Data learning

Курс “Машинное обучение”

Лабораторная работа

Weighted Least Squares

Леонов В.В., М23-524

Вариант 1-06

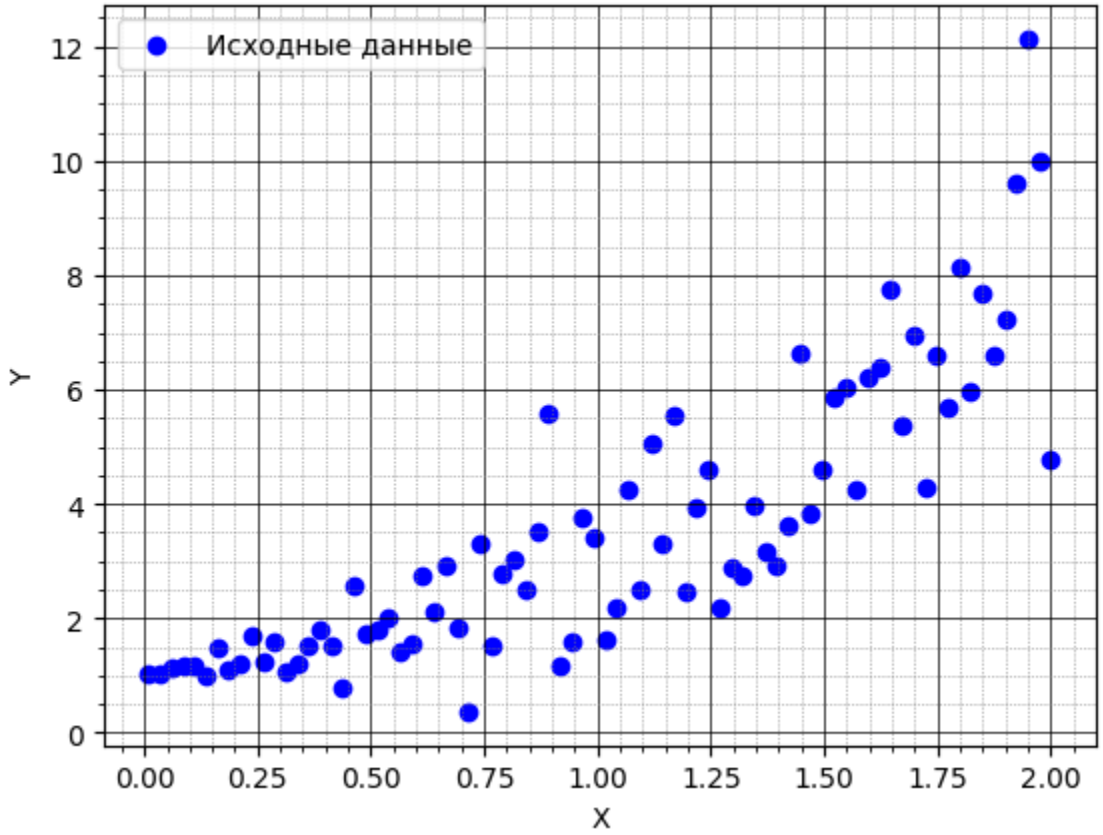
2023

Исходные данные

Исходные данные представлены в виде файла **data_v1-06.csv**, который содержит две переменные *x*, *y*

Объем выборки составляет **80 записей**.

Исходные данные



Используемые методы и формулы

Кросс-валидация – это метод оценки производительности модели машинного обучения, который помогает уменьшить влияние случайности в процессе разделения данных на обучающую и тестовую выборки. Основная идея заключается в разделении данных на несколько подмножеств и последующем обучении и тестировании модели на их разных комбинациях. Кросс-валидация позволяет более точно оценить обобщающую способность модели, уменьшает риск переобучения и обнаруживает стабильность модели на различных подмножествах данных.

Этапы:

Разбиение данных: Исходные данные разделяются на K подмножеств.

Обучение и тестирование: Модель обучается K раз, каждый раз используя K-1 фолдов в качестве обучающего набора данных и оставшийся 1 фолд в качестве тестового набора данных.

Оценка производительности: За каждую итерацию вычисляются метрики производительности модели, и в конце процесса получается усредненная оценка.

Используемые методы и формулы

Метод наименьших квадратов (OLS, Ordinary Least Squares) – это статистический метод оценки параметров линейной регрессии. Он минимизирует сумму квадратов разностей между фактическими и предсказанными значениями зависимой переменной. Формула для оценки коэффициентов линейной регрессии в случае одномерной зависимой переменной выглядит следующим образом:

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

где $\hat{\beta}$ – вектор оценок коэффициентов регрессии,

X – матрица признаков,

Y – вектор зависимых переменных.

Используемые методы и формулы

R-квадрат является мерой, используемой в регрессионном анализе для измерения объяснительной способности модели. Он предоставляет информацию о том, насколько хорошо зависимая переменная (целевая переменная) объясняется независимыми переменными (признаками) в модели регрессии. R-квадрат находится в пределах от 0 до 1, где:
$R^2 = 0$: модель не объясняет вариацию зависимой переменной,
$R^2 = 1$: модель полностью объясняет вариацию зависимой переменной.

$$R^2_{test} = 1 - \frac{\left(1 - R^2_{train}\right)(size(Y_{test}) - 1)}{size(Y_{test}) - shape(X) - 1}$$

Используемые методы и формулы

t-статистика используется для проверки гипотез относительно среднего значения в выборке. Формула для t-статистики в случае проверки гипотезы о среднем значении для одной выборки выглядит следующим образом:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

p-значение (p-value) – это вероятность получить результаты, не менее экстремальные, чем фактические результаты, при условии, что нулевая гипотеза верна. В контексте t-теста, формула для p-значения выглядит следующим образом:

$$p_{value} = P(|T| \geq |t|)$$

Используемые методы и формулы

Тест Бройша–Пагана используется для проверки гипотезы о гетероскедастичности ошибок в регрессионной модели.

$$y_t = x_t^T b + \varepsilon_t$$

$$\hat{\sigma}^2 = \frac{ESS}{n}$$

$$\frac{e_t^2}{\widehat{\sigma^2}} = \gamma_0 + z_t^T \gamma + \nu_t$$

$$\frac{e_t^2}{\widehat{\sigma^2}} = \gamma_0 + x_t^T \gamma + \nu_t$$

Используемые методы и формулы

Критерий хи-квадрат используется для проверки статистической значимости связи между двумя категориальными переменными. Этот тест сравнивает фактическое распределение частот в наблюдаемых данных с тем, которое можно было бы ожидать при условии независимости между переменными.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}},$$

где O_{ij} – наблюдаемая частота,

E_{ij} – ожидаемая частота.

Используемые методы и формулы

Метод взвешенных наименьших квадратов (WLS) является обобщением метода наименьших квадратов (OLS), который учитывает веса для каждого наблюдения. Эти веса отражают степень доверия, которую мы придаём каждому наблюдению. Формула для оценки параметров WLS выглядит следующим образом:

$$\hat{\beta} = (X^T \mathrm{W} \mathrm{X})^{-1} X^T W \mathrm{Y},$$

где $\hat{\beta}$ – вектор оценок коэффициентов регрессии,

X – матрица признаков,

W – диагональная матрица весов,

Y – вектор зависимых переменных.

После обучения остатки необходимо скорректировать с учетом весов:

$$\varepsilon'(x_i) = \varepsilon(x_i) w(x_i)$$

Результаты исследований

Построение простейшей линейной регрессионной модели методом наименьших квадратов

OLS Regression Results

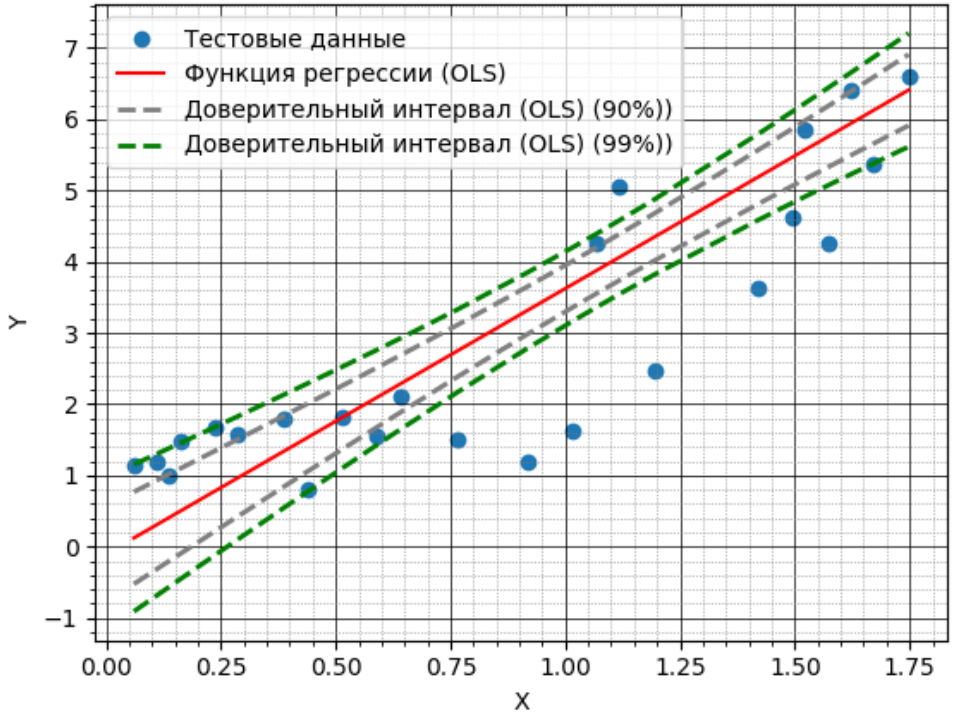
Dep. Variable:	y	R-squared:	0.699
Model:	OLS	Adj. R-squared:	0.693
Method:	Least Squares	F-statistic:	125.3
Date:	Mon, 25 Dec 2023	Prob (F-statistic):	1.08e-15
Time:	06:32:00	Log-Likelihood:	-99.062
No. Observations:	56	AIC:	202.1
Df Residuals:	54	BIC:	206.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1037	0.404	-0.257	0.798	-0.914	0.706
x1	3.7265	0.333	11.195	0.000	3.059	4.394

Omnibus:	7.193	Durbin-Watson:	1.404
Prob(Omnibus):	0.027	Jarque-Bera (JB):	7.100
Skew:	0.568	Prob(JB):	0.0287
Kurtosis:	4.324	Cond. No.	4.01

Результаты исследований

Построение простейшей линейной регрессионной модели методом OLS. Диаграмма рассеяния с рассчитанной функцией регрессии и границами доверительных интервалов интервалов.



Результаты исследований

Коэффициент детерминации на обучающей выборке (OLS):0.6989

Коэффициент детерминации на тестовой выборке (OLS):0.6702

Результаты исследований

Результаты исследований

Результаты исследований

Построение модели остатков на входную переменную

OLS Regression Results

Dep. Variable:	y	R-squared:	0.219
Model:	OLS	Adj. R-squared:	0.204
Method:	Least Squares	F-statistic:	15.12
Date:	Mon, 25 Dec 2023	Prob (F-statistic):	0.000279
Time:	06:49:12	Log-Likelihood:	-69.300
No. Observations:	56	AIC:	142.6
Df Residuals:	54	BIC:	146.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2489	0.237	1.048	0.299	-0.227	0.725
x1	0.7607	0.196	3.888	0.000	0.368	1.153

Omnibus:	15.550	Durbin-Watson:	1.330
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20.958
Skew:	1.003	Prob(JB):	2.81e-05
Kurtosis:	5.226	Cond. No.	4.01

Результаты исследований

Для остатков модели

T-статистика: 3.8882

P-значение: 0.0003

Т.к. t-статистика отлична от нуля, то модель значима.

Т.к. p-значение мало, то соответствующий коэффициент значим.

В результате качественного анализа остатков можно сделать вывод, что остатки гетероскедастичны.

Результаты исследований

Breusch-Pagan тест на гетероскедастичность:

LM статистика: 9.2124

P-значение: 0.0024

F-статистика: 10.6325

P-значение для F-статистики: 0.0019

Т.к. P-значение < 0.05, то можно считать, что гетероскедастичность присутствует в данных.

Результаты исследований

Построение модели методом WLS с величинами, обратными модельным значениям функции регрессии

WLS Regression Results

Dep. Variable:	y	R-squared:	0.735
Model:	WLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	149.6
Date:	Mon, 25 Dec 2023	Prob (F-statistic):	3.43e-17
Time:	03:04:06	Log-Likelihood:	-91.055
No. Observations:	56	AIC:	186.1
Df Residuals:	54	BIC:	190.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3627	0.256	1.416	0.163	-0.151	0.876
x1	3.2891	0.269	12.231	0.000	2.750	3.828

Omnibus:	4.623	Durbin-Watson:	1.431
Prob(Omnibus):	0.099	Jarque-Bera (JB):	3.596
Skew:	0.505	Prob(JB):	0.166
Kurtosis:	3.721	Cond. No.	2.95

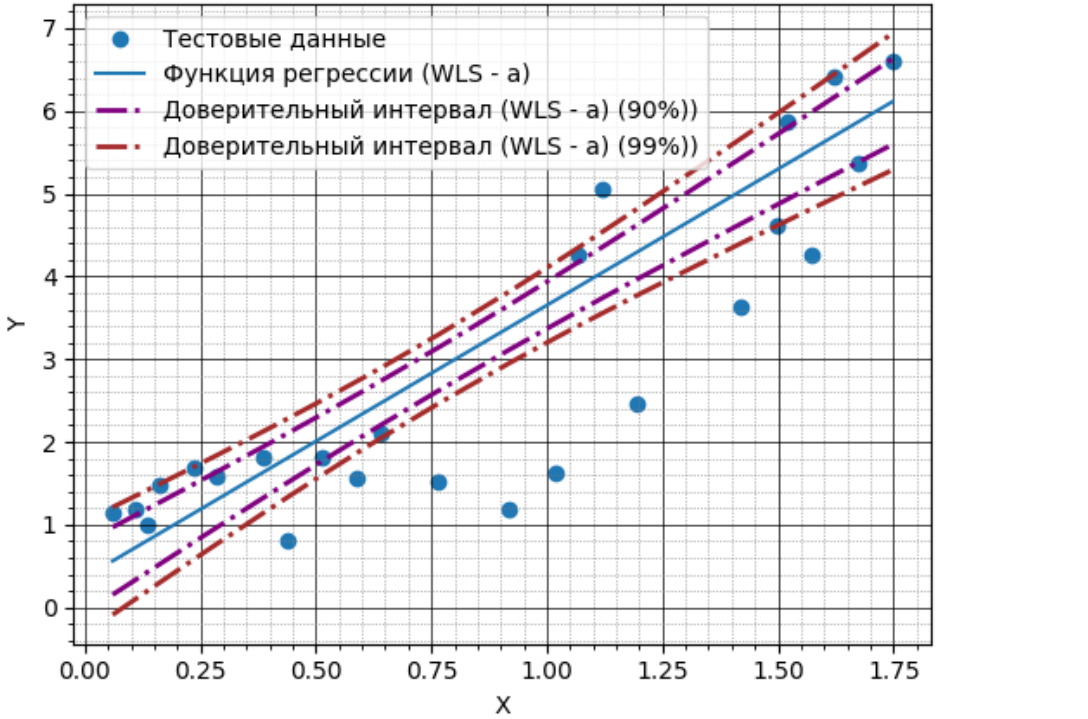
Результаты исследований

Коэффициент детерминации на обучающей выборке (WLS - a):0.7348
Коэффициент детерминации на тестовой выборке (WLS - a):0.7095

Результаты исследований

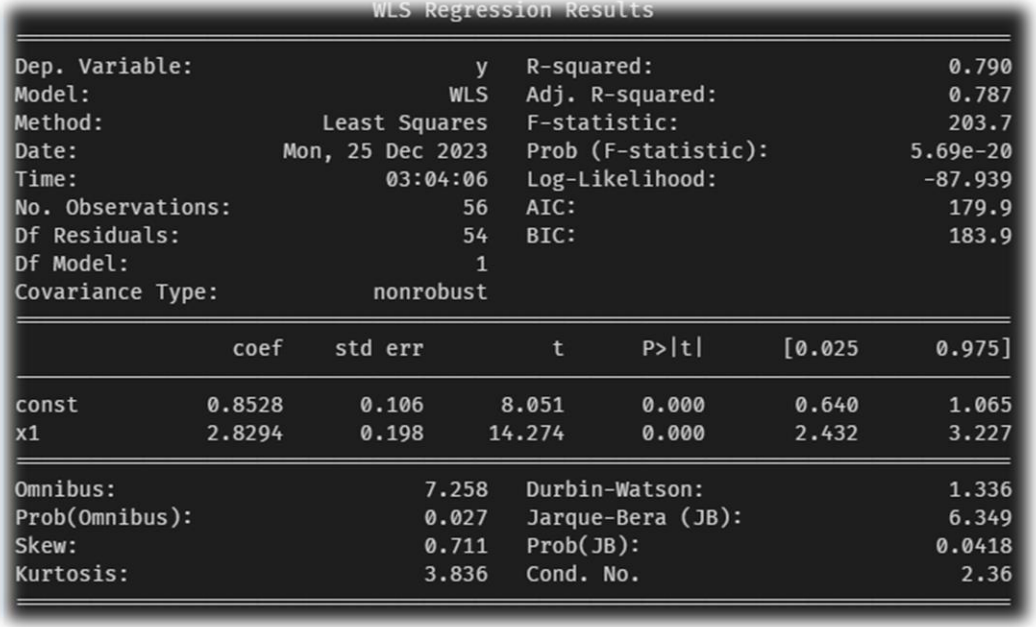
Результаты исследований

Построение модели методом WLS с величинами, обратными модельным значениям функции регрессии. Диаграмма рассеяния с рассчитанной функцией регрессии и границами доверительных интервалов интервалов.



Результаты исследований

Построение модели методом WLS с величинами, равными $1/x$



WLS Regression Results

Dep. Variable:	y	R-squared:	0.790
Model:	WLS	Adj. R-squared:	0.787
Method:	Least Squares	F-statistic:	203.7
Date:	Mon, 25 Dec 2023	Prob (F-statistic):	5.69e-20
Time:	03:04:06	Log-Likelihood:	-87.939
No. Observations:	56	AIC:	179.9
Df Residuals:	54	BIC:	183.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8528	0.106	8.051	0.000	0.640	1.065
x1	2.8294	0.198	14.274	0.000	2.432	3.227

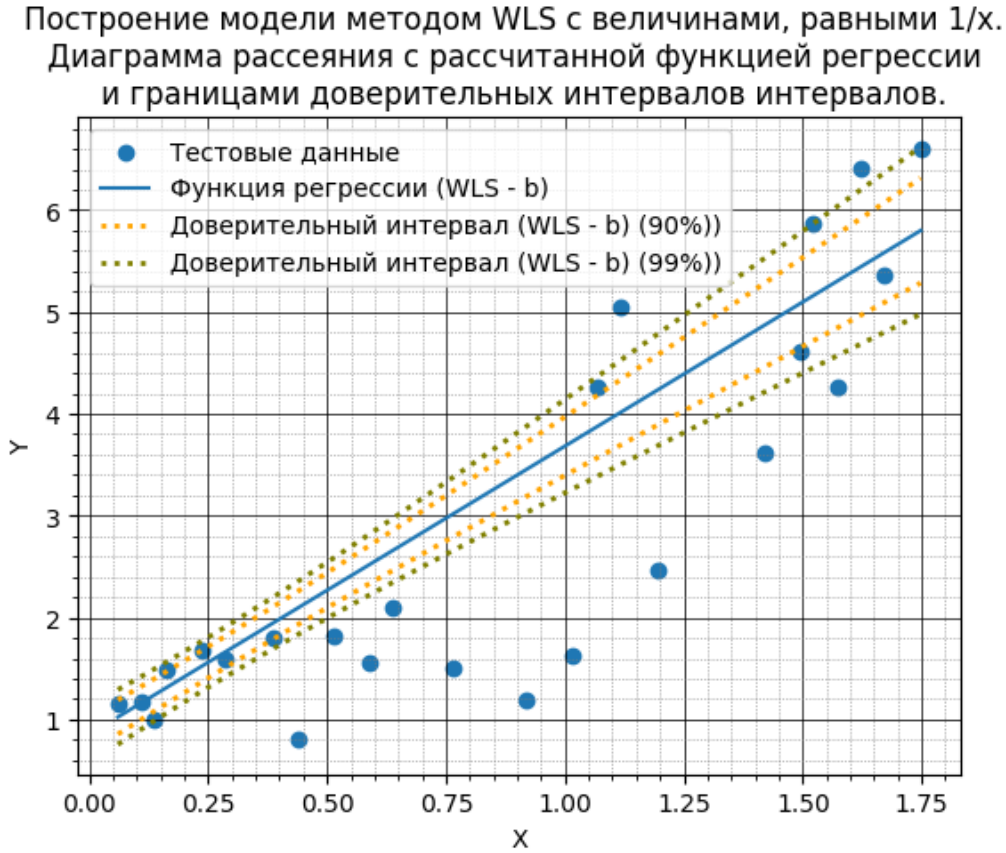
Omnibus:	7.258	Durbin-Watson:	1.336
Prob(Omnibus):	0.027	Jarque-Bera (JB):	6.349
Skew:	0.711	Prob(JB):	0.0418
Kurtosis:	3.836	Cond. No.	2.36

Результаты исследований

Коэффициент детерминации на обучающей выборке (WLS - b):0.7905
Коэффициент детерминации на тестовой выборке (WLS - b):0.7705

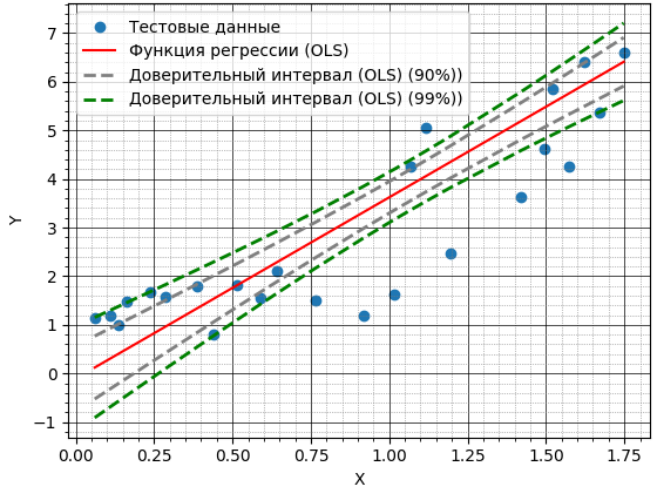
Результаты исследований

Результаты исследований

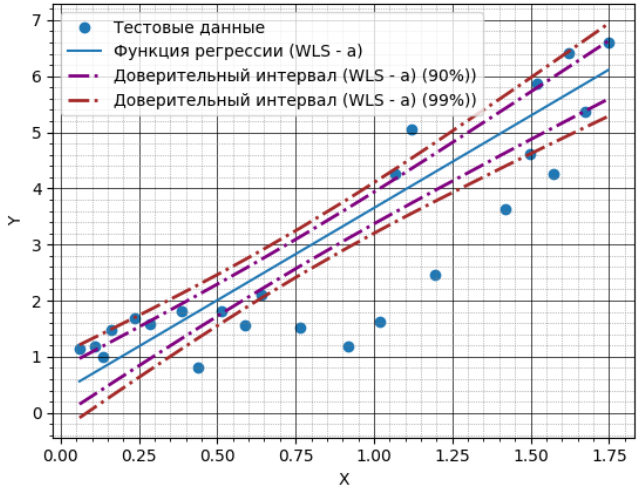


Результаты исследований

Построение простейшей линейной регрессионной модели методом OLS. Диаграмма рассеяния с рассчитанной функцией регрессии и границами доверительных интервалов интервалов.

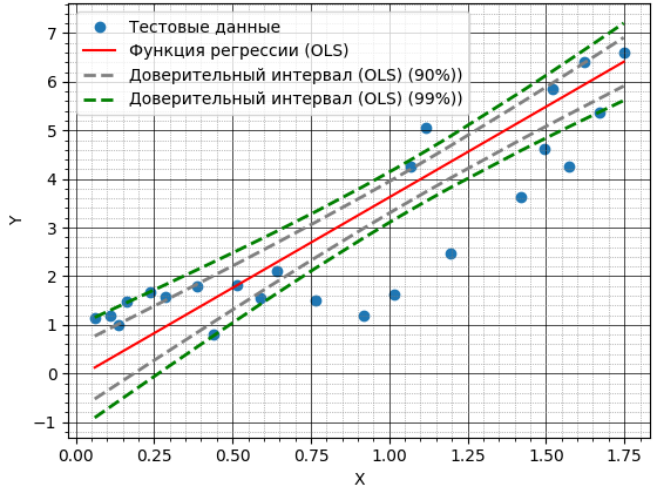


Построение модели методом WLS с величинами, обратными модельным значениям функции регрессии. Диаграмма рассеяния с рассчитанной функцией регрессии и границами доверительных интервалов интервалов.

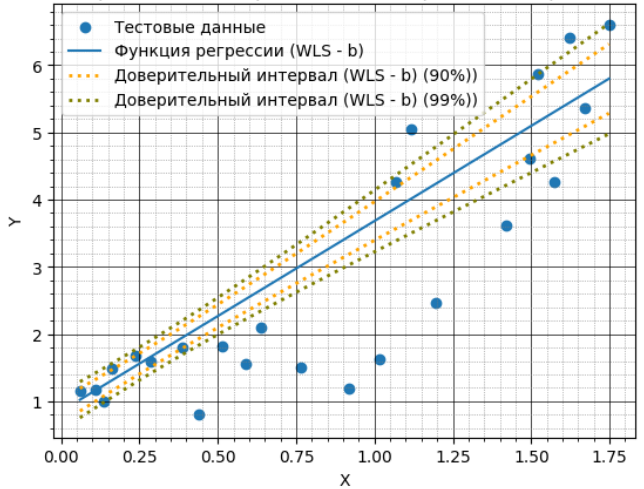


Результаты исследований

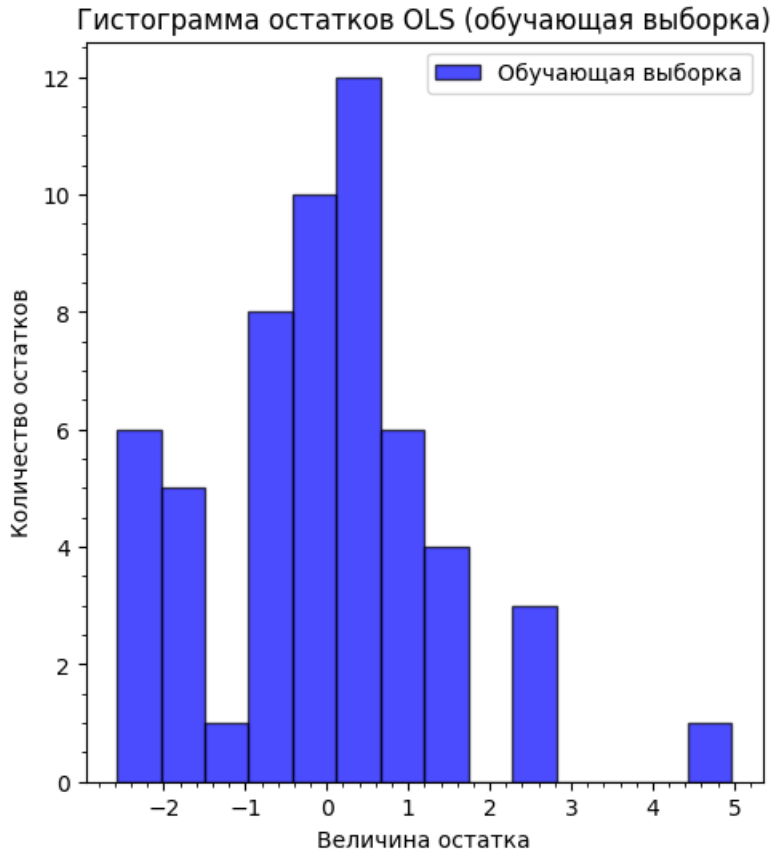
Построение простейшей линейной регрессионной модели методом OLS. Диаграмма рассеяния с рассчитанной функцией регрессии и границами доверительных интервалов интервалов.

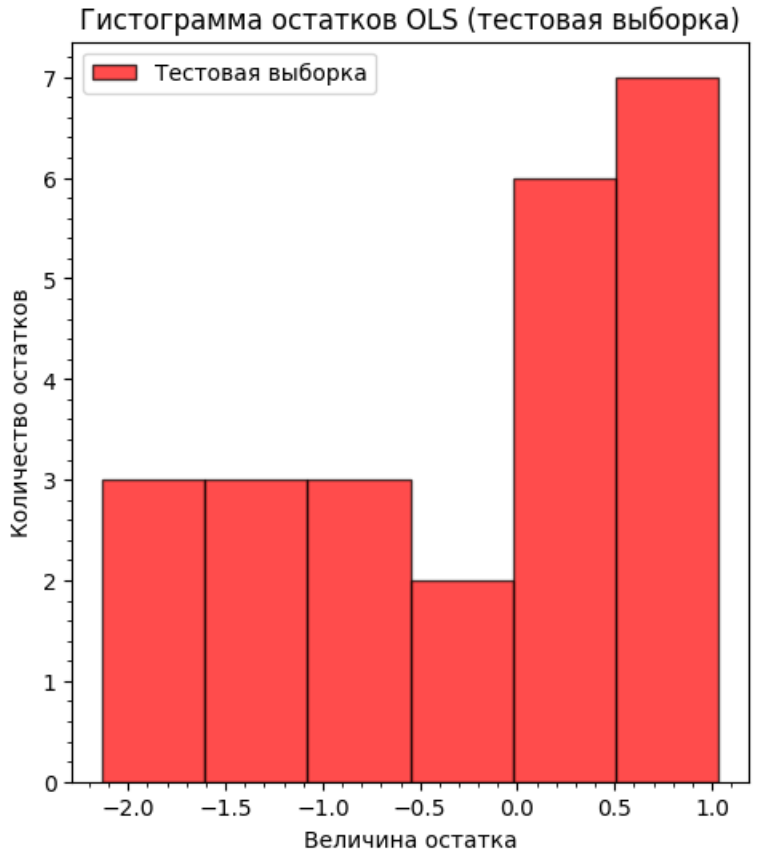


Построение модели методом WLS с величинами, равными 1/x. Диаграмма рассеяния с рассчитанной функцией регрессии и границами доверительных интервалов интервалов.

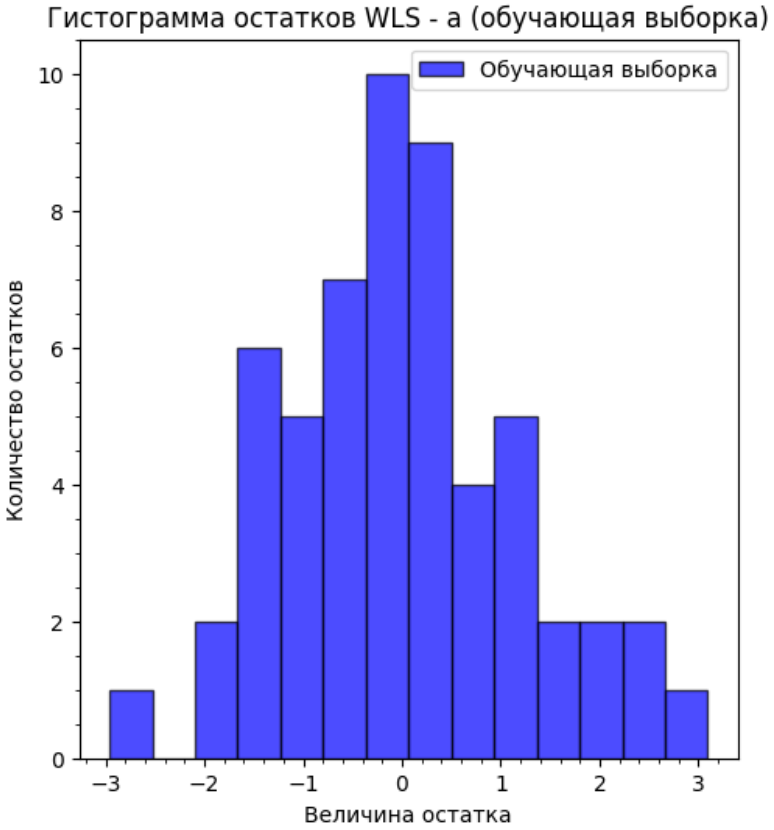


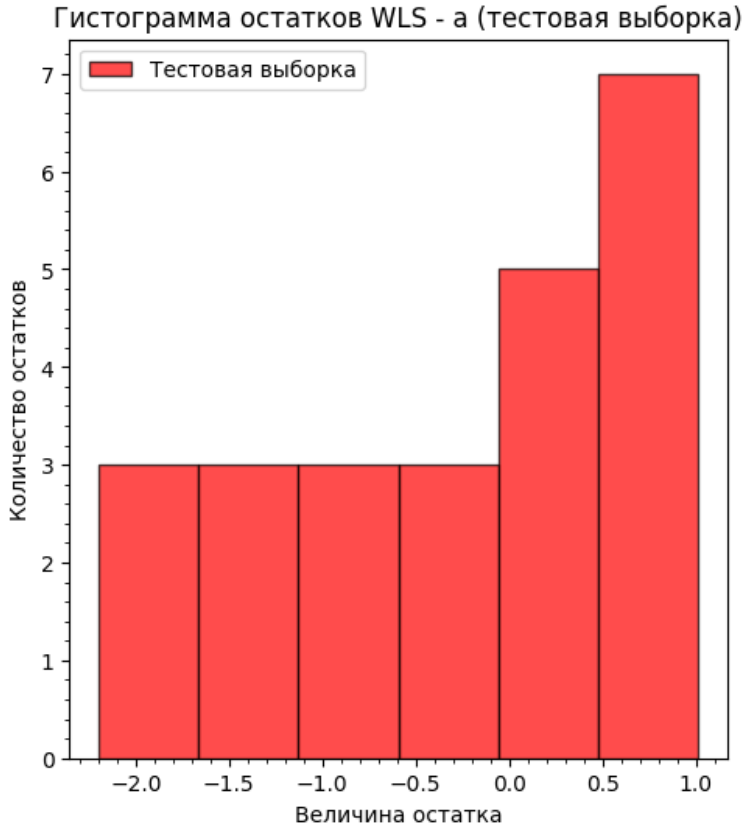
Результаты исследований



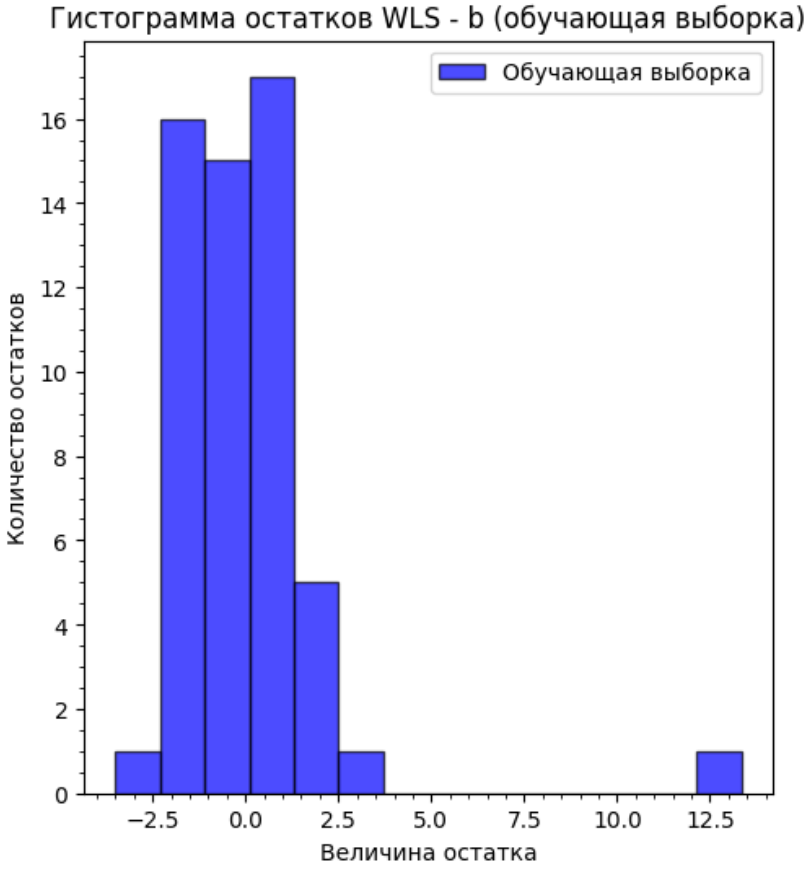


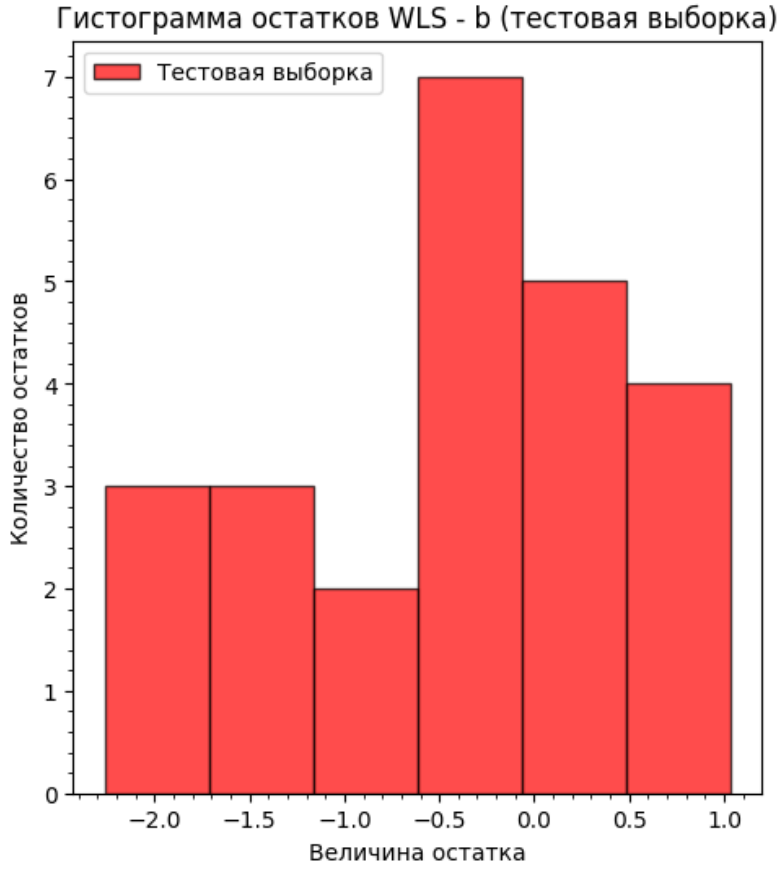
Результаты исследований





Результаты исследований





Результаты исследований

Тест на нормальность остатков

OLS (обучающая выборка):

Статистика критерия: 56.0000

P-значение: 0.4371

Тест на нормальность остатков

OLS (тестовая выборка):

Статистика критерия: 25.2445

P-значение: 0.3378

Т.к. p-значение > 0.05, то распределение нормальное

Результаты исследований

Тест на нормальность остатков

WLS - a (обучающая выборка):

Статистика критерия: 56.0000

P-значение: 0.4371

Тест на нормальность остатков

WLS - a (тестовая выборка):

Статистика критерия: 26.8178

P-значение: 0.2638

Т.к. p-значение > 0.05, то распределение нормальное

Результаты исследований

Тест на нормальность остатков

WLS - b (обучающая выборка):

Статистика критерия: 56.0000

P-значение: 0.4371

Тест на нормальность остатков

WLS - b (тестовая выборка):

Статистика критерия: 28.9454

P-значение: 0.1821

Т.к. p-значение > 0.05, то распределение нормальное

Выводы

Метод наименьших квадратов (OLS) и метод взвешенных наименьших квадратов (WLS) представляют собой два подхода к оценке параметров в регрессионном анализе, но они отличаются в обработке гетероскедастичности.

Среди преимуществ WLS:
- WLS более устойчив к наличию гетероскедастичности, так как он позволяет учесть различия в дисперсии ошибок.
- Если гетероскедастичность присутствует, и веса выбраны правильно, WLS может обеспечить более эффективные оценки параметров.