Урок 12. Регрессия (часть 1)

Хакимов Р.И. + ChatGPT

Простая линейная регрессия

Простая линейная регрессия — это метод статистического анализа, который используется для моделирования и анализа линейной зависимости между двумя переменными: одной независимой (или объясняющей) переменной и одной зависимой (или откликовой) переменной.
Цель простого линейного регрессийного анализа — построить линейную модель, которая наилучшим образом описывает связь между этими переменными.

Простая линейная регрессия

Модель простой линейной регрессии

Модель простой линейной регрессии имеет следующий вид:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

где:

Y — зависимая переменная (отклик).

X — независимая переменная (объясняющая).

β_0 — свободный член (пересечение линии с осью Y).

β_1 — коэффициент наклона (показывает, насколько изменяется Y при изменении X на единицу).

ϵ — случайная ошибка (различие между предсказанным и фактическим значением Y).

Простая линейная регрессия

Процесс построения модели

1. Сбор данных. Соберите данные для зависимой и независимой переменных.

2. Построение диаграммы рассеяния. Постройте диаграмму рассеяния для визуализации взаимосвязи между переменными. Это поможет определить, насколько хорошо данные могут быть представлены линейной моделью.

3. Расчет коэффициентов регрессии. Используйте метод наименьших квадратов для оценки коэффициентов β_0 и β_1. Метод наименьших квадратов минимизирует сумму квадратов отклонений предсказанных значений от фактических данных.

4. Оценка модели. Проверьте качество модели с помощью статистических метрик, таких как коэффициент детерминации R^2, стандартная ошибка регрессии, и тестирование значимости коэффициентов.

Простая линейная регрессия

Формулы для коэффициентов

1. Коэффициент наклона (β_1):

$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

2. Свободный член (β_0):

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

где $\bar{X}$ и $\bar{Y}$ — средние значения переменных X и Y соответственно.

Простая линейная регрессия

Интерпретация результатов

Коэффициент наклона (β_1) показывает, на сколько единиц изменяется зависимая переменная Y при увеличении независимой переменной X на одну единицу. Если β_1 положителен, связь между переменными положительная; если отрицателен — связь отрицательная.

2. Свободный член (β_0) - это значение зависимой переменной Y, когда независимая переменная X равна нулю. Это может быть интерпретировано как начальная точка на оси Y.

3. Коэффициент детерминации (R^2) показывает, какая доля вариации зависимой переменной объясняется моделью. Значение R^2 варьируется от 0 до 1. Чем ближе R^2 к 1, тем лучше модель объясняет вариацию данных.

Простая линейная регрессия

Коэффициент детерминации R^2 (R-squared) — это статистическая метрика, которая показывает, насколько хорошо модель регрессии объясняет изменчивость зависимой переменной. Значение R^2 варьируется от 0 до 1. Чем ближе R^2 к 1, тем лучше модель описывает данные.

Формула для вычисления R^2 следующая:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

где:

- SS_{res} — сумма квадратов остатков (residual sum of squares), которая вычисляется как:

$$SS_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

где y_i — наблюдаемые значения зависимой переменной, $\hat{y}_i$ — предсказанные моделью значения.

Простая линейная регрессия

- SS_{tot} — полная сумма квадратов (total sum of squares), которая выражается как:

$$SS_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

где $\bar{y}$ — среднее значение наблюдаемых данных.

Таким образом, коэффициент детерминации R^2 показывает долю общей вариации зависимой переменной, которая объясняется моделью. Если $R^2 = 1$, то модель точно предсказывает все наблюдаемые значения. Если $R^2 = 0$, модель не объясняет изменчивость данных вообще.

Простая линейная регрессия

- SS_{tot} — полная сумма квадратов (total sum of squares), которая выражается как:

$$SS_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

где $\bar{y}$ — среднее значение наблюдаемых данных.

Таким образом, коэффициент детерминации R^2 показывает долю общей вариации зависимой переменной, которая объясняется моделью. Если $R^2 = 1$, то модель точно предсказывает все наблюдаемые значения. Если $R^2 = 0$, модель не объясняет изменчивость данных вообще.

Простая линейная регрессия. Пример 1

Рассмотрим пример расчета коэффициентов простой линейной регрессии. Пусть у нас есть следующие данные о зависимости переменной y от переменной x:

x	y
1	2
2	3
3	4
4	5
5	6

Простая линейная регрессия. Пример 1

Мы хотим найти коэффициенты простой линейной регрессии вида:

$$y = \beta_0 + \beta_1 x$$

Шаг 1: Найдем средние значения $\bar{x}$ и $\bar{y}$

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{y} = \frac{2+3+5+4+6}{5} = 4$$

Простая линейная регрессия. Пример 1

Шаг 2: Найдем коэффициент наклона β_1

Коэффициент наклона β_1 вычисляется по формуле:

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Вычислим сумму $\sum(x_i - \bar{x})(y_i - \bar{y})$:

$$(1-3)(2-4) + (2-3)(3-4) + (3-3)(5-4) + (4-3)(4-4) + (5-3)(6-4)$$

$$= (-2)(-2) + (-1)(-1) + (0)(1) + (1)(0) + (2)(2)$$

$$= 4 + 1 + 0 + 0 + 4 = 9$$

Простая линейная регрессия. Пример 1

Вычислим сумму $\sum(x_i - \bar{x})^2$:

$$\begin{aligned}(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 \\ = (-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2 \\ = 4 + 1 + 0 + 1 + 4 = 10\end{aligned}$$

Теперь можно найти β_1:

$$\beta_1 = \frac{9}{10} = 0.9$$

Простая линейная регрессия. Пример 1

Шаг 3: Найдем свободный коэффициент β_0
Свободный коэффициент β_0 вычисляется по формуле:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Подставим значения:

$$\beta_0 = 4 - 0.9 \times 3 = 4 - 2.7 = 1.3$$

Итоговая модель
Итак, уравнение регрессии имеет вид:

$$y = 1.3 + 0.9x$$

Теперь можно использовать это уравнение для предсказания значений y на основе значений x.

Простая линейная регрессия. Пример 2

Рассмотрим пример расчета коэффициентов простой линейной регрессии. Пусть у нас есть небольшие данные о зависимости роста человека y от его веса x:

Вес (x), кг	Рост (y), см
60	155
65	160
70	168
75	175
80	180

Задача заключается в нахождении коэффициентов β_0 (свободный член) и β_1 (наклон прямой) для уравнения линейной регрессии:

$$y = \beta_0 + \beta_1 x$$

Простая линейная регрессия. Пример 2

Формулы для расчета коэффициентов

1. Коэффициент наклона β_1:

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

2. Свободный член β_0:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

где:

- $\bar{x}$ — среднее значение x,
- $\bar{y}$ — среднее значение y,
- x_i и y_i — значения каждой пары данных.

Простая линейная регрессия. Пример 2

Шаг 1. Найдем средние значения $\bar{x}$ и $\bar{y}$:

$$\bar{x} = \frac{60 + 65 + 70 + 75 + 80}{5} = 70$$

$$\bar{y} = \frac{155 + 160 + 168 + 175 + 180}{5} = 167.6$$

Простая линейная регрессия. Пример 2

Шаг 2. Найдем β_1:

Найдём числитель и знаменатель для формулы β_1:

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = (60-70)(155-167.6) + (65-70)(160-167.6) +$$

$$+(70-70)(168-167.6) + (75-70)(175-167.6) + (80-70)(180-167.6) =$$

$$= (-10)(-12.6) + (-5)(-7.6) + (0)(0.4) + (5)(7.4) + (10)(12.4) =$$

$$= 126 + 38 + 0 + 37 + 124 = 325$$

$$\sum(x_i - \bar{x})^2 = (60-70)^2 + (65-70)^2 + (70-70)^2 + (75-70)^2 + (80-70)^2$$

$$= (-10)^2 + (-5)^2 + (0)^2 + (5)^2 + (10)^2 = 100 + 25 + 0 + 25 + 100 = 250$$

Теперь найдём β_1:

$$\beta_1 = \frac{325}{250} = 1.3$$

Простая линейная регрессия. Пример 2

Шаг 3. Найдем β_0:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 167.6 - 1.3 \times 70 = 167.6 - 91 = 76.6$$

Шаг 4. Итоговое уравнение регрессии:
Подставив коэффициенты, получаем уравнение линейной регрессии:

$$y = 76.6 + 1.3x$$

Это уравнение позволяет предсказывать рост человека на основании его веса. Например, если вес человека $x = 68$ кг, предсказанный рост будет:

$$y = 76.6 + 1.3 \times 68 = 76.6 + 88.4 = 165 \text{ см}$$

Множественная линейная регрессия

Множественная линейная регрессия — это метод статистического анализа, который используется для моделирования зависимости одной зависимой переменной (или отклика) от нескольких независимых переменных (или предикторов). Модель множественной линейной регрессии может быть выражена следующим уравнением:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

где:

Y — зависимая переменная,

$X_1, X_2, ..., X_n$ — независимые переменные (факторы),

β_0 — свободный член (константа),

$\beta_1, \beta_2, ..., \beta_n$ — коэффициенты регрессии, показывающие вклад каждой независимой переменной в предсказание зависимой переменной,

ε — ошибка модели (остаток), отражающая разницу между фактическими и предсказанными значениями Y.

Множественная линейная регрессия

Основные предпосылки множественной линейной регрессии:
1. Линейность: связь между зависимой и независимыми переменными должна быть линейной.
2. Независимость: остатки (ошибки модели) должны быть независимы.
3. Нормальность остатков: ошибки модели должны быть нормально распределены.
4. Гомоскедастичность: дисперсия остатков должна быть одинаковой для всех значений независимых переменных.
5. Отсутствие мультиколлинеарности: независимые переменные не должны сильно коррелировать между собой.

Множественная линейная регрессия позволяет оценить, насколько сильно каждый из факторов влияет на зависимую переменную и как изменяется прогнозируемая переменная при изменении независимых переменных.

Множественная линейная регрессия

Коэффициенты множественной линейной регрессии можно рассчитать с помощью метода наименьших квадратов. Формула для расчета коэффициентов регрессии в множественной линейной регрессии (векторные формы) выглядит следующим образом:

$$\beta = (X^T X)^{-1} X^T y$$

где:
- β — вектор коэффициентов регрессии,
- X — матрица признаков (независимых переменных), где каждая строка соответствует наблюдению, а каждый столбец — переменной. Первой колонкой обычно добавляется столбец единиц для учета свободного члена β_0,
- y — вектор зависимой переменной.

Множественная линейная регрессия

Шаги для расчета коэффициентов:

1. Подготовка данных: Создайте матрицу X, добавив в нее столбец единиц для свободного члена β_0.

2. Вычисление матрицы $X^T X$: Транспонируйте матрицу X и умножьте на саму матрицу X.

3. Вычисление обратной матрицы: Найдите обратную матрицу для результата, полученного на предыдущем шаге.

4. Вычисление $X^T y$: Умножьте транспонированную матрицу X на вектор y.

5. Умножение: Умножьте обратную матрицу на результат, полученный на предыдущем шаге.

Множественная линейная регрессия. Пример

Предположим, у нас есть следующие данные:

Площадь (x1)	Количество комнат (x2)	Цена (y)
50	2	15000
70	3	20000
100	4	30000

Множественная линейная регрессия. Пример

Шаги для расчета коэффициентов:

1. Создание матрицы X:

$$X = \begin{bmatrix} 1 & 50 & 2 \\ 1 & 70 & 3 \\ 1 & 100 & 4 \end{bmatrix}$$

2. Создание вектора y:

$$y = \begin{bmatrix} 150000 \\ 200000 \\ 300000 \end{bmatrix}$$

3. Вычисление:
   $X^T X$
   $(X^T X)^{-1}$
   $X^T y$
   Наконец, $\beta = (X^T X)^{-1} X^T y$