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 $f(h) = 1.5h - 2 - \frac{1}{2h}$, $h \in [0.1; 2]$, $\varepsilon_k$ – 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 $\alpha$) и оценить их качество.
Построить график, на котором отобразить наблюдения, исходную функцию и линии регрессий.

# 2 Байесовская регрессия

Байесовская регрессия

# 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 $\varepsilon$ 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.

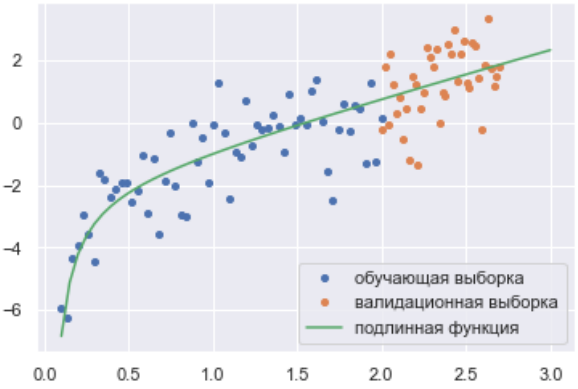



Рис. 1: Обучающая и валидационная выборки

## 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.

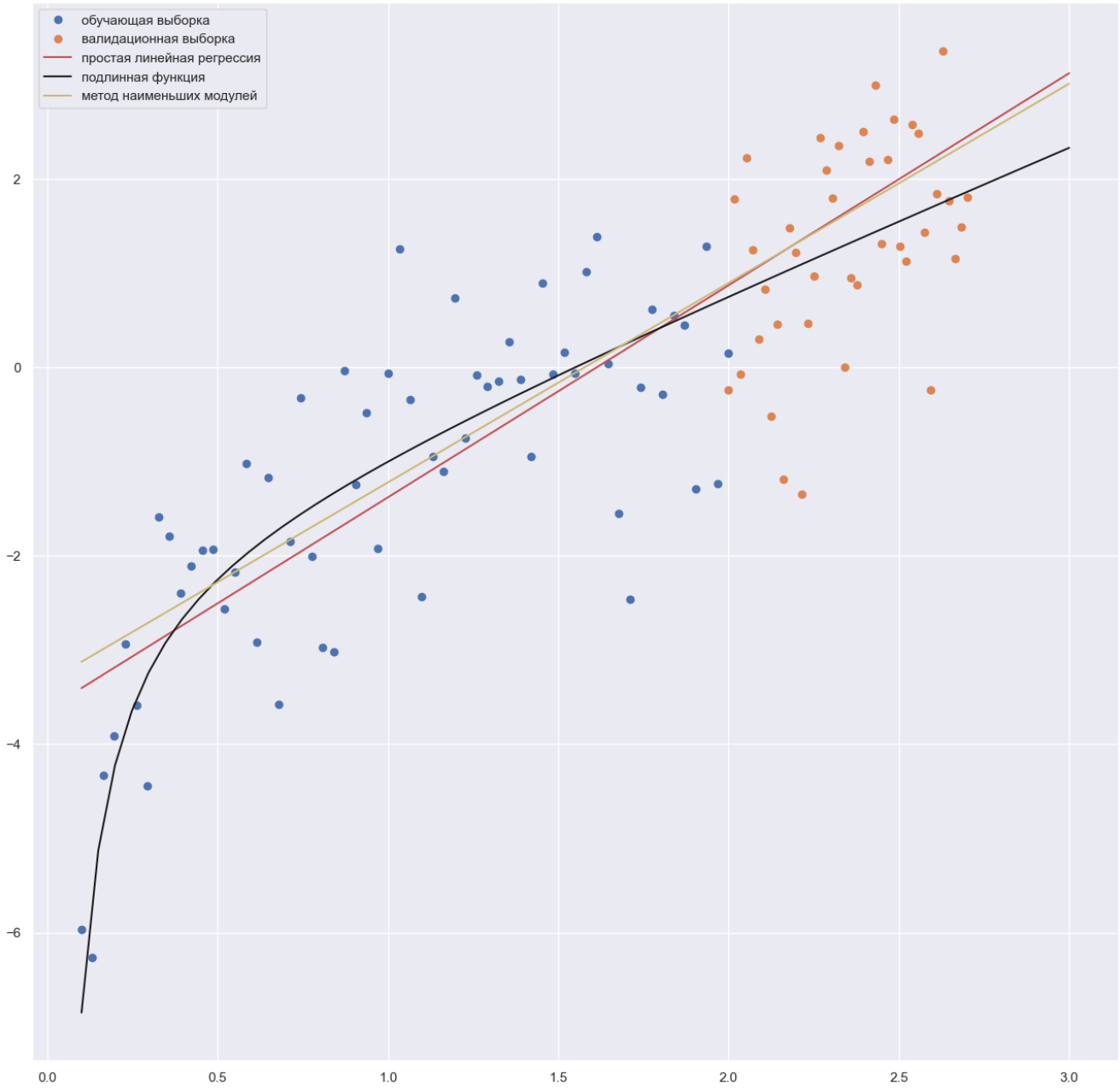



Рис. 2: Линии регрессии МНК и МНМ простой линейной регрессии

| | 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.

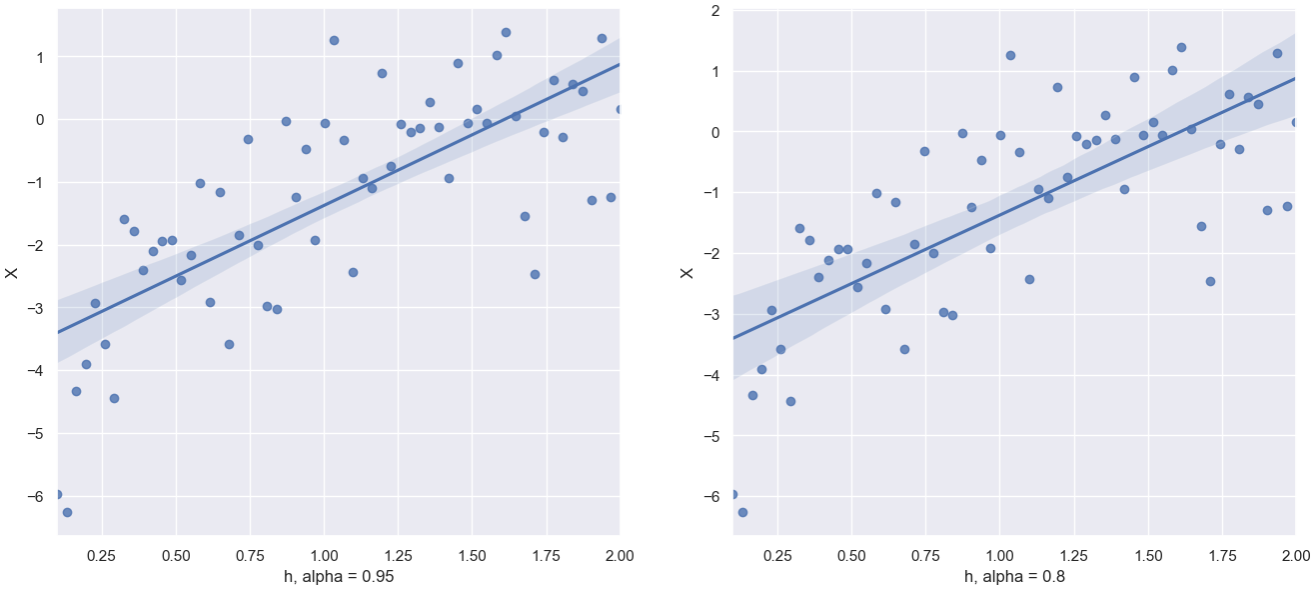



Рис. 3: Доверительные интервалы для $X$ с надежностью 0.8 и 0.95

**Анализ остатков.**

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

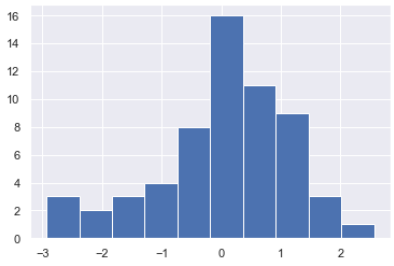


Рис. 4: Гистограмма ошибок

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.

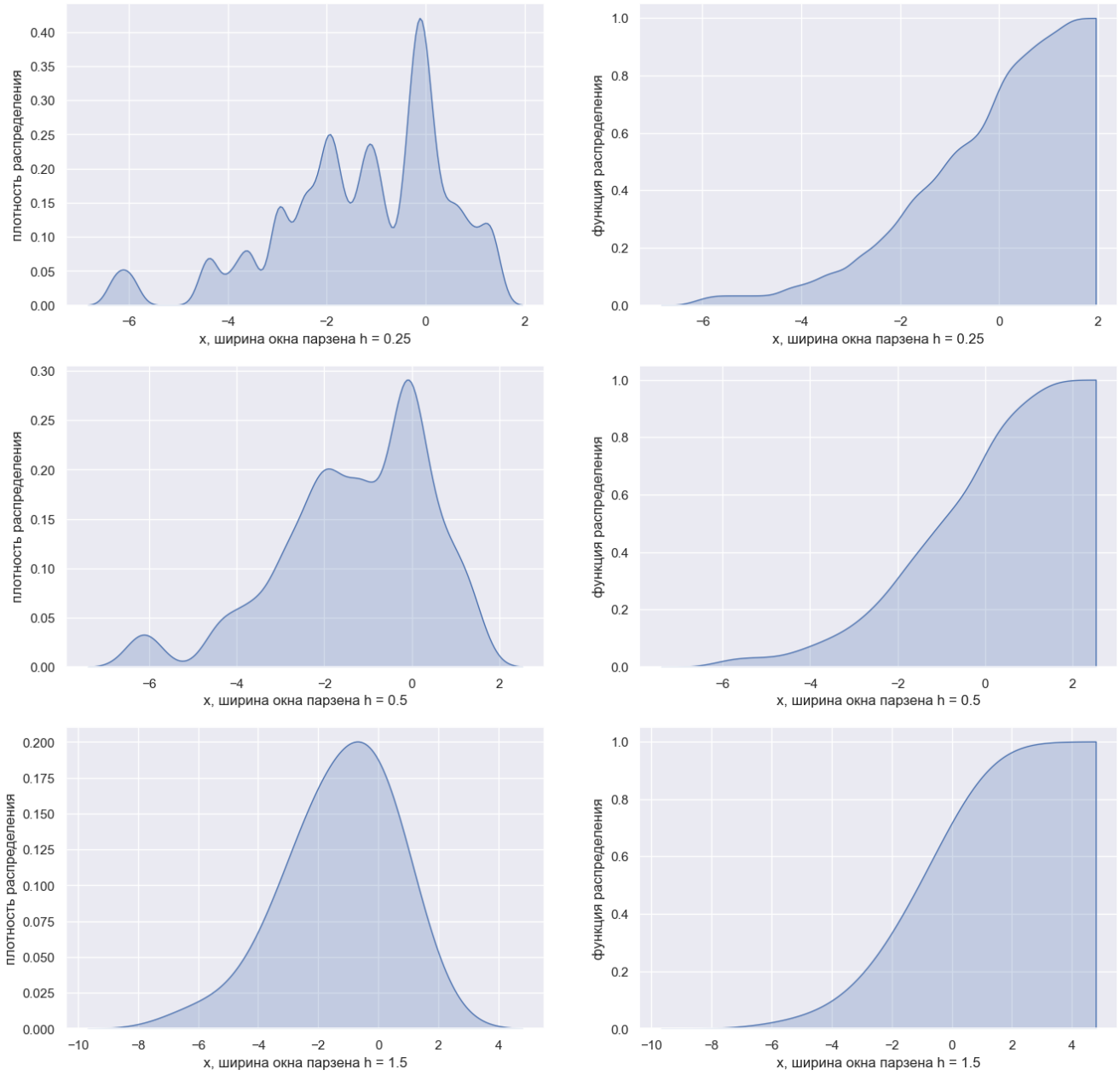



Рис. 5: Ядерные оценки плотности и распределения вероятности

**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 $\theta_0$): | $p = 1$ | $p = 2$ | $p = 3$ | $p = 4$ | $p = 5$ | $p = 6$ |
|---|---|---|---|---|---|---|
| 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: $[-7.78, -0.14, 36.02, -83.64, 91.56, -45.23, 8.18]$.

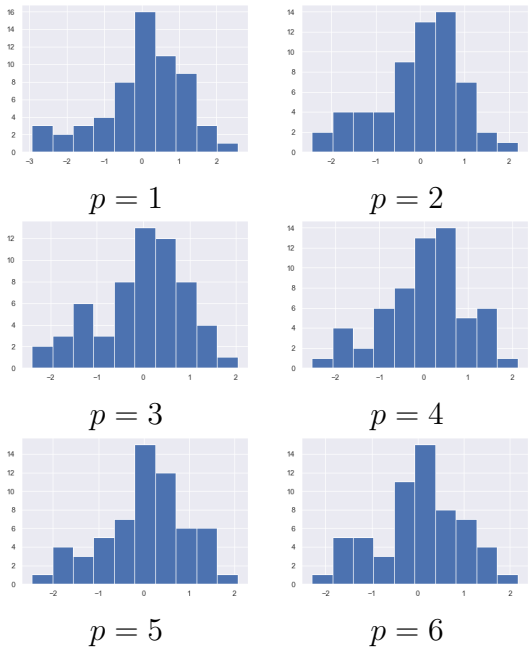
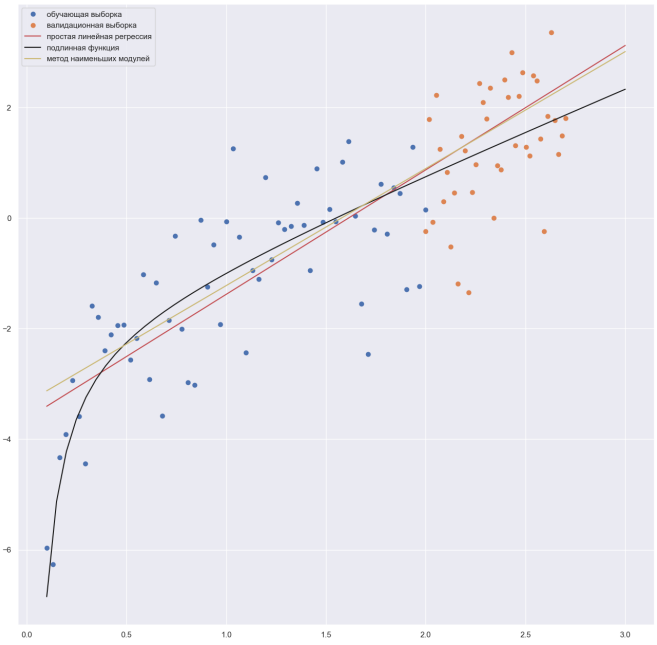



Рис. 6: Гистограммы ошибок полиномиальной регрессии

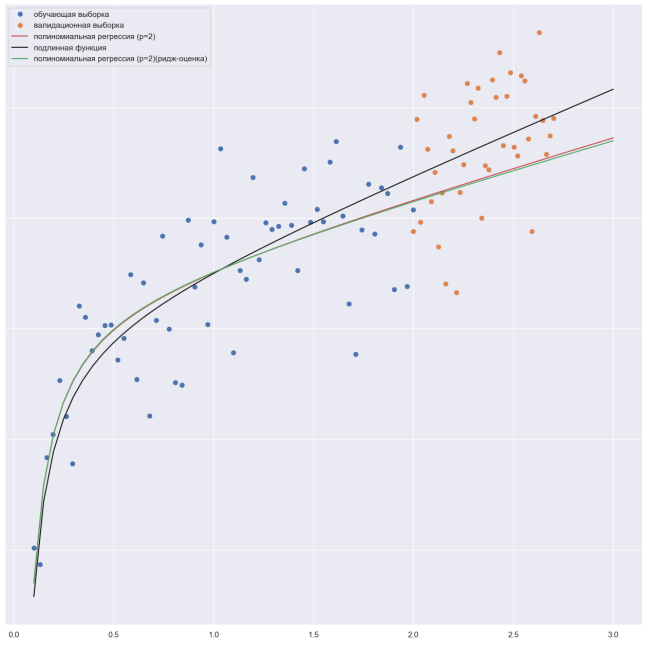
Сводная таблица с гипотезами и метриками качества моделей:

| 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 | + | + | + | + | + | + |

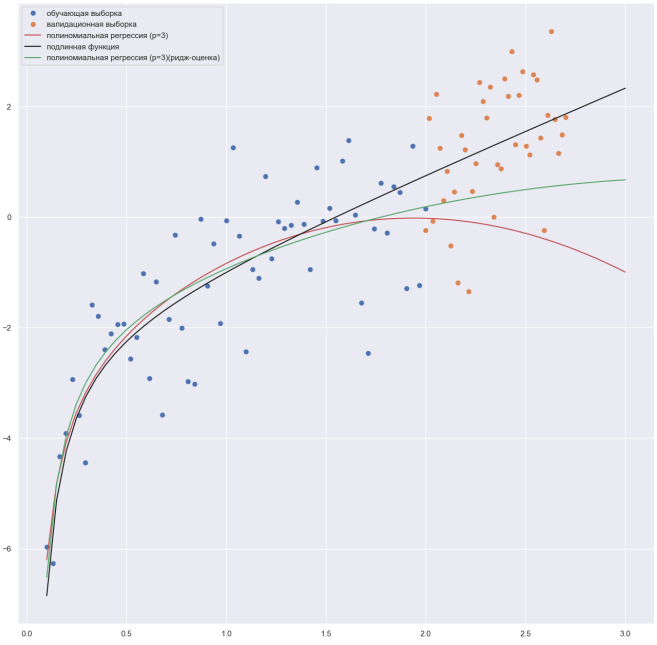
Таблица 3: Сводная таблица полиномиальных регрессий



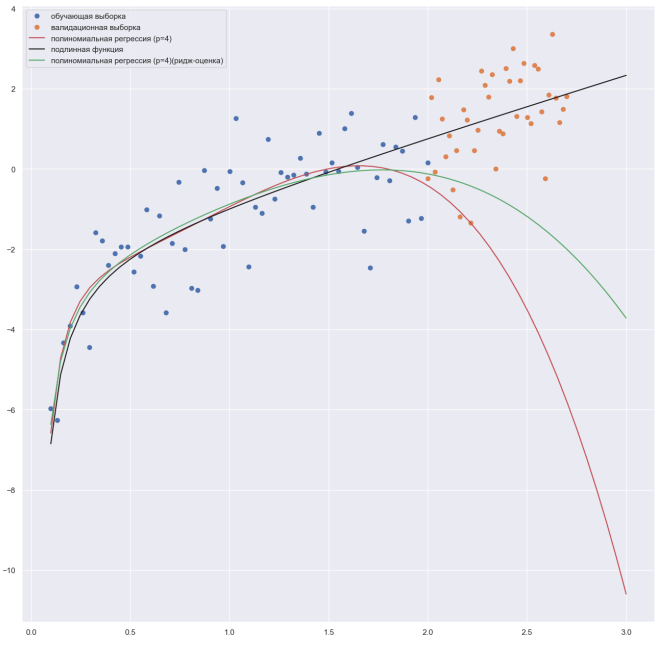

$p = 1$



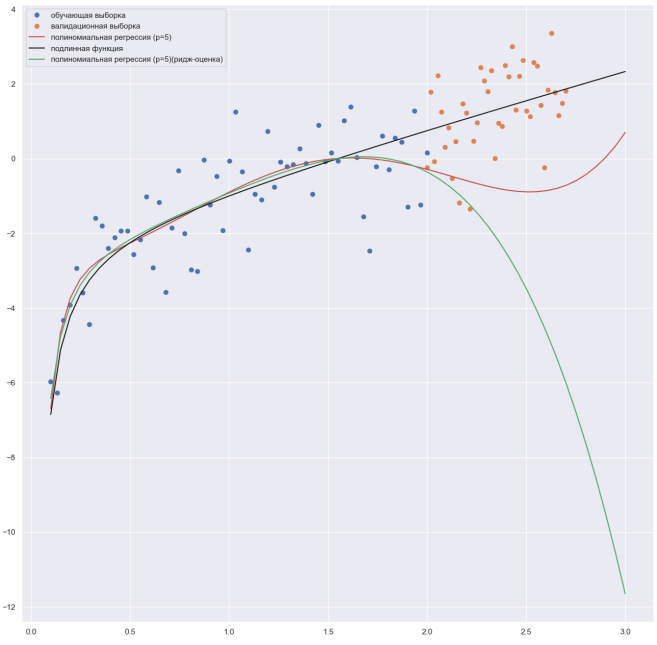

$p = 2$



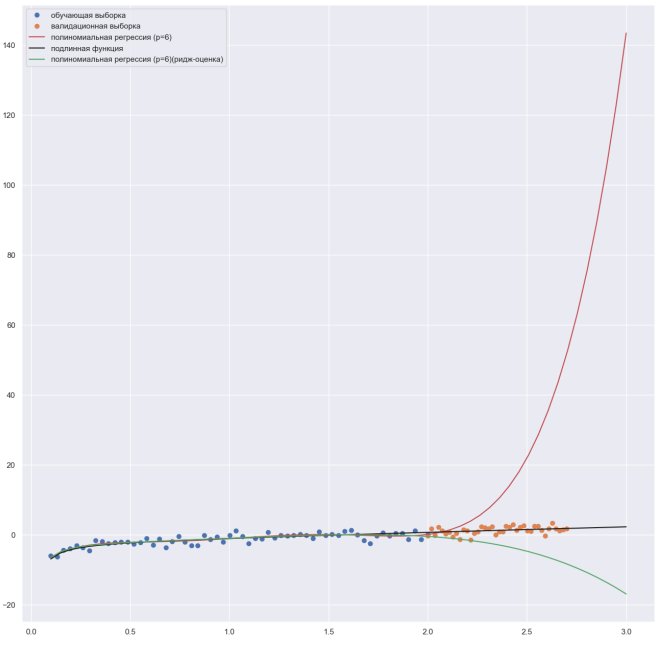

$p = 3$




$p = 4$




$p = 5$




$p = 6$



Рис. 7: 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 7 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 7 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

### 3.5 Квантильная регрессия

## 4 Выводы

## 5 Список источников