Лекция 4. Регрессионный анализ

Понятие регрессии

Регрессия – это зависимость среднего значения какой-либо величины от некоторой другой величины или от нескольких других величин. В отличие от чисто функциональной зависимости *y=f(x),* когда каждому значению независимой переменной *x* соответствует одно определённое значение зависимой переменной *y,* при регрессионной связи одному и тому же значению независимой переменной (фактору) *x* могут соответствовать в зависимости от конкретного случая различные значения зависимой переменной (отклика) *y*. Если при каждом значении $x=x_i$ наблюдается n_i значений y_{ij}; $j=\overline{1,n_i}$, то зависимость средних арифметических значений: $\overline{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ от x_i и является регрессией в статистическом понимании этого термина.

Изучение регрессии основано на том, что случайные величины *X* и *Y* связаны между собой вероятностной зависимостью: при каждом конкретном значении $X = x$ величина *Y* является случайной величиной со вполне определённым распределением вероятностей. Зависимость зависимой переменной – отклика от одной независимой переменной – фактора или нескольких факторов называется уравнением регрессии. По количеству факторов выделяют парную (однофакторную) и множественную (многофакторную) регрессию. Для парной будем рассматривать следующие методы регрессии: линейную, показательную, экспоненциальную, гиперболическую и параболическую.

Регрессионный анализ – это раздел математической статистики, изучающий регрессионную зависимость между случайными величинами по статистическим данным. Цель регрессионного анализа состоит в определении общего вида уравнения регрессии, вычислении оценок неизвестных параметров, входящих в уравнение регрессии проверке статистических гипотез о регрессионной связи.

В этом случае для обработки результатов рекомендуется применять регрессионный анализ, обладающий свойствами сравнительной простоты и конструктивности, которые заключаются в возможности использования регрессионных уравнений для генерации эффективных решений на основе оптимизационных методов. Отметим, что если переменные не количественные, а качественные, то рекомендуется использовать дисперсионный анализ. Если же часть переменных количественная, а часть качественная, то рекомендуется корреляционный анализ.

Таким образом, *регрессионный анализ* – набор статистических методов исследования влияния одной или нескольких независимых переменных $X_1, \dots, X_n$ на зависимую переменную Y. Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные – критериальными переменными.

Линейная регрессия

Линейная регрессия (Linear regression) – модель зависимости переменной *x* от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости.

Линейная регрессия относится к задаче определения «линии наилучшего соответствия» через набор точек данных и стала простым предшественником нелинейных методов, которые используют для обучения нейронных сетей.
Предположим, нам задан набор из 7 точек (рисунок 1).

Цель линейной регрессии — поиск линии, которая наилучшим образом соответствует этим точкам. Напомним, что общее уравнение для прямой есть $f(x) = b_0 + b_1 \cdot x$, где b_1 – наклон линии, а b_0 – его сдвиг. Таким образом, решение линейной регрессии определяет значения для b_0 и b_1, так что $f(x)$ приближается как можно ближе к y (рисунок 2).

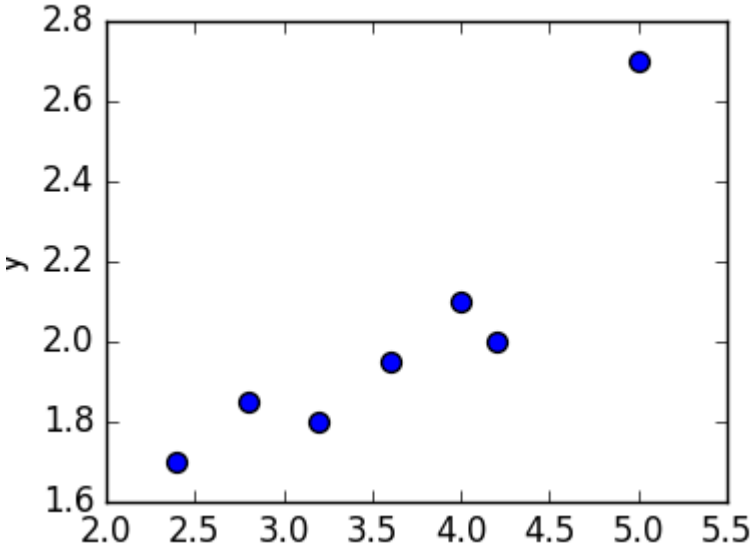

Рисунок 1 – Данные для линейной регрессии

Рассмотрим несколько графиков, потенциально соответствующих функции линейной регрессии (рисунок 2).

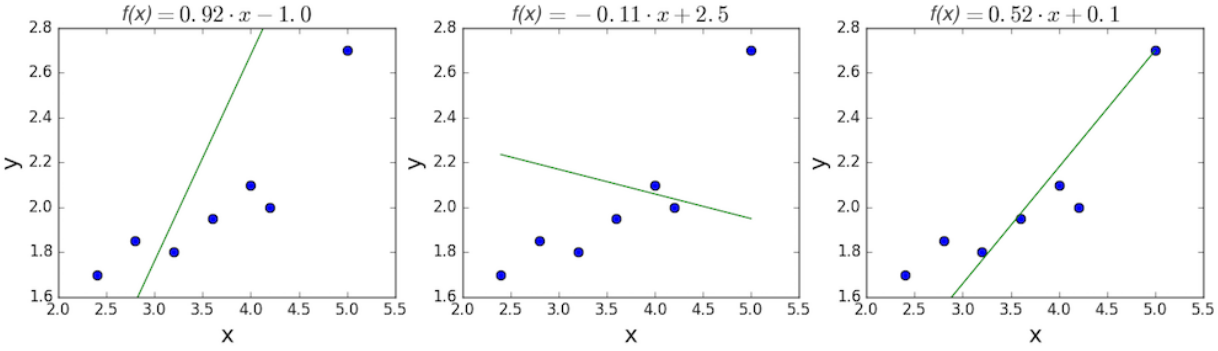

Рисунок 2 – Примеры графиков линейной регрессии

Из графиков видно, что первые две линии не соответствуют данным. Третья, похоже, лучше, чем две другие. Но как мы можем это проверить? Формально нам необходимо выразить, насколько хорошо подходит линия, и мы можем это сделать, определив функцию потерь.

Функция потерь — метод наименьших квадратов

Функция потерь – это мера количества ошибок, которые наша линейная регрессия делает на наборе данных. Хотя есть разные функции потерь, все они вычисляют расстояние между предсказанным значением *y(x)* и его

фактическим значением. Например, взяв строку из среднего примера выше, $f(x)=-0.11 \cdot x+2.5$, мы выделяем дистанцию ошибки между фактическими и прогнозируемыми значениями красными пунктирными линиями (рисунок 3).

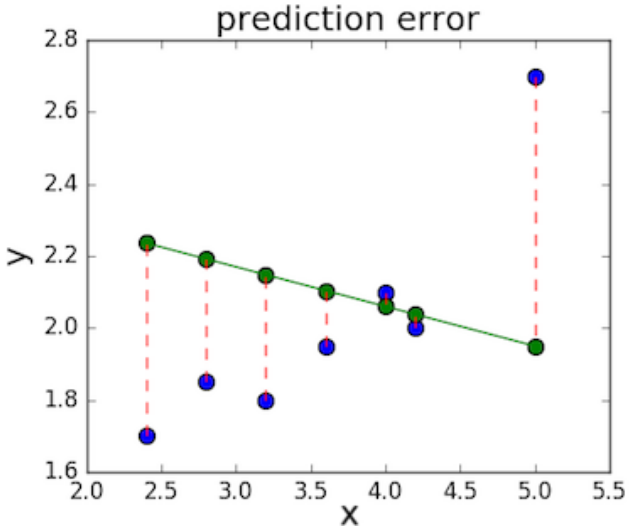

Рисунок 3 – Ошибки при линейной регрессии

Распространеной функцией потерь является функция средней квадратичной ошибки – Mean squared error (MSE). Чтобы вычислить MSE, необходимо рассчитать квадраты значения ошибок, как разность между эмпирическими данными и соответствующими им модельным, а после – усреднить:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2.$$

В машинном обучении MSE является широко представленным подходом к оценке ошибки благодаря простоте расчета и применимости в оценке точности регрессионной модели.

Множественная регрессия

Множественной называют линейную регрессию, в модели которой число независимых переменных две или более. Уравнение множественной линейной регрессии имеет вид:

$$f(x_1, x_2, \ldots, x_n) = b_0+b_1x_1+b_2x_2+\cdots+b_nx_n,$$

Как и в простой линейной регрессии, параметры модели вычисляются при помощи метода наименьших квадратов.

Отличие между простой и множественной линейной регрессией заключается в том, что вместо линии регрессии в ней используется гиперплоскость.

Преимущество множественной линейной регрессии по сравнению с простой заключается в том, что использование в модели нескольких входных переменных позволяет увеличить долю объяснённой дисперсии выходной переменной, и таким образом улучшить соответствие модели данным. Т.е. при добавлении в модель каждой новой переменной коэффициент детерминации растёт.

Метод наименьших квадратов

Выше говорилось, что регрессионный анализ основан на методе наименьших квадратов, который требует, чтобы сумма квадратов отклонений экспериментальных значений от вычисленных по аппроксимирующей зависимости была минимальной.

Метод наименьших квадратов (МНК) – один из наиболее часто используемых методов при обработке эмпирических данных, построении и анализе физических, биологических, технических, экономических и социальных моделей.

С помощью МНК решают задачу выбора параметров функции (заранее заданного вида) для приближённого описания зависимости величины y от величины x.

Исходные данные могут носить самый разнообразный характер и относиться к различным отраслям науки или техники. Например, зависимость температура воздуха (y) от высоты над уровнем моря (x) и другие зависимости.

Пусть необходимо установить функциональную зависимость между двумя эмпирическими данными x и y, значения которых занесены в следующую таблицу (таблица 1).

Таблица 1 – Пример начальных данных в задаче регрессии

x	x_1	x_2	…	x_i	…	x_n
y	y_1	y_2	…	y_i	…	y_n

Точки $(x_i; y_i)$ координатной плоскости принято называть *экспериментальными*.

Установим вид функции $y = f(x)$ по характеру расположения на координатной плоскости экспериментальных точек.

Если точки расположены так, как показано на рисунке 4, то разумно предположить, что между x и y существует линейная зависимость, выражающаяся формулой:

$$y = kx + b. \qquad (4.1)$$

Рассмотрим случай такой зависимости (рисунок 4).

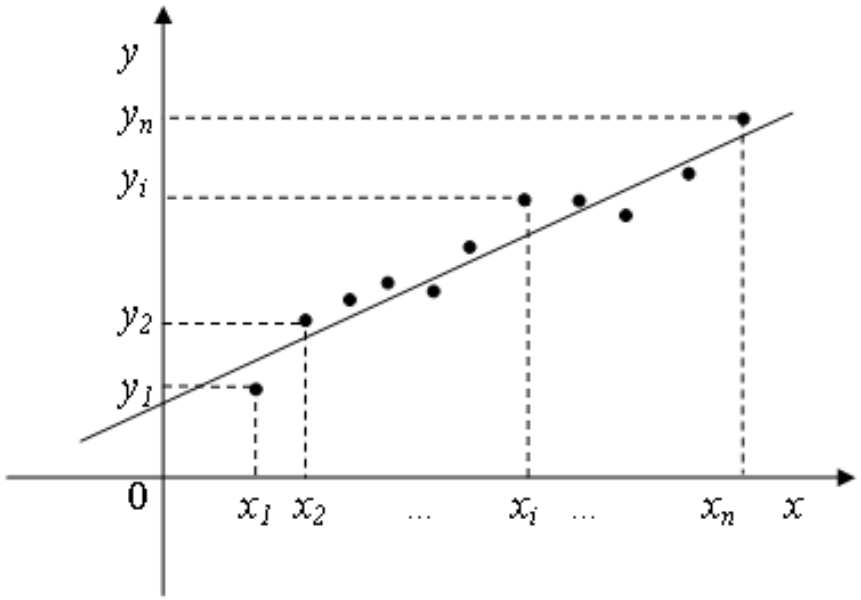

Рисунок 4 – Пример дискретной зависимости, близкой к линейной

Уравнение (4.1) можно представить в виде

$$y - (kx + b) = 0.$$

Так как точки $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$ не обязательно лежат на одной прямой, то, подставляя вместо x и y значения координат этих точек в выражение $y - (kx + b)$, получаем равенства:

$$y_1-(kx_1+b)=\delta_1,\ \ y_2-(kx_2+b)=\delta_2,\ \ldots,\ y_n-(kx_n+b)=\delta_n,$$

где $\delta_1,\ \delta_2,\ \ldots,\ \delta_n$ – некоторые числа, которые называют *погрешностями* (*отклонениями, невязками*).

Понятно, что чем меньше эти погрешности по абсолютной величине, тем лучше прямая, задаваемая уравнением $y=kx+b$, описывает зависимость между экспериментально полученными значениями *x* и *y*.

Сущность метода наименьших квадратов заключается в подборе коэффициентов *k* и *b* таким образом, чтобы сумма квадратов погрешностей была как можно меньшей:

$$S=\delta_1^2+\delta_2^2+\ldots+\delta_n^2=\sum_{i=1}^{n}\delta_i^2=\sum_{i=1}^{n}(y_i-(kx_i+b))^2\to\min \qquad (4.2)$$

Отметим, что в равенстве (4.2) находится сумма именно квадратов погрешностей, так как в случае суммирования самих погрешностей δ_i сумма может оказаться малой за счет разных знаков погрешностей.

Так как в равенстве (4.2) x_i и y_i – заданные числа, а *k* и *b* – неизвестные, то сумму *S* можно рассмотреть как функцию двух переменных *k* и *b*: $S=S(k,b)$. Исследуем ее на экстремум:

Необходимое условие существования экстремума функции двух переменных:

$$\begin{cases}\dfrac{\partial S}{\partial k}=0,\\[2ex] \dfrac{\partial S}{\partial b}=0;\end{cases}$$

$$\frac{\partial S}{\partial k}=2\sum_{i=1}^{n}(y_i-(kx_i+b))(-x_i)=-2\sum_{i=1}^{n}(y_i-(kx_i+b))\,x_i,$$

$$\frac{\partial S}{\partial b}=2\sum_{i=1}^{n}(y_i-(kx_i+b))(-1)=-2\sum_{i=1}^{n}(y_i-(kx_i+b)).$$

Приравнивая эти частные производные к нулю, получаем линейную систему двух уравнений с двумя переменными *k* и *b*:

$$\begin{cases} -2\sum_{i=1}^{n}(y_i-(kx_i+b))\,x_i=0, \\ -2\sum_{i=1}^{n}(y_i-(kx_i+b))=0. \end{cases}$$

Преобразуя первое уравнение системы, получим

$$-\sum_{i=1}^{n}y_ix_i+k\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i=0\,.$$

Преобразуя второе уравнение системы, получим

$$-\sum_{i=1}^{n}y_i+k\sum_{i=1}^{n}x_i+bn=0\,.$$

Откуда имеем систему:

$$\begin{cases} k\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_ix_i, \\ k\sum_{i=1}^{n}x_i+bn=\sum_{i=1}^{n}y_i. \end{cases}$$

Система (4.3) называется *нормальной системой*.

Из этой системы находим k и b, которые затем подставляем в уравнение (1) и получаем искомое уравнение прямой.

Тот факт, что функция $S=S(k,b)$ в найденной точке (k,b) имеет именно минимум, устанавливается с помощью частных производных второго порядка.

$$\frac{\partial^2 S}{\partial k^2}=-2\sum_{i=1}^{n}(-x_i)\,x_i=2\sum_{i=1}^{n}(x_i)^2,$$

$$\frac{\partial^2 S}{\partial b^2}=-2\cdot\sum_{i=1}^{n}(-1)=2n,$$

$$\frac{\partial^2 S}{\partial k\,\partial b}=-2\sum_{i=1}^{n}(-x_i)=2\sum_{i=1}^{n}x_i\,.$$

Вычислим $\Delta=\dfrac{\partial^2 S}{\partial k^2}\cdot\dfrac{\partial^2 S}{\partial b^2}-\left(\dfrac{\partial^2 S}{\partial k\,\partial b}\right)^2$.

$$\Delta=4n\sum_{i=1}^{n}(x_i)^2-\left(2\sum_{i=1}^{n}x_i\right)^2=2\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2.$$

Очевидно, $\Delta > 0$, следовательно, в найденной точке (k,b) функция $S = S(k,b)$ имеет экстремум; а так как $\frac{\partial^2 S}{\partial k^2} > 0$, то, согласно достаточному условию экстремума функции двух переменных, в точке (k,b) функция имеет минимум.

Полученная функция $y = kx + b$ называется *линейной регрессией*, а коэффициенты *k* и *b* – *коэффициентами регрессии* (величины *y* на *x*).

Зависимость между экспериментально полученными величинами может быть близка к квадратичной (рисунок 5). В этом случае задача состоит в нахождении коэффициентов a_2, a_1, a_0 для составления уравнения вида $y = a_2x^2 + a_1x + a_0$ (рисунок 5).

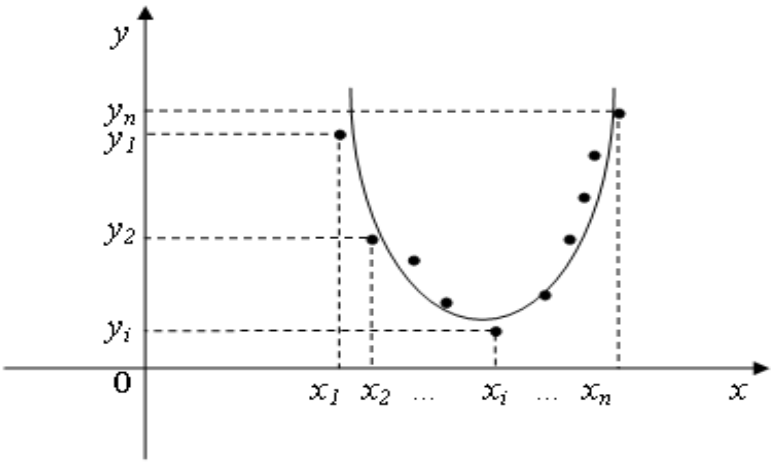

Рисунок 5 – Пример дискретной зависимости, близкой к квадратичной

Можно доказать, что для определения коэффициентов a_2, a_1, a_0 следует решить систему уравнений:

$$\begin{cases} n a_0 + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i, \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i, \\ a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i. \end{cases}$$

В экспериментальной практике в качестве приближающих функций, помимо линейной $y = kx + b$ и квадратичной $y = a_2 x^2 + a_1 x + a_0$, в зависимости от характера точечного графика часто используются следующие приближающие функции:

$$y = ax^m,\ \ y = ae^{mx},\ \ y = \frac{1}{ax+b},\ \ y = \frac{a}{x} + b,\ \ \ y = \frac{x}{ax+b},$$

$$y = a\ln x + b.$$

Очевидно, что когда вид приближающей функции установлен, задача сводится только к отысканию значений параметров.

Пример

Д.И. Менделеев в труде «Основы химии» приводит данные растворимости *y* натриевой селитры $NaNO_3$ на 100 г воды в зависимости от температуры t^0 (таблица 2).

Таблица 2 – Начальные условия задачи

t_i^0	0	4	10	15	21	29	35	51	68
y_i	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1

Соответствующая зависимость может быть представлена линейной функцией $y = kt + b$.

Требуется найти аппроксимирующую (приближаемую) функцию в предположении, что она является линейной.

Найдем коэффициенты *k* и *b*.

Для этого составим и решим нормальную систему уравнений

$$\begin{cases} k\sum_{i=1}^{n} t_i^2 + b\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i t_i, \\ k\sum_{i=1}^{n} t_i + bn = \sum_{i=1}^{n} y_i. \end{cases}$$

n – число эмпирических точек, $n = 9$.

Выполним предварительные расчеты и для удобства занесем их в таблицу (столбцы t_i, y_i, t_i^2, $t_i y_i$)

Таблица 3 – Пример рассчитанных значений

№	t_i	y_i	t_i^2	$t_i y_i$	$y_{pac.i} = kt_i + b_i$	δ_i	δ_i^2
1	0	66,7	0	0	67,55	-0,85	0,7225
2	4	71,0	16	284	71,03	-0,03	0,0009
3	10	76,3	100	763	76,25	0,05	0,0025
4	15	80,6	225	1209	80,6	0	0
5	21	85,7	441	1799,7	85,82	-0,12	0,0144
6	29	92,9	841	2694,1	92,78	0,12	0,0144
7	35	99,4	1225	3479	98	1,4	1,96
8	51	113,6	2601	5793,6	111,92	1,68	2,8224
9	68	125,1	4624	8506,8	126,71	-1,61	2,5921
∑	**233**	**811,3**	**10073**	**24529,2**			**8,19**

Таким образом, нормальная система принимает вид

$$\begin{cases} k \cdot 10073 + b \cdot 233 = 24529{,}2 \\ k \cdot 233 + b \cdot 9 = 811{,}3. \end{cases}$$

Решая систему, находим

$$k \approx 0{,}87$$

$$b \approx 67{,}55$$

Следовательно, уравнение искомой прямой

$$y = 0{,}87t + 67{,}55$$

Вычислим теперь для исходных значений t_i расчетные значения $y_{pac.i} = kt_i + b_i$ и занесем полученные результаты в таблицу (столбец $y_{pac.i} = kt_i + b_i$)

Найдем $\delta_i = y_i - (kx_i + b)$ и занесем результаты в таблицу (столбец δ_i). Вычислим сумму квадратов отклонений

$$S = \sum_{i=1}^{n} \delta_i^2 \approx 8{,}19\,.$$

В результате получим решение задачи, включающее сумму квадратов отклонений, определяющую точность, согласно выбранной метрике.

Контрольные вопросы по теме:

1. Приведите общую постановку задачи регрессионного анализа.
2. Приведите прикладной пример задачи регрессионного анализа.
3. В чём состоит задача парной линейной регрессии?
4. Сформулируйте и запишите постановку простейшей задачи парной линейной регрессии.
5. Как можно использовать линейную регрессию для построения прогнозов?
6. Запишите выражение для суммы квадратов отклонений от линии регрессии, когда искомая функциональная зависимость – многочлен второй степени.
7. Решите аналитически простейшую задачу парной линейной регрессии методом наименьших квадратов.
8. В чем суть множественной регрессии?