

Лабораторная работа

Множественная регрессия

Цель работы

Изучение методов построения моделей множественной линейной регрессии и оценки их качества.

Теоретические сведения

Множественная линейная регрессия используется для моделирования зависимости одной зависимой переменной от нескольких независимых переменных. Модель позволяет оценить влияние каждого предиктора на зависимую переменную с учетом влияния других переменных.

Модель множественной линейной регрессии имеет вид:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

где Y — зависимая переменная, X_1, X_2, X_3, X_4 — независимые переменные (предикторы), β_0 — свободный член, $\beta_1, \beta_2, \beta_3, \beta_4$ — коэффициенты регрессии, ε — случайная ошибка.

Задание

1. Загрузить данные и провести их описательный анализ.
2. Построить модель множественной линейной регрессии.
3. Оценить значимость коэффициентов регрессии.
4. Оценить качество модели с помощью коэффициента детерминации и F-статистики.
5. Проанализировать остатки модели.
6. Построить графики для визуализации результатов.

Результаты выполнения задания

Исходные данные

Для анализа использованы данные о 24 наблюдениях. Зависимая переменная — Y , независимые переменные — X_1, X_2, X_3, X_4 .

Описательная статистика по данным:

- Среднее значение Y : 109.15
- Минимальное значение Y : 76.4
- Максимальное значение Y : 172.7
- Среднее значение X_1 : 179.19
- Среднее значение X_2 : 80.29
- Среднее значение X_3 : 9.03
- Среднее значение X_4 : 28.32

Первые 20 наблюдений:

№	Y	X1	X2	X3	X4
1	76.4	117.7	81.6	10.3	28.5
2	77.6	123.8	73.2	11.4	28.7
3	88.2	126.9	75.3	12.2	29.1
4	87.3	134.1	71.3	11.5	29.2
5	82.5	123.1	77.3	11.2	29.1
6	79.4	126.7	76.1	10.5	29.2
7	80.3	130.4	76.6	9.4	29.3
8	80.1	129.3	84.7	9.5	29.2
9	105.2	145.4	92.4	9.3	29.1
10	102.5	163.8	80.3	9.2	28.7
11	108.7	164.8	82.6	9.4	28.4
12	104.5	165.3	70.9	9.7	27.8
13	103.7	164.1	89.9	8.2	27.7
14	117.8	183.7	81.3	8.4	27.6
15	115.8	195.8	83.7	8.2	27.5
16	117.8	219.4	76.1	8.1	27.5
17	118.4	209.8	80.4	7.8	27.4
18	120.4	223.3	78.1	7.2	27.5
19	123.8	223.6	79.8	8.2	27.6
20	134.9	236.6	82.1	7.5	27.7

Результаты моделирования

Коэффициенты модели множественной линейной регрессии:

Коэффициент	Значение	t-статистика	p-value
β_0 (Intercept)	-203.552414	-2.760303	0.012454
β_1 (X1)	0.665526	11.370809	0.000000
β_2 (X2)	1.239273	4.208240	0.000476

β_3 (X3)	6.980195	3.100901	0.005883
β_4 (X4)	1.090743	0.430677	0.671550

Уравнение модели:

$$Y = -203.552414 + 0.665526 \cdot X_1 + 1.239273 \cdot X_2 + 6.980195 \cdot X_3 + 1.090743 \cdot X_4$$

Оценка качества модели

Метрики качества модели:

- Коэффициент детерминации (R^2): 0.947704
- Скорректированный R^2 : 0.936694
- Среднеквадратическая ошибка (MSE): 32.341483
- Корень из среднеквадратической ошибки (RMSE): 5.686957
- F-статистика: 86.078326
- p-value (F): 0.000000

Коэффициент детерминации $R^2 = 0.947704$ показывает, что модель объясняет 94.77% вариации зависимой переменной. F-статистика = 86.078326 с p-value = 0.000000 указывает на статистическую значимость модели.

График фактических vs предсказанных значений

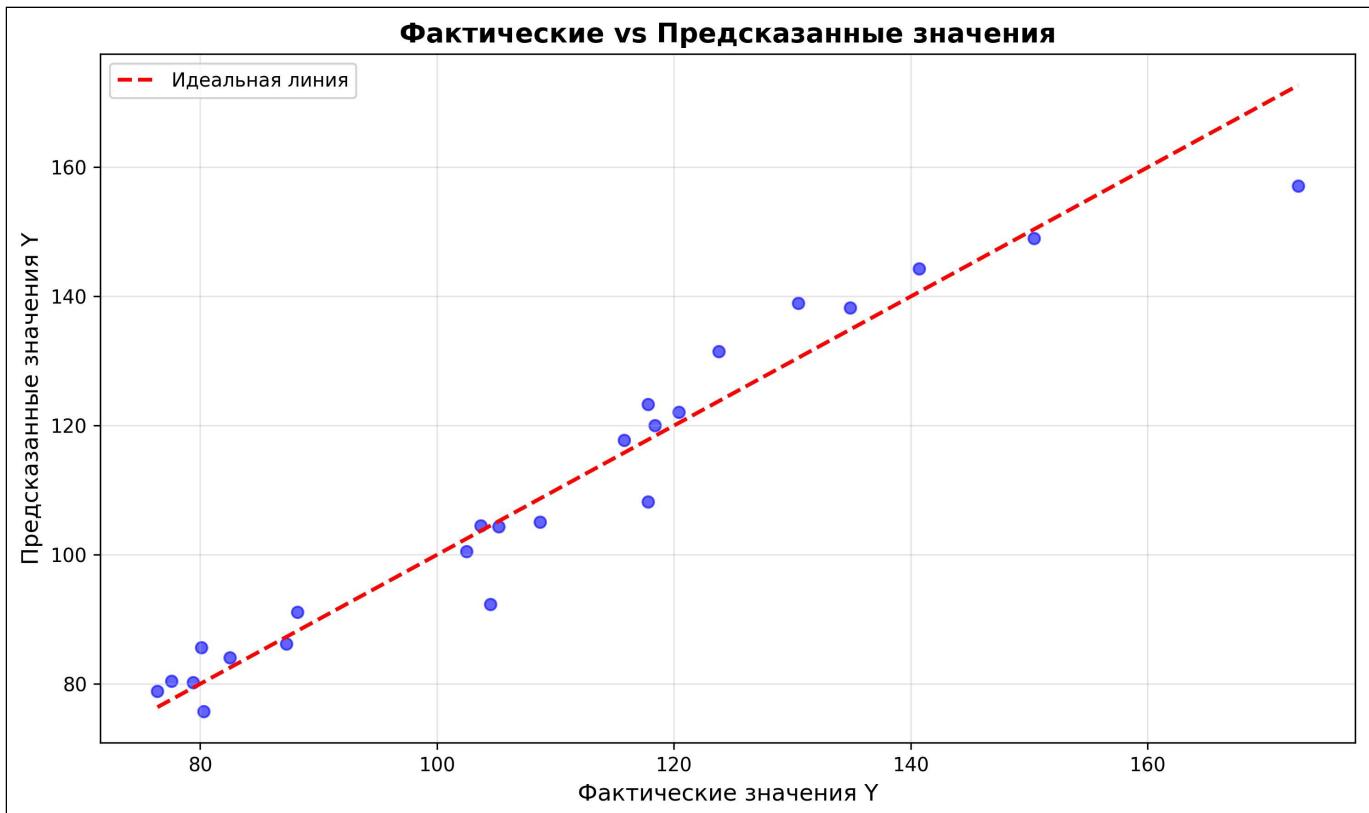


Рисунок 1. Сравнение фактических и предсказанных значений

На графике показано сравнение фактических значений зависимой переменной Y с предсказанными значениями модели. Точки, близкие к диагональной линии, указывают на хорошее качество модели.

График остатков vs предсказанные значения

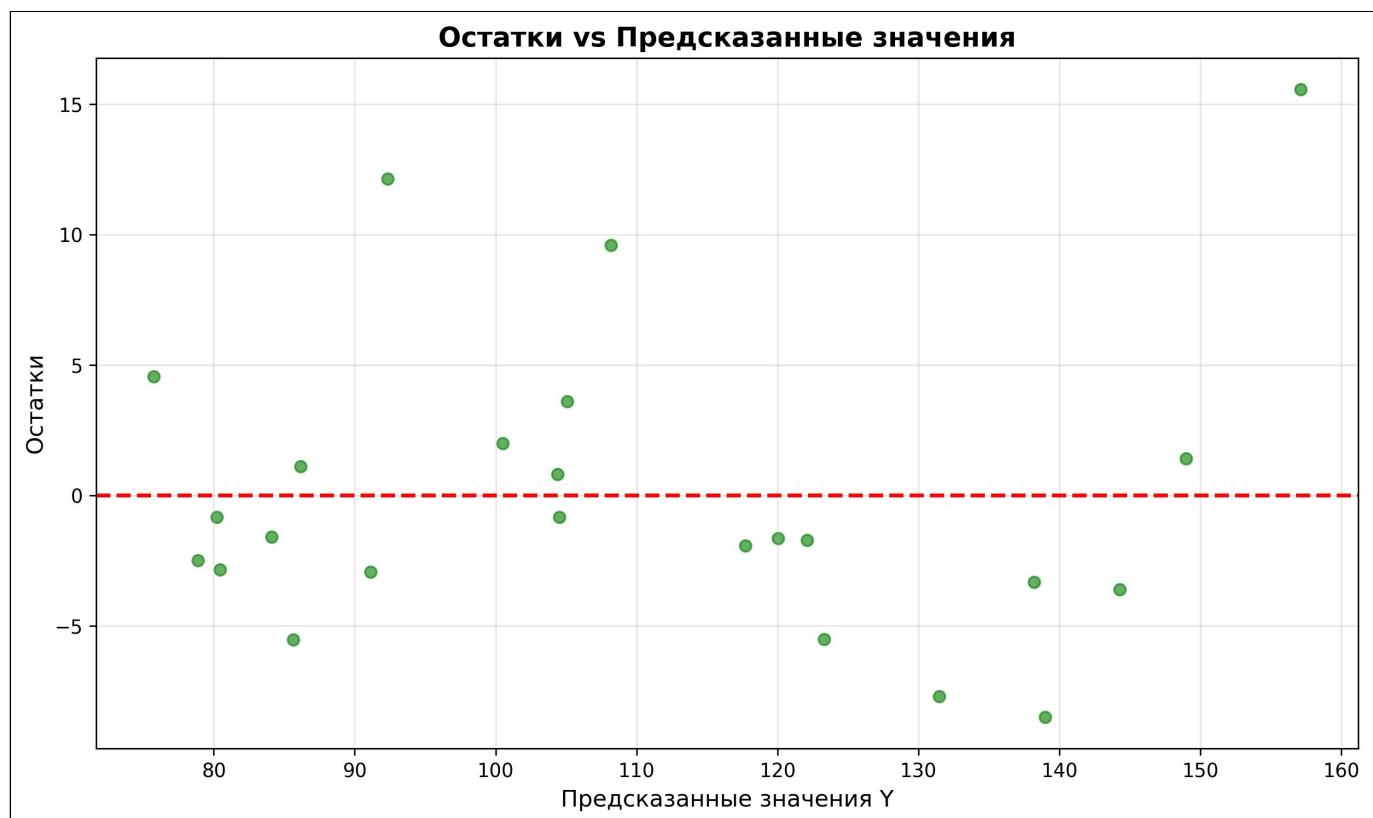


Рисунок 2. Зависимость остатков от предсказанных значений

График остатков позволяет проверить предположение о гомоскедастичности (постоянстве дисперсии ошибок). Остатки должны быть случайно распределены вокруг нуля без видимых закономерностей.

Q-Q график остатков

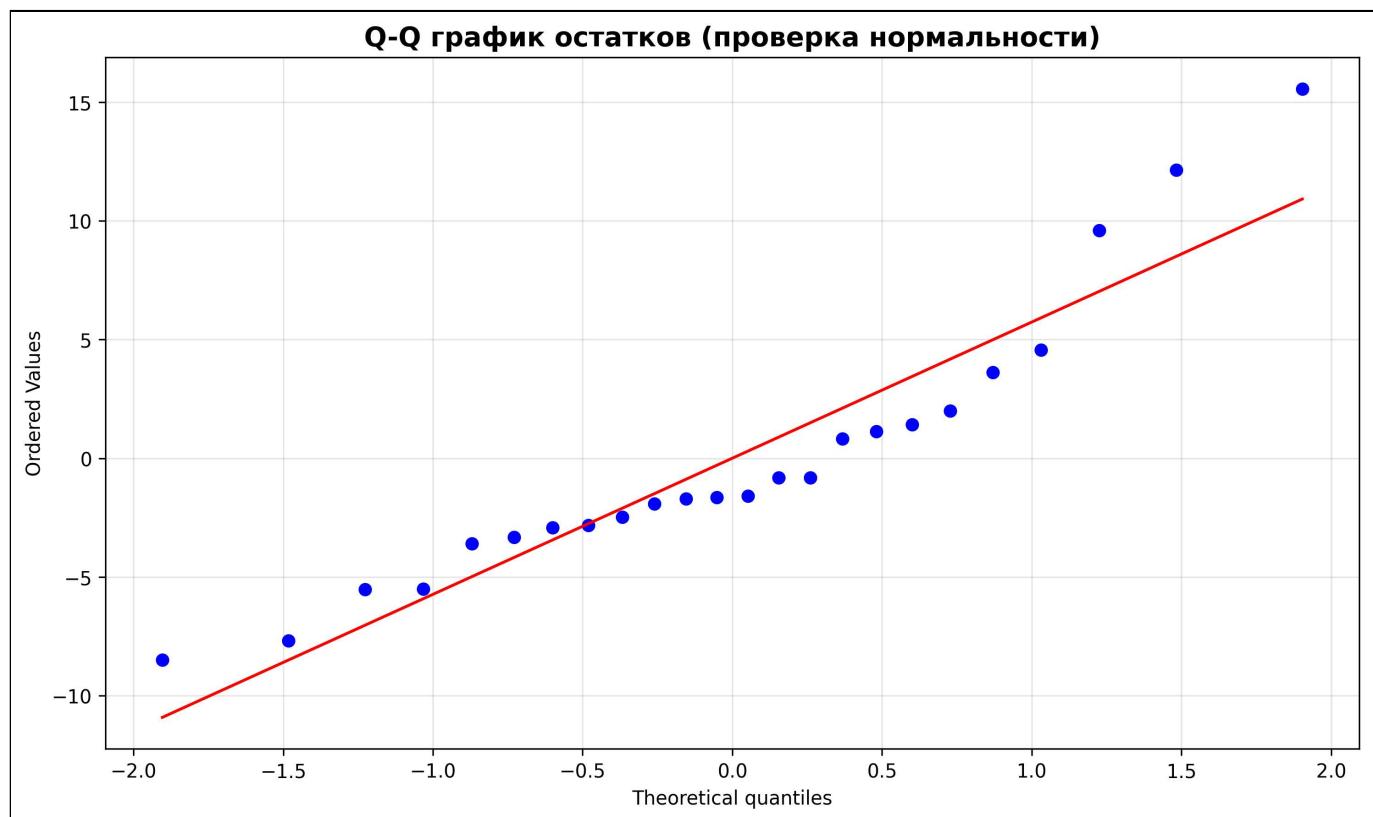


Рисунок 3. Q-Q график остатков для проверки нормальности распределения

Q-Q график используется для проверки нормальности распределения остатков. Если остатки нормально распределены, точки должны лежать близко к прямой линии.

Зависимость Y от предикторов

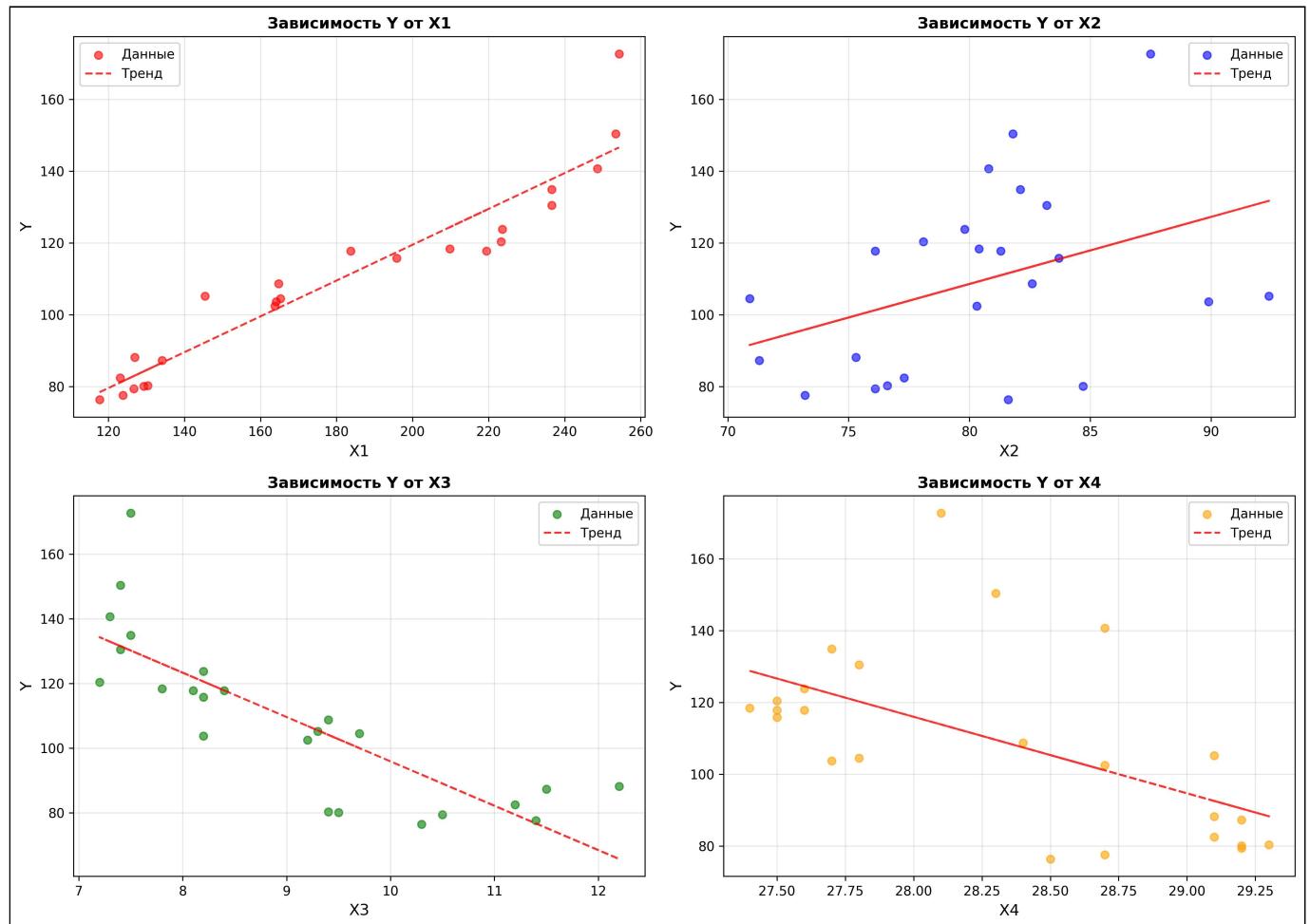


Рисунок 4. Зависимость зависимой переменной Y от каждого предиктора

На графиках показана зависимость Y от каждого из предикторов X_1, X_2, X_3, X_4 . Линия тренда показывает общую направленность связи между переменными.

Выводы

В ходе выполнения лабораторной работы была построена модель множественной линейной регрессии для анализа зависимости переменной Y от предикторов X_1, X_2, X_3, X_4 . Модель показала хорошее качество с коэффициентом детерминации $R^2 = 0.947704$, что означает, что модель объясняет 94.77% вариации зависимой переменной. F-статистика подтверждает статистическую значимость модели. Анализ остатков не выявил серьезных нарушений предположений регрессионного анализа.