\hat{K}$	0.18	0.39	4.91	66.94	978.41	14265.62
R^2	0.55	0.68	0.69	0.7	0.7	0.71
Гипотеза о нормальном распределении ошибок	-	-	-	-	-	-
Гипотеза об автокорреляции	+	-	-	-	-	-
Гипотеза о гетероскедастичности	+	+	+	+	+	+

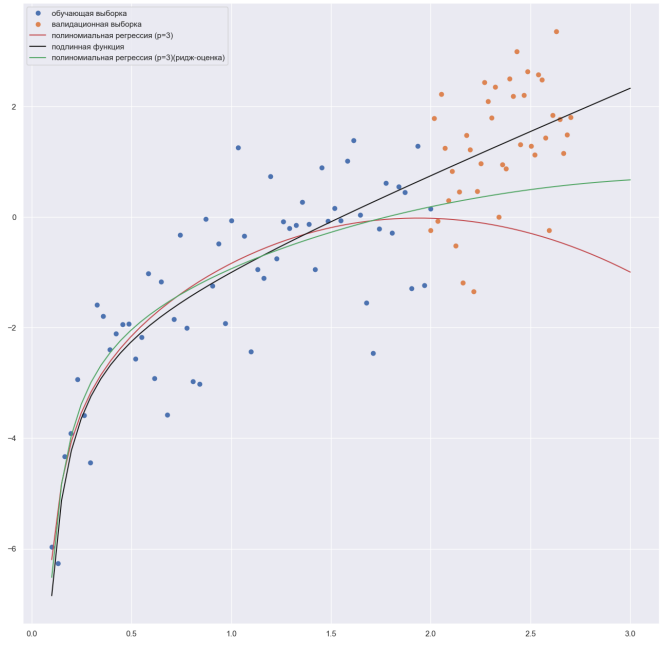
Таблица 3: Сводная таблица полиномиальных регрессий



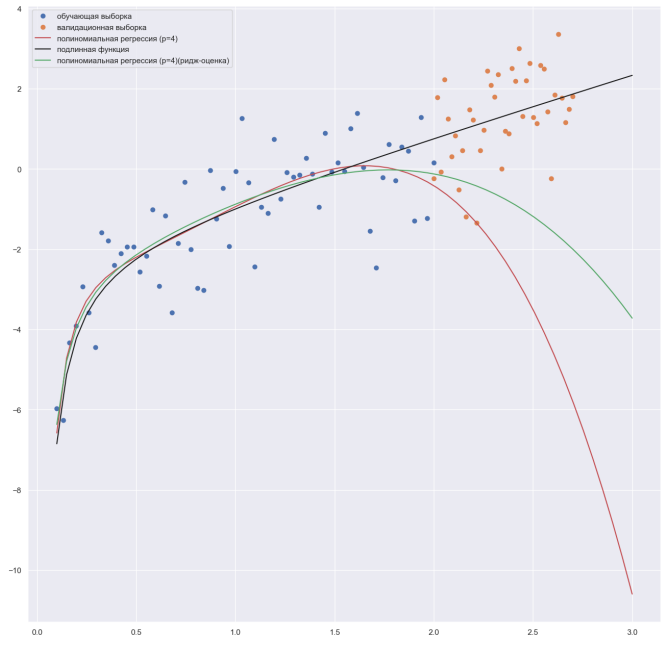
$p = 1$



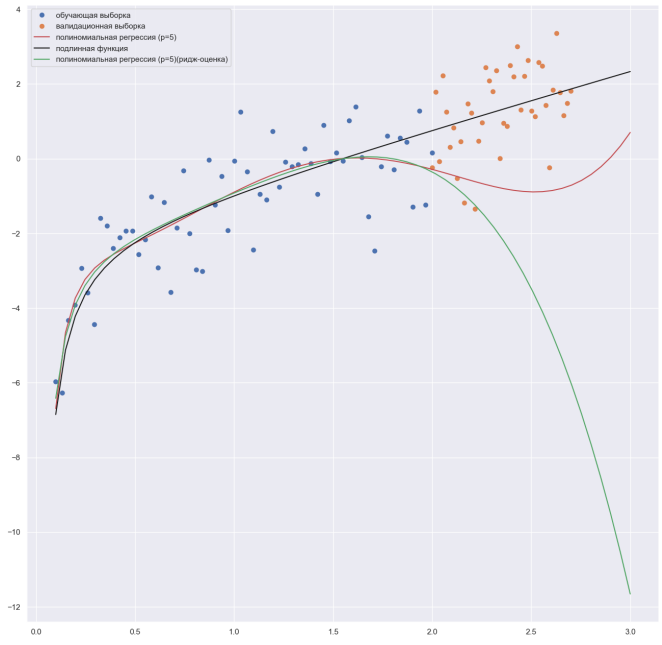
$p = 2$



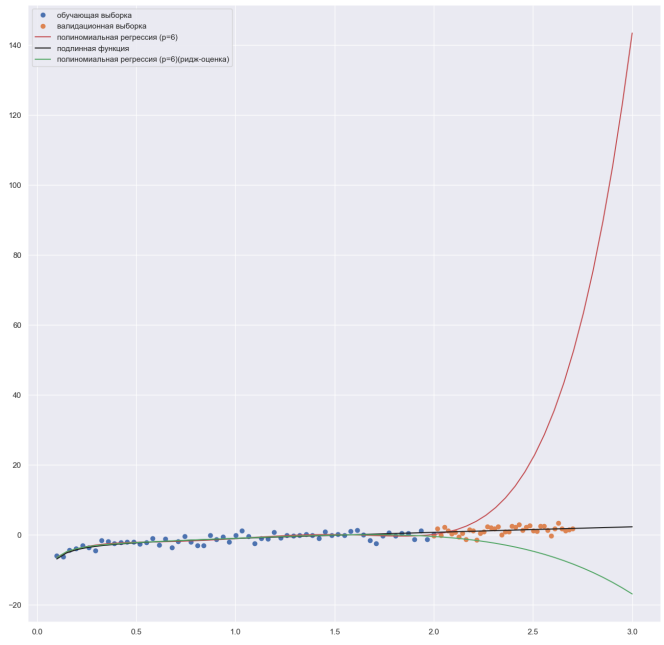
$p = 3$



$p = 4$



$p = 5$



$p = 6$

Рис. 11: Линии регрессий

Промежуточные итоги.

В данной части было проанализированно 6 моделей полиномиальной регрессии. По данным таблицы 3 можно заметить классический пример эффекта переобучения: MSE на обучающей выборке с ростом числа параметров незначительно уменьшается, при этом MSE на валидационной выборке быстро растет. Во всех моделях, кроме первой матрица $H^T H$ мультиколлинеарна, для таких моделей были построены ridge-оценки. На обучающей выборке MSE-оценки ridge моделей примерно совпадают с MSE-оценками МНК моделей, однако на валидационной выборке данные модели имеют существенно более низкие показатели. Следует заметить, что для пятой модели MSE ridge хуже MSE МНК, это можно легко объяснить по графику 11 при $p = 5$: ridge-оценка привела к уменьшению значения параметра при старшей степени, что повлекло к развороту многочлена вниз (зеленая кривая), в то время расстояние от кривой, соответствующей МНК-оценке, до валидационной выборки многократно меньше.

Начиная с $p = 3$ trK, MSE, MSE ridge на тестовой выборке начинают ухудшаться. R^2 практически не изменяется. Следовательно, оптимальным значением для параметра полиномиальной регрессии является $p = 2$. Однако, судя по графикам на рисунке 11 для $p = 2$ модели не хватило данных для обучения, поскольку кривая с приблизительно $h = 1.5$ начинает отклоняться вниз от искомой зависимости, а простая линейная регрессия не смогла описать зависимость данных около $h = 0$ и также отклоняется от искомой зависимости, но уже в большую сторону. Возможно, имеет смысл скорректировать параметр при h у модели с $p = 2$ и взять его как среднее арифметическое параметров при h у моделей $p = 1$ и $p = 2$.

Гипотезу о гетероскедастичности не удалось отвергнуть ни у одной модели. Гипотеза о нормальном распределении ошибок была отвергнута во всех моделях. Гипотеза об автокорреляции была отвергнута во всех моделях, кроме первой.

3.4 Регрессия для наблюдений с выбросами

Смоделируем распределение Тьюки. Для этого сгенерируем 3 массива случайных чисел: u, v, w, выберем долю выбросов $\delta = 0.08$ и силу выбросов $sigma_1 = 10\sigma$.

- $u = (u_1, \ldots, u_{60}), \quad u_i \sim \mathcal{N}(0, \sigma^2)$ – обычные наблюдения;
- $v = (v_1, \ldots, v_{60}), \quad v_i \sim \mathcal{N}(0, (10\sigma)^2)$ – выбросы;
- $w = (w_1, \ldots, w_{60}), \quad w_i \sim \mathcal{R}(0, 1)$ – вектор вероятностей.

Сформируем из сгенерированных массивов выборку по следующему алгоритму:

```
X = [ ]
for i in range(len(u)):
    if w[i] < delta:
        X.append(f(h[i]) + u[i])
    else:
        X.append(f(h[i]) + v[i])
```

Аналогично сгенерируем тестовую выборку.
Результат моделирования (график приведен в логарифмическом масштабе по вертикальной оси); также построим модели простой линейной модели для МНК и МНМ на данных с выбросами:

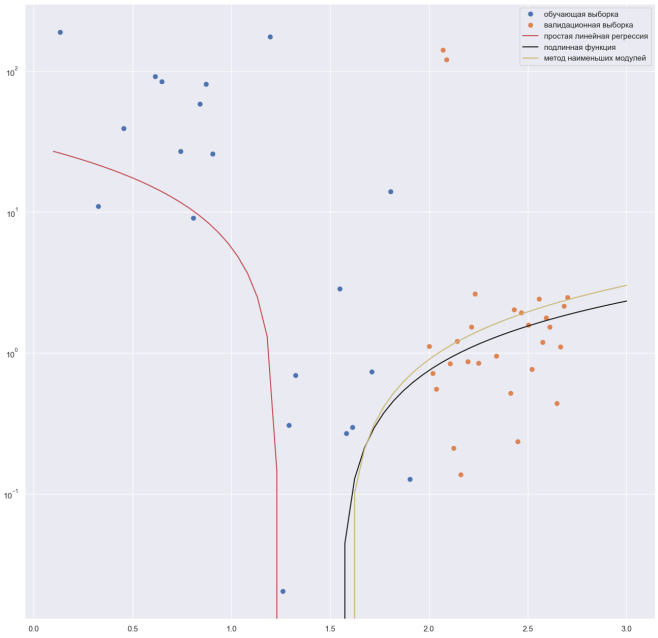


Рис. 12: Линии регрессий для наблюдений с выбросами

	МНК	МНМ
Уравнения прямых	$x = 29.34 - 23.73h$	$x = -3.36 + 2.11h$
R^2	0.07	0.07
RMSE	47.9	50.33
$\sum_i \varepsilon_i^2$ (на обуч. выборке)	137649.99	151957.64
$\sum_i \varepsilon_i^2$ (на тест. выборке)	96192.08	73400.67

Таблица 4: Сравнение моделей МНК и МНМ на данных с выбросами

Некоторые измерения над моделью с МНК.

Оценка ковариационной матрицы:

$$\hat{K} = \begin{pmatrix} 182.92 & -135.87 \\ -135.87 & 129.4 \end{pmatrix}$$

След оценки ковариационной матрицы $tr = 312.32$.

Функция логарифмического правдоподобия $l = -318.82$; информационный критерий Акаике $AIC = 7.86$; скорректированный (для малых выборок) $AIC_c = 8.07$; критерий Шварца $BIC = 645.83$.

Гипотеза, что $\forall i\ \theta_i = 0$ принялась критерием Фишера на уровне значимости $\alpha = 0.005$. Гипотеза, что $\theta_n = 0$ принялась критерием Фишера на уровне значимости $\alpha = 0.005$.

$VIF = (4.54, 1.0)$, следовательно проблема мультиколлинеарности методом VIF не обнаружена.

Анализ остатков.

Критерий Шапиро-Уилка: $T = 0.82$, $pvalue = 3.81 \cdot 10^{-7}$.
Распределение ошибок нормальное на уровне значимости 0.05.

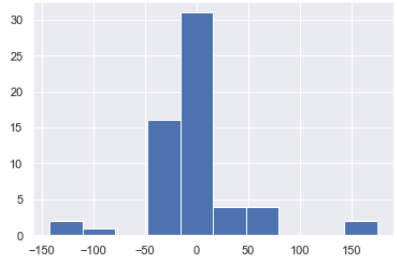


Рис. 13: Гистограмма ошибок для наблюдений с выбросами

Значение статистики Дарбина-Уотсона $T = 1.99$.
Выборочный коэффициент корреляции $r = 0.003$.
Гипотеза о некоррелированности принимается.

Критерий Бройша-Пагана: $(T_1 = 2.11, pvalue_1 = 0.35), (T_2 = 1.04, pvalue_2 = 0.36)$.
Гипотеза о гетероскедастичности принимается на уровне значимости 0.05.

Отбраковка выбросов.

Отбраковку выбросов будем проводить следующим образом:

1 построение регрессии Тьюки или Хьюбера;

2 отбраковка тех значений, которые лежат более чем на 3σ от регрессионной кривой Тьюки или Хьюбера;

Линейная регрессия называется регрессией Тьюки, если в качестве функции потерь взята функция потерь Тьюки:

$$\begin{cases} \frac{\delta^2}{6} \cdot \left(1 - \left(1 - \left(\frac{u}{\delta}\right)^2\right)^3\right), & |u| < \delta \\ \frac{\delta^2}{6}, & \text{иначе} \end{cases}$$

Также рассмотрим регрессию Хьюбера. Линейная регрессия называется регрессией Хьюбера, если в качестве функции потерь взята функция потерь Хьюбера:

$$\begin{cases} \frac{u^2}{2}, & |u| < \delta \\ \delta \cdot \left(|u| - \frac{\delta}{2}\right), & \text{иначе} \end{cases}$$

Построим регрессии. Регрессия Хьюбера строилась для параметра $\delta = 0.1$, регрессия Тьюки строилась для параметра $\delta = 3$.

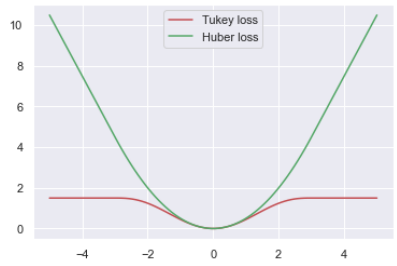


Рис. 14: Функции потерь Хьюбера и Тьюки для параметра $\delta = 3$

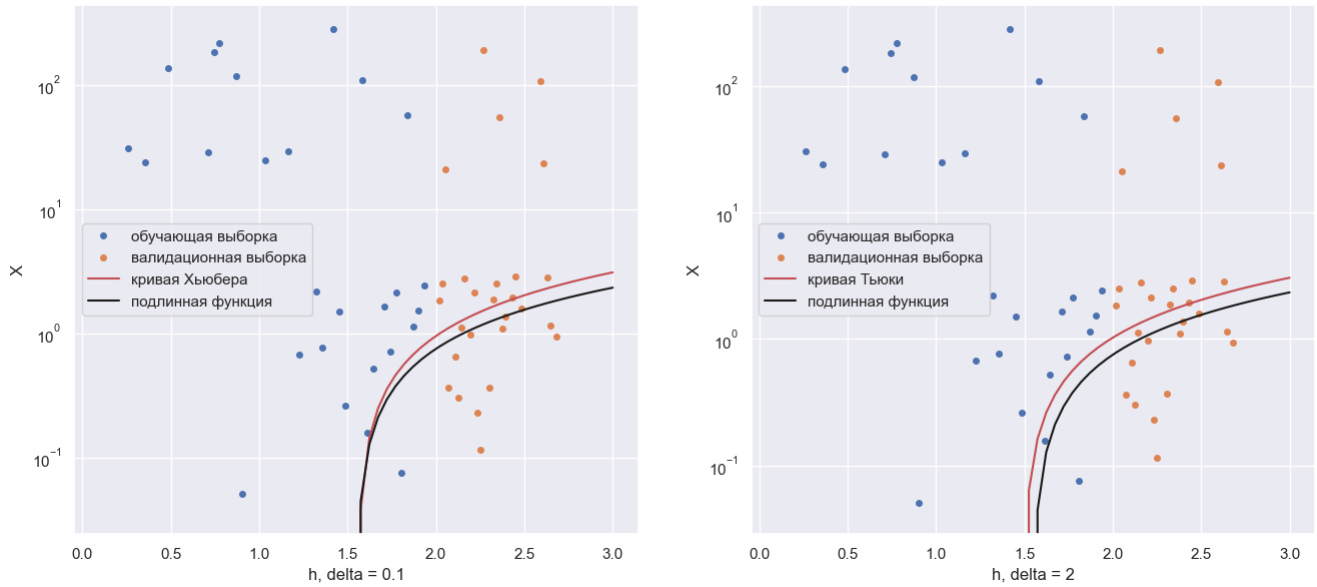


Рис. 15: Линии регрессий Хьюбера и Тьюки в логарифмическом по X масштабе

Для построенных регрессий высчитаем характеристики качества:

	Хьюбер	Тьюки
Уравнения прямых	$x = -3.34 + 2.15h$	$x = -3.02 + 2.02h$
R^2	0.55	0.55
RMSE	1.16	1.21
$\sum_i \varepsilon_i^2$ (на обуч. выборке)	80.94	87.9
$\sum_i \varepsilon_i^2$ (на тест. выборке)	45.49	46.52

Таблица 5: Сравнение моделей Хьюбера и Тьюки на данных с выбросами

Сделаем отбраковку выбросов по построенной кривой Хьюбера и построим на получившимся наборе данных МНК и МНМ оценки:

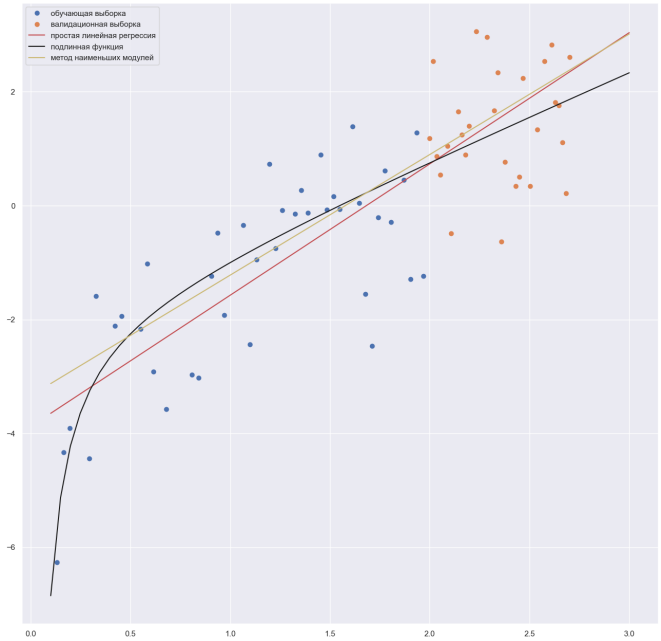


Рис. 16: МНК и МНМ оценки для данных, очищенных от выбросов

	МНК	МНМ
Уравнения прямых	$x = -3.88 + 2.31h$	$x = -3.37 + 2.12h$
R^2	0.55	0.55
RMSE	1.14	1.19
$\sum_i \varepsilon_i^2$ (на обуч. выборке)	52.08	56.77
$\sum_i \varepsilon_i^2$ (на тест. выборке)	31.22	31.54

Таблица 6: Сравнение моделей МНК и МНМ на данных, очищенных от выбросов

Некоторые измерения над моделью с МНК.

Оценка ковариационной матрицы:

$$\hat{K} = \begin{pmatrix} 0.19 & -0.13 \\ -0.13 & 0.12 \end{pmatrix}$$

След оценки ковариационной матрицы $tr = 0.3$.

Функция логарифмического правдоподобия $l = -95.39$; информационный критерий Акаике $AIC = 0.41$; скорректированный (для малых выборок) $AIC_c = 0.62$; критерий Шварца $BIC = 198.97$.

Гипотеза, что $\forall i\ \theta_i = 0$ не принялась критерием Фишера на уровне значимости $\alpha = 0.05$. Гипотеза, что $\theta_n = 0$ не принялась критерием Фишера на уровне значимости $\alpha = 0.05$.

$VIF = (5.28, 1.0)$, следовательно проблема мультиколлинеарности методом VIF не обнаружена.

Анализ остатков.

Критерий Шапиро-Уилка: $T = 0.96$, $pvalue = 0.13$.
Гипотезу о нормальном распределении ошибок на уровне 0.05 не удается принять.

Значение статистики Дарбина-Уотсона $T = 1.35$.
Выборочный коэффициент корреляции $r = 0.33$.
Гипотеза о некоррелированности отвергается.

Критерий Бройша-Пагана: $(T_1 = 0.13, pvalue_1 = 0.7), (T_2 = 0.12, pvalue_2 = 0.72)$.
Гипотеза о гетероскедастичности принимается на уровне значимости 0.05.

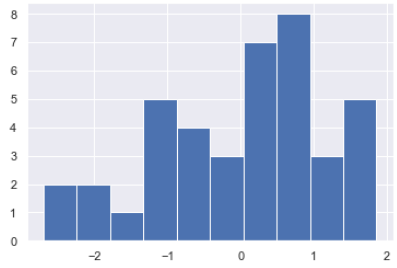


Рис. 17: Гистограмма ошибок для наблюдений, очищенных от выбросов

Промежуточные итоги.

В данной части были смоделированы данные с выбросами с помощью распределения Тьюки. Для данных с выбросами были получены МНК и МНМ оценки. Качество таких моделей оказалось крайне плохим, однако по графику видно, что МНМ-оценка находится близко к подлинной функции. Были построены модели регрессии Тьюки и Хьюбера, данные модели смогли распознать выбросы и добиться приемлемого качества как на обучающей выборке, так и на валидационной. По графику с функциями потерь Хьюбера и Тьюки можно сделать выводы: функция потерь Тьюки используется при наличие больших по модулю выбросов, поскольку штраф за ошибки, по модулю больше чем δ (параметр функции Тьюки), является константой, если же известно, что доля выбросов не велика, то имеет смысл применить функцию потерь Хьюбера, для которого вне δ-окрестности фактически применяется МНМ, в внутри окрестности МНК.

Отбраковка выбросов была сделана с помощью оценки Хьюбера: данные, отстающие более, чем на 3σ от кривой Хьюбера исключаются из выборки.

3.5 Квантильная регрессия

Смоделируем несимметричные ошибки на исходных данных, заменив 90% отрицательных ошибок на их абсолютное значение. И построим МНК и МНМ оценки:

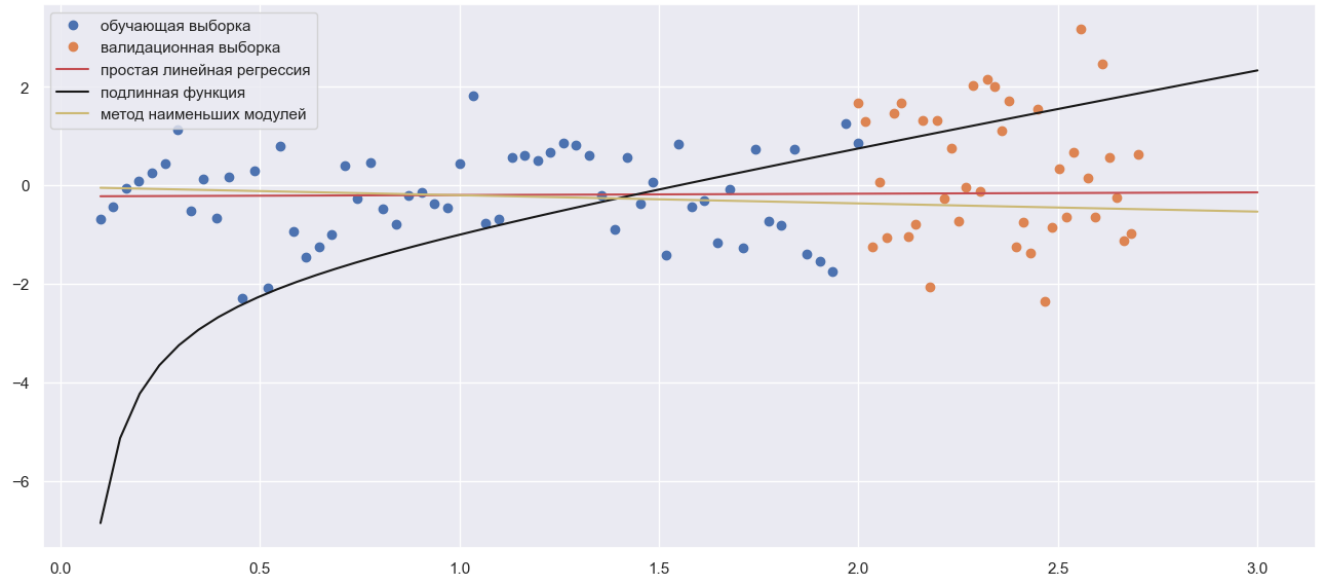


Рис. 18: МНК и МНМ оценки для данных с несимметричными ошибками

	МНК	МНМ
Уравнения прямых	$x = -0.22 + 0.03h$	$x = -0.03 - 0.17h$
R^2	0.0003	0.0003
RMSE	0.86	0.87
$\sum_i \varepsilon_i^2$ (на обуч. выборке)	45.08	45.79
$\sum_i \varepsilon_i^2$ (на тест. выборке)	77.08	88.66

Таблица 7: Сравнение моделей МНК и МНМ на данных с несимметричными ошибками

Рис. 19: Квантильная регрессия для различных значений уровня квантиля

Квантиль q	0.15	0.3	0.45	0.6	0.75	0.9
Pseudo R^2	0.04	0.011	0.003	0.0004	0.062	0.017

Таблица 8: Качество квантильных регрессия для разных уровней квантилей

Промежуточные итоги.

МНК и МНМ модели не смогли распознать несимметричные ошибки, судя по графику. Были построены модели квантильной регрессии для набора параметров квантилей (0.15, 0.3, 0.45, 0.6, 0.75, 0.9), квантильная регрессия также не смогла распознать такие ошибки для данного набора данных, что видно по графику и по данным таблицы с pseudo R^2.

4 Выводы

В ходе работы был смоделирован набор данных по заданной функции, для него были построенны МНК и МНМ оценки, вычислены метрики качества, построены доверительные интервалы, а также был проведен анализ остатков, построены ядерные оценки.

Повысить качество модели удалось за счет применения полиномиальной регрессии, ожидаемо модель $X = \sum_{i=-1}^{1} \theta_i h^i$ оказалась наиболее наилучшей. Всего было построенно 6 моделей полиномиальной регрессии, для каждой модели были вычислены метрики качества, проверены параметрические гипотезы, также был сделан анализ остатков для каждой модели.

Были смоделированы наблюдения с выбросами по распределению Тьюки. МНК и МНМ оценки оказались неудачными, поэтому были построены оценки Тьюки и Хьюбера, которые дали хорошие показатели качества на обучающей и тестовой выборках. На основе регрессии Хьюбера был реализован алгоритм по фильтрации выбросов. Полученные новые МНК и МНМ оценки оказались более точными, чем оценка Хьюбера.

Также была создана выборка с несимметричными ошибками. Для данного набора данных МНК и МНМ оценки не смогли восстановить искомую зависимость, при построение квантильной регрессии модели также не распознали асимметричные ошибки, при увеличении h разница между прогнозным значением и ожидаемым также растет.

5 Список источников

1 Лекции по Эконометрике. Е.Н.Платонов;

2 Лекции по Статистическому анализу данных. А.В.Горяинов;

3 Прикладные методы анализа статистических данных.
Е.Р.Горяинова, А.Р.Панков, Е.Н.Платонов;

4 statsmodels.org/dev/examples/notebooks/generated/;

5 mathisonian.github.io/kde/;

6 towardsdatascience.com/everything-you-need-to-know-about-multicollinearity-2f21f082d6dc;

7 ru.wikipedia.org/wiki/Информационный_критерий;

8 youtube.com/watch?v=TfIkloxgKKU;

9 towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7;

10 statswithr.github.io/book/introduction-to-bayesian-regression.html;

11 en.wikipedia.org/wiki/Bayesian_linear_regression;