Методы машинного обучения

Лекция 4

Регрессионный анализ

Регрессионный анализ

Набор методов моделирования измеряемых данных и исследования их свойств, относится к разделам математической статистики и машинного обучения. Осуществляет исследование влияния одной или нескольких независимых переменных X_1, X_2, …, X_p на зависимую переменную Y.

Независимые переменные называют регрессорами, предикторами или объясняющими переменными, а зависимая переменная является результирующей, критериальной или регрессантом. *Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.*

Регрессионный анализ очень тесно связан с корреляционным анализом. В корреляционном анализе исследуется направление и теснота связи между количественными переменными. В регрессионном анализе исследуется форма зависимости между количественными переменными.

Наиболее распространенный вид регрессионного анализа — линейная регрессия, когда находят линейную функцию, которая, согласно определённым математическим критериям, наиболее соответствует данным. Например, в методе наименьших квадратов вычисляется прямая (или гиперплоскость), сумма квадратов отклонений которой от элементов данных минимальна.

Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных.

Цели и задачи регрессионного анализа

Цель регрессионного анализа – с помощью уравнения регрессии предсказать ожидаемое среднее значение результирующей переменной, т.е. определение степени детерминированности вариации критериальной (зависимой) переменной от независимых переменных (предикторов).

- Основные задачи регрессионного анализа следующие:
- определения вида и формы зависимости;
- оценка параметров уравнения регрессии;
- проверка значимости уравнения регрессии;
- проверка значимости отдельных коэффициентов уравнения (определение вклада отдельных независимых переменных в вариацию зависимой);
- построение интервальных оценок коэффициентов;
- исследование характеристик точности модели;
- построение точечных и интервальных прогнозов результирующей переменной (предсказание значения зависимой переменной с помощью независимых).

Как и корреляционный анализ, регрессионный анализ отражает только количественные зависимости между переменными. Причинно-следственные зависимости регрессионный анализ не отражает. Гипотезы о причинно-следственной связи переменных должны формулироваться и обосновываться исходя из теоретического анализа содержания изучаемого явления.

Математическое определение регрессии

- Пусть Y, X_1, X_2, …, X_p — случайные величины с заданным совместным распределением вероятностей.
- Если для каждого набора значений $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$ определено условное математическое ожидание (**уравнение регрессии в общем виде**):

$$\boldsymbol{f(x_1, x_2, \dots, x_p) = E(Y | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p)},$$

то функция $f(x_1, x_2, \dots, x_p)$ называется регрессией величины Y по величинам X_1, X_2,…, X_p, а ее график – **линией регрессии**.

- Зависимость Y от X_1, X_2, …, X_p проявляется в изменении средних значений Y при изменении X_1, X_2, …, X_p.
- При каждом фиксированном наборе значений $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$ величина Y остаётся случайной величиной с определённым распределением.
- Для выяснения вопроса, насколько точно регрессионный анализ оценивает изменение Y при изменении X_1, X_2, …, X_p, используется средняя величина дисперсии Y при разных наборах значений X_1, X_2, …, X_p (фактически речь идет о мере рассеяния зависимой переменной вокруг линии регрессии).

Регрессионная модель

Регрессионный анализ осуществляет поиск функции регрессионной зависимости $f(x) = E(y|x)$. Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих:

$$y = f(x) + \varepsilon$$

где f — функция регрессионной зависимости, а ε — аддитивная случайная величина с нулевым матожиданием. Обычно предполагается, что величина ε имеет гауссово распределение с нулевым средним и дисперсией σ_ε^2 .

Задача нахождения регрессионной модели нескольких свободных переменных:

Задана выборка: множество $\{x_1, x_2, \dots, x_N \,|x \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1, y_2, \dots, y_N \,|y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной. Совместно эти множества обозначаются как D - множество исходных данных $\{(x, y)_i\}$.

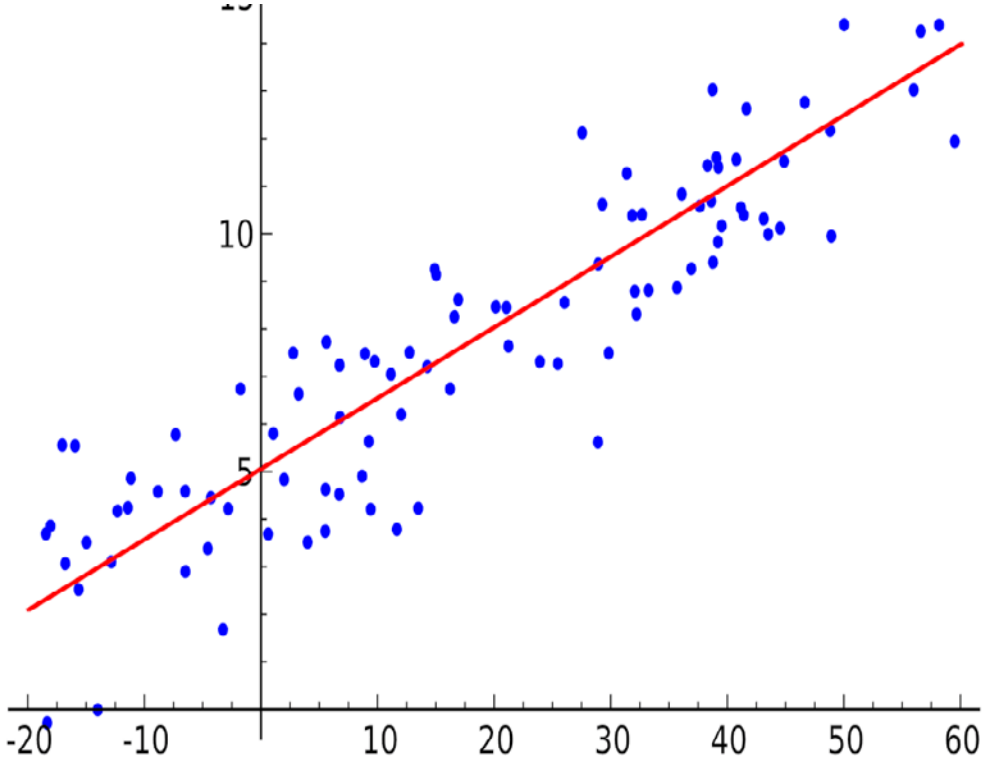
Регрессионная модель — параметрическое семейство функций $f(w, x)$ зависящая от параметров $w \in \mathbb{R}$ и свободных переменных x. Требуется найти наиболее вероятные параметры $\bar{w}$:

$$\bar{w} = \underset{w \in \mathbb{R}^w}{argmax}\ p(y|x, w, f) = p(D|w, f)$$

Функция вероятности p зависит от гипотезы порождения данных и задается Байесовским выводом или методом наибольшего правдоподобия.

Линейная регрессия

- используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной y от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости.
- относится к задаче определения «линии наилучшего соответствия» через набор точек данных и стала простым предшественником нелинейных методов, которые используют для обучения нейронных сетей.



Модель линейной регрессии

Линейная регрессия предполагает, что функция f зависит от параметров модели w линейно. При этом линейная зависимость от свободной переменной x необязательна:

$$y = f(w, x) + \varepsilon = \sum_{j=0}^{N} w_j \cdot g_j(x) + \varepsilon.$$

В случае, когда функция $g(x) = x$ линейная регрессия имеет вид:

$$f(w, x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_N x_N \text{ и } y = \sum_{j=0}^{N} w_j \cdot x_j + \varepsilon,$$

где x_j — компоненты вектора x (регрессоры или факторы модели), N – количество факторов модели, w_j - параметры модели (коэффициенты регрессии), ε – случайная ошибка модели.

Коэффициенты линейной регрессии показывают скорость изменения зависимой переменной по данному фактору, при фиксированных остальных факторах (в линейной модели эта скорость постоянна): $\forall j \; w_j = \frac{\partial f}{\partial x_j} = const$

Параметр w_0, при котором нет факторов, называют часто константой. Формально — это значение функции при нулевом значении всех факторов. Для аналитических целей удобно считать, что константа — это параметр при «факторе», равном 1, т.е. $x_0 = 1$ (или другой произвольной постоянной, поэтому константой называют также и этот «фактор»).

Парная и множественная регрессии

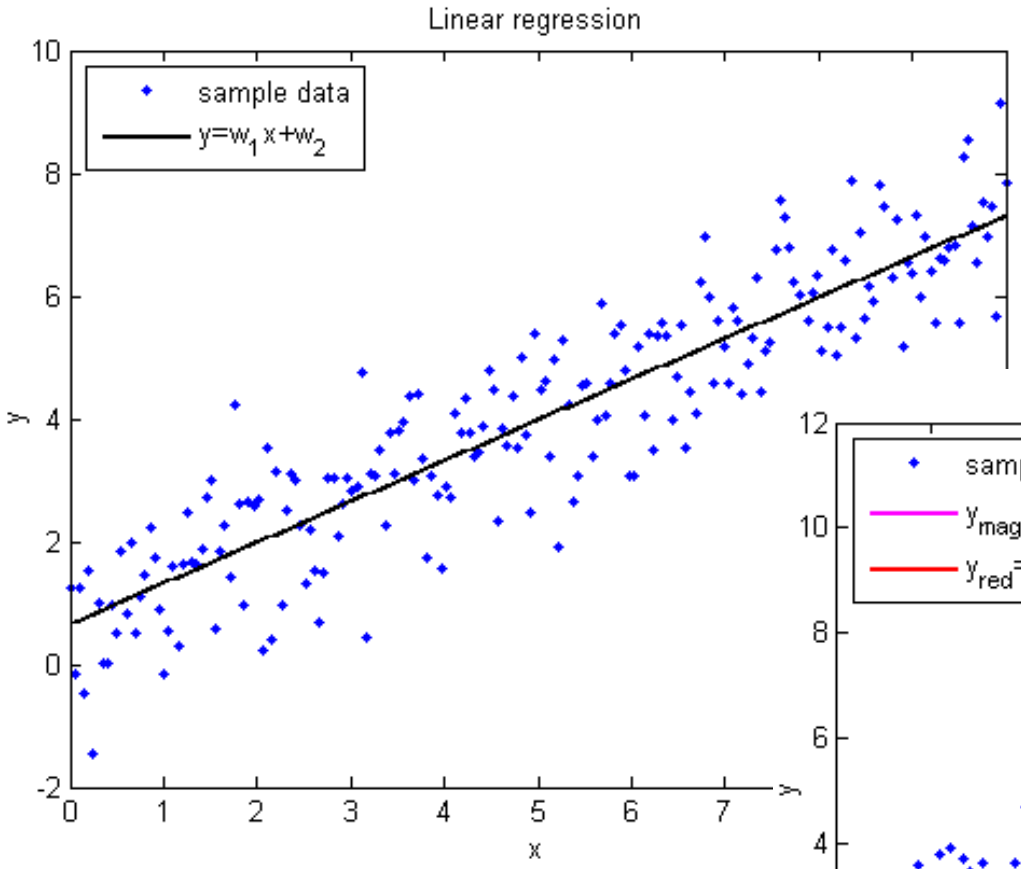
- В частном случае, когда фактор единственный (без учёта константы), говорят о парной или простейшей линейной регрессии:

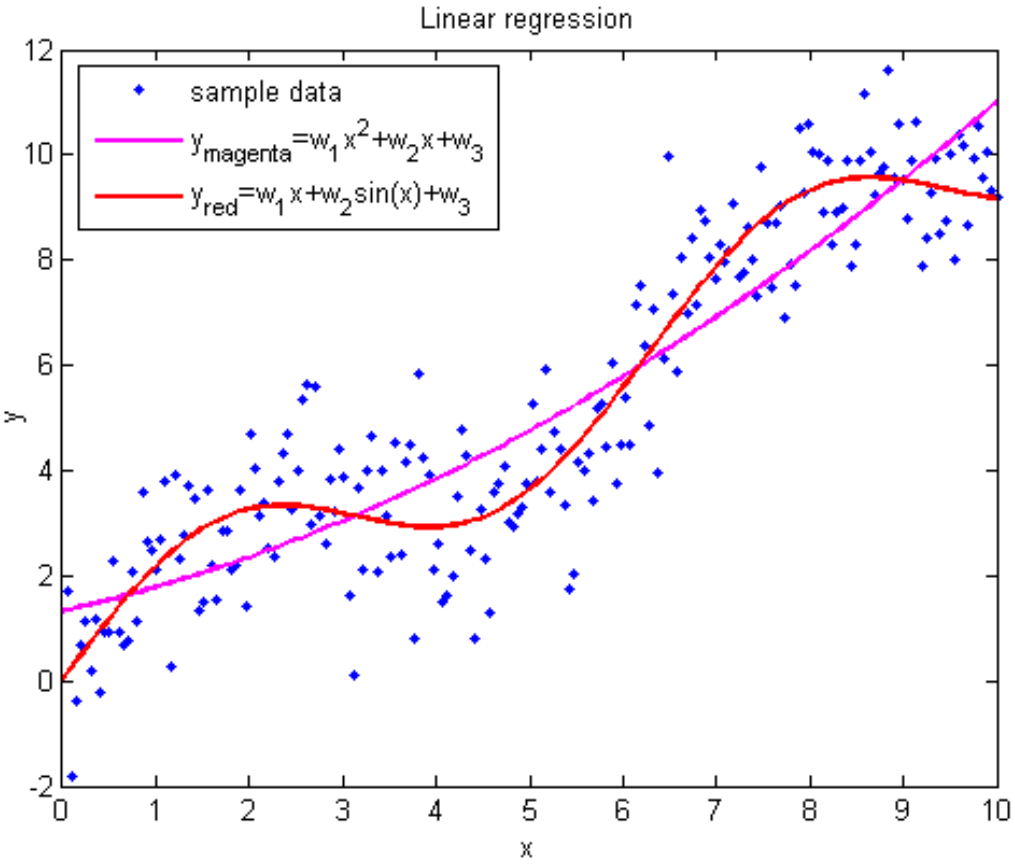
$$y_t = a + bx_t + \varepsilon_t$$

- Когда количество факторов (без учёта константы) больше одного, то говорят о множественной регрессии:

$$y_t = w_0 + w_1 x_{t1} + w_2 x_{t2} + \cdots + w_N x_{tN} + \varepsilon_t$$

Примеры моделей линейной регрессии

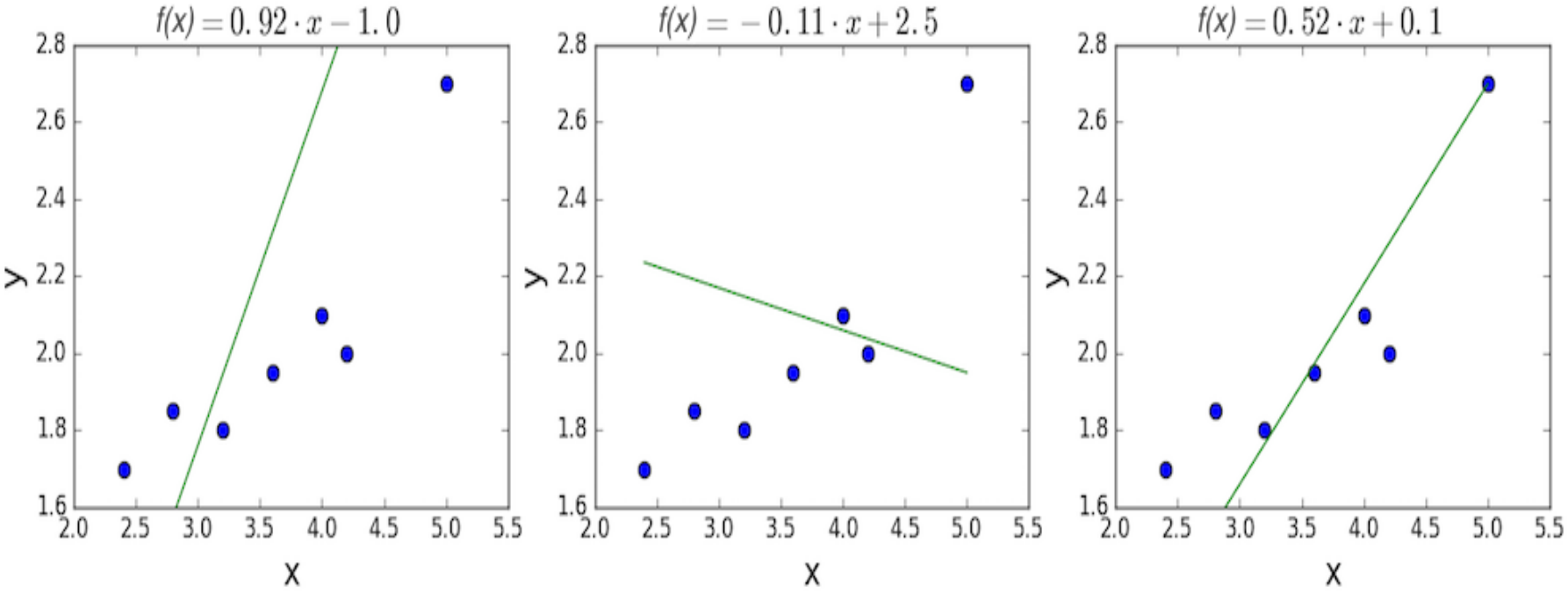




Применение линейной регрессии

Предположим, нам задан набор из 7 точек. Цель — поиск линии, которая наилучшим образом соответствует этим точкам. Попробуем несколько случайных кандидатов. Довольно очевидно, что первые две линии не соответствуют нашим данным. Третья, похоже, лучше, чем две другие. Но как мы можем это проверить?

Формально нам нужно выразить, насколько хорошо подходит линия, и мы можем это сделать, определив функцию потерь.

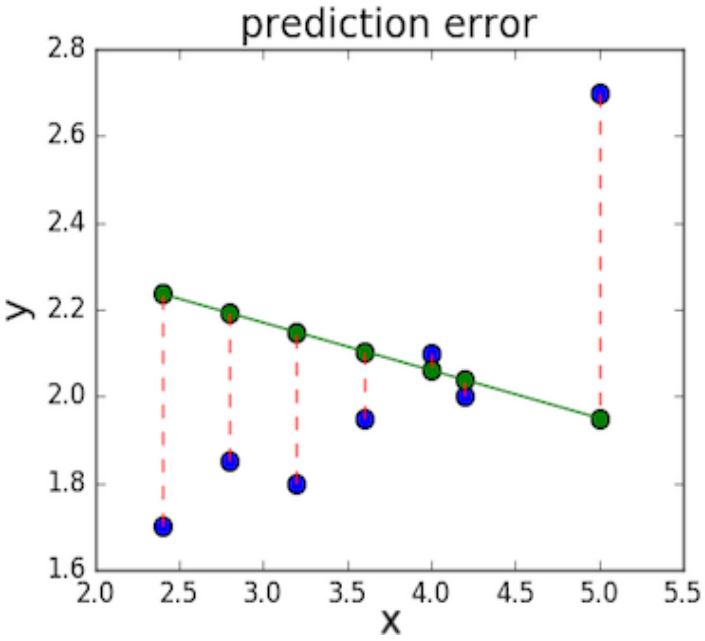


Сумма квадратов отклонений и среднеквадратичная ошибка

Разности $y_i - f(x_i)$ между фактическими значениями зависимой переменной и восстановленными называются регрессионными остатками (residuals). В литературе используются также синонимы: *невязки* и *ошибки*. Одной из важных оценок критерия качества полученной зависимости является **сумма квадратов остатков (**Sum of Squared Errors - SSE**)**:

$$\mathrm{SSE} = \sum_{i=1}^{M} (y_i - f(w, x_i))^2 = \sum_{i=1}^{M} (y_i - \hat{y}_i)^2$$

Сумма квадратов отклонений реально наблюдаемых y_i от их оценок $\hat{y}_i$.



Дисперсия остатков или среднеквадратичная ошибка (Mean Square Error – MSE) вычисляется по формуле:

$$\bar{\sigma}_{\varepsilon}^2 = \frac{\mathrm{SSE}}{N} = MSE$$

Метод наименьших квадратов - МНК

Метод нахождения оптимальных параметров линейной регрессии, таких, что сумма квадратов ошибок (регрессионных остатков) минимальна.

Есть система линейных уравнений: $Aw = y$, где A прямоугольная матрица размера $m \times n$, $m > n$ (то есть число строк матрицы А больше количества искомых переменных). Такая система уравнений в общем случае не имеет решения. Поэтому «решение» заключается в выборе вектора w, который минимизирует «расстояние» между векторами Aw и y. Для этого можно применить критерий минимизации суммы квадратов разностей левой и правой частей уравнений системы:

$$(Aw - y)^T (Aw - y) \to \min_w .$$

Решение этой задачи минимизации приводит к решению следующей системы уравнений: $A^T A\, w = A^T y$, которое называется *нормальным уравнением*. Если столбцы матрицы A линейно независимы, то матрица $A^T A$ обратима и единственное решение:

$$w = (A^T A)^{-1} A^T y$$

Отыскание решения w по методу наименьших квадратов эквивалентно задаче отыскания такой точки $p = Aw$, которая лежит ближе всего к y и находится при этом в пространстве столбцов матрицы A.

$$p = Aw = A(A^T A)^{-1} A^T y = Py$$

Матрица $P = A(A^T A)^{-1} A^T y$ называется матрицей проектирования вектора y на пространство столбцов матрицы A.

Пример построения линейной регрессии

Заданы: выборка (таблица) $D = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_M & y_M \end{pmatrix}$, регрессионная модель (линейная) — квадратичный полином $f = w_3 x^2 + w_2 x + w_1 = \sum_{j=1}^{3} w_j x^{j-1}$

Для нахождения оптимального значения вектора параметров $w = \langle w_1, \dots, w_3 \rangle^T$ выполняется следующая подстановка: $x_i^0 \to a_{i1}, x_i^1 \to a_{i2}, x_i^2 \to a_{i3}$.

Тогда матрица A значений подстановок свободной переменной x_i будет иметь вид:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \dots & \dots & \dots \\ a_{M1} & a_{M2} & a_{M3} \end{pmatrix}$$

Задан критерий качества модели (функция ошибки):

$S(w) = \sum_{i=1}^{M} (y_i - f(w, x_i))^2 = |Aw - y|^2 \to min$, где вектор $y = \langle y_1, \dots, y_M \rangle$.

Требуется найти такие параметры w, которые бы доставляли минимум $S(w)$: $w = \underset{w \in \mathbb{R}^3}{\operatorname{argmin}}(S)$

$$S(w) = |Aw - y|^2 = (Aw - y)^T (Aw - y) = y^T y - y^T Aw - w^T A^T y + w^T A^T Aw$$
$$= y^T y - 2y^T Aw + w^T A^T Aw$$

Для того, чтобы найти минимум функции невязки, требуется приравнять ее производные к нулю. Производные данной функции по w составляют:

$$\frac{\partial S}{\partial w} = -2A^T y + 2A^T Aw = 0$$

Это выражение совпадает с нормальным уравнением. Решение этой задачи должно удовлетворять системе линейных уравнений: $A^T A\, w = A^T y$, то есть, $w = (A^T A)^{-1} A^T y$.

МНК в регрессионном анализе (аппроксимация данных)

Сущность МНК (обычного, классического) заключается в том, чтобы найти такие параметры w, при которых сумма квадратов отклонений будет минимальной:

$$S = \sum_{k=1}^{M}(y_k - f(w, x_k))^2 = \sum_{k=1}^{M}(y_k - \hat{y}_k)^2 = \mathrm{SSE} \to min,\ w = \underset{w \in \mathbb{R}^3}{\mathrm{argmin}}(S)$$

В общем случае решение этой задачи может осуществляться численными методами оптимизации (минимизации). В этом случае говорят о *нелинейном МНК* (NLS или NLLS —*Non-Linear Least Squares*).

Для аналитического решения задачи минимизации необходимо найти стационарные точки функции $\mathrm{SSE(w)}$, продифференцировав её по неизвестным параметрам w, приравняв производные к нулю и решив полученную систему уравнений.

Для этого определим функцию невязки: $\sigma(\vec{w}) = \frac{1}{2}\sum_{k=1}^{M}(y_k - \hat{y}_k)^2$, где y_k - наблюдаемое значение, $\hat{y}_k = w_0 + w_1 x_{k1} + w_2 x_{k2} + \cdots + w_N x_{kN}$ - оценка y_k согласно модели, M— объём выборки, *N* – количество регрессоров (факторов).

Условие минимума функции невязки:

$$\begin{cases} \dfrac{\partial\sigma(\vec{w})}{\partial w_i} = 0 \\ i = 0 \ldots N \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{M} y_i = \sum_{i=1}^{M}\sum_{j=1}^{N} w_j x_{ij} + w_0 M \\ \displaystyle\sum_{i=1}^{M} y_i x_{ik} = \sum_{i=1}^{M}\sum_{j=1}^{N} w_j x_{ij} x_{ik} + w_0 \sum_{i=1}^{M} x_{ik} \\ k = 1, \ldots, N \end{cases}$$

Полученная система является системой $N + 1$ линейных уравнений с $N + 1$ неизвестными $w_0, \ldots, w_N$.

МНК - нахождение коэффициентов линейной регрессии

Если представить свободные члены левой части уравнений (предыдущий слайд) матрицей B, а коэффициенты при неизвестных в правой части — матрицей A, то получаем матричное уравнение: $B = A \times W$, для решения которого применимы методы решения СЛАУ, например метод Гаусса. Полученная матрица будет матрицей, содержащей коэффициенты уравнения линии регрессии W.

$$B = \begin{pmatrix} \sum_{i=1}^{M} y_i \\ \sum_{i=1}^{M} y_i x_{i,1} \\ \vdots \\ \sum_{i=1}^{M} y_i x_{i,N} \end{pmatrix}, \quad A = \begin{pmatrix} M & \sum_{i=1}^{M} x_{i,1} & \sum_{i=1}^{M} x_{i,2} & \dots & \sum_{i=1}^{M} x_{i,N} \\ \sum_{i=1}^{M} x_{i,1} & \sum_{i=1}^{M} x_{i,1} x_{i,1} & \sum_{i=1}^{M} x_{i,2} x_{i,1} & \dots & \sum_{i=1}^{M} x_{i,N} x_{i,1} \\ \sum_{i=1}^{M} x_{i,2} & \sum_{i=1}^{M} x_{i,1} x_{i,2} & \sum_{i=1}^{M} x_{i,2} x_{i,2} & \dots & \sum_{i=1}^{M} x_{i,N} x_{i,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{M} x_{i,N} & \sum_{i=1}^{M} x_{i,1} x_{i,N} & \sum_{i=1}^{M} x_{i,2} x_{i,N} & \dots & \sum_{i=1}^{M} x_{i,N} x_{i,N} \end{pmatrix}, \quad W = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_N \end{pmatrix}$$

МНК - случай полиномиальной модели

Если данные аппроксимируются полиномиальной функцией регрессии одной переменной $f(x) = w_0 + \sum_{i=1}^{k} w_i x^i$, то, воспринимая степени x^i как независимые факторы для каждого i можно оценить параметры модели исходя из общей формулы оценки параметров линейной модели.

Для этого в общей формуле достаточно учесть, что при такой интерпретации $x_{ti}x_{tj} = x_t^i x_t^j = x_t^{i+j}$ и $x_{tj}y_t = x_t^j y_t$. Следовательно, матричные уравнения в данном случае примут вид:

$$\begin{pmatrix} n & \sum\limits_n x_t & \dots & \sum\limits_n x_t^k \\ \sum\limits_n x_t & \sum\limits_n x_t^2 & \dots & \sum\limits_n x_t^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum\limits_n x_t^k & \sum\limits_n x_t^{k+1} & \dots & \sum\limits_n x_t^{2k} \end{pmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} \sum\limits_n y_t \\ \sum\limits_n x_t y_t \\ \vdots \\ \sum\limits_n x_t^k y_t \end{bmatrix}$$

Интерпретация параметров регрессии

- Параметры w_i являются частными коэффициентами корреляции x_i и y;
- $(w_i)^2$ - измеряет индивидуальный вклад x_i в объяснение y, при закреплении влияния остальных предикторов.
- В случае коррелирующих предикторов возникает проблема неопределённости в оценках, которые становятся зависимыми от порядка включения предикторов в модель.
- В нелинейных моделях регрессионного анализа, важно обращать внимание на тип нелинейности:
  - по независимым переменным (с формальной точки зрения легко сводящейся к линейной регрессии),
  - по оцениваемым параметрам (вызывающей серьёзные вычислительные трудности).

Нелинейная регрессия

Частный случай регрессионного анализа, в котором рассматриваемая регрессионная модель является функцией, зависящей от параметров и свободных переменных. Главное, что зависимость от параметров предполагается нелинейной.

Задана выборка из M пар (x_i, y_i) и регрессионная модель $f(w, x)$, которая зависит от параметров $w = (w_1, \dots, w_W)$ и свободной переменной x. Требуется найти такие значения параметров, которые доставляли бы минимум сумме квадратов регрессионных остатков:

$$S = \sum_{i=1}^{M} r_i^2 \text{, где } r_i = y_i - f(w, x_i) \text{ для } i = 1, \dots, M.$$

Для нахождения минимума функции S, приравняем к нулю её первые частные производные параметрам w:

$$\frac{\partial S}{\partial w_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial w_j} = 0, \text{где } j = 1, \dots, n.$$

Так как функция S в общем случае не имеет единственного минимума, то сначала назначается начальное значение вектора параметров w_0 и далее приближаться к оптимальному вектору по шагам:

$$w_j \approx w_j^{k+1} = w_j^k + \Delta w_j$$

где *k*- номер итерации, Δw_j- вектор шага.

Для нахождения оптимальных параметров нелинейных регрессионных моделей используются метод сопряжённых градиентов, метод Ньютона-Гаусса или алгоритм Левенберга-Марквардта.

Оценки качества регрессии (MSE, RMSE, MAE)

Средняя квадратичная ошибка (Mean Squared Error, MSE)

MSE применяется в ситуациях, когда нам надо подчеркнуть большие ошибки и выбрать модель, которая дает меньше больших ошибок прогноза (ошибки прогноза возводятся в квадрат).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$$

Корень из средней квадратичной ошибки (Root Mean Squared Error, RMSE)

RMSE получается из MSE путем извлечения корня.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2}$$

Средняя абсолютная ошибка (англ. Mean Absolute Error, MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f(x_i) - y_i\right|$$

Среднеквадратичный функционал сильнее штрафует за большие отклонения по сравнению со среднеабсолютным, и поэтому более чувствителен к выбросам. При использовании любого из этих функционалов может быть полезно проанализировать, какие объекты вносят наибольший вклад в общую ошибку.

Коэффициент детерминации *R-квадрат*

Коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Фактически, данная мера качества — это нормированная среднеквадратичная ошибка.

$$\boldsymbol{R^2} = 1 - \frac{\sum_{i=1}^{n}(f(x_i) - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Коэффициент детерминации принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется как соответствие модели данным.

Для приемлемых моделей $\boldsymbol{R^2}$ > 50 %. Модели с $\boldsymbol{R^2}$ > 80 % можно признать достаточно хорошими. Значение $\boldsymbol{R^2}$ = 1 означает функциональную зависимость между переменными.

Оценки качества регрессии (MAPE, SMAPE, MASE)

Средняя абсолютная процентная ошибка (Mean Absolute Percentage Error, MAPE)

$$MAPE = 100\% \times \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - f(x_i)|}{|y_i|}$$

Это коэффициент, не имеющий размерности, с простой интерпретацией. Его можно измерять в долях или процентах. Если MAPE=11.4%, то это говорит о том, что ошибка составила 11,4% от фактических значений.

Симметричная MAPE (Symmetric MAPE, SMAPE)

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{2|y_i - f(x_i)|}{|y_i| + |f(x_i)|}$$

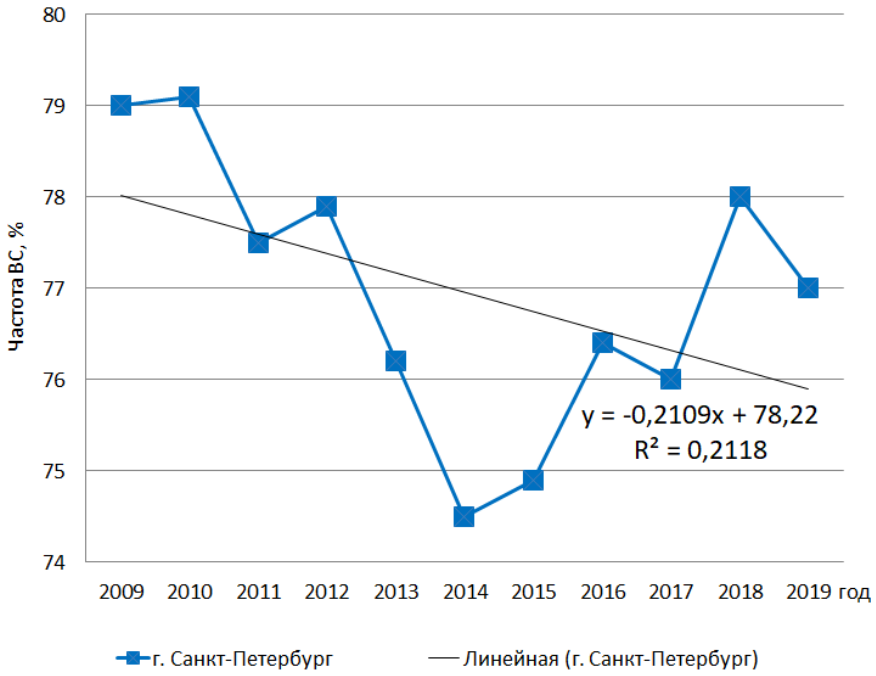
Средняя абсолютная масштабированная ошибка (Mean absolute scaled error, MASE)

$$MASE = \frac{T-1}{h}\frac{\sum_{j=1}^{h}|e_{T+j}|}{\sum_{t=2}^{T}|Y_i - Y_{i-1}|}$$

MASE имеет дело с двумя суммами: числитель соответствует тестовой выборке, знаменатель - обучающей. Ошибка не зависит от масштабов данных и является симметричной: то есть положительные и отрицательные отклонения от факта рассматриваются в равной степени.

Проблема MASE в том, что её тяжело интерпретировать. Например, MASE=1.21 ни о чём, по сути, не говорит. Это просто означает, что ошибка прогноза оказалась в 1.21 раза выше среднего абсолютного отклонения ряда в первых разностях, и ничего более.

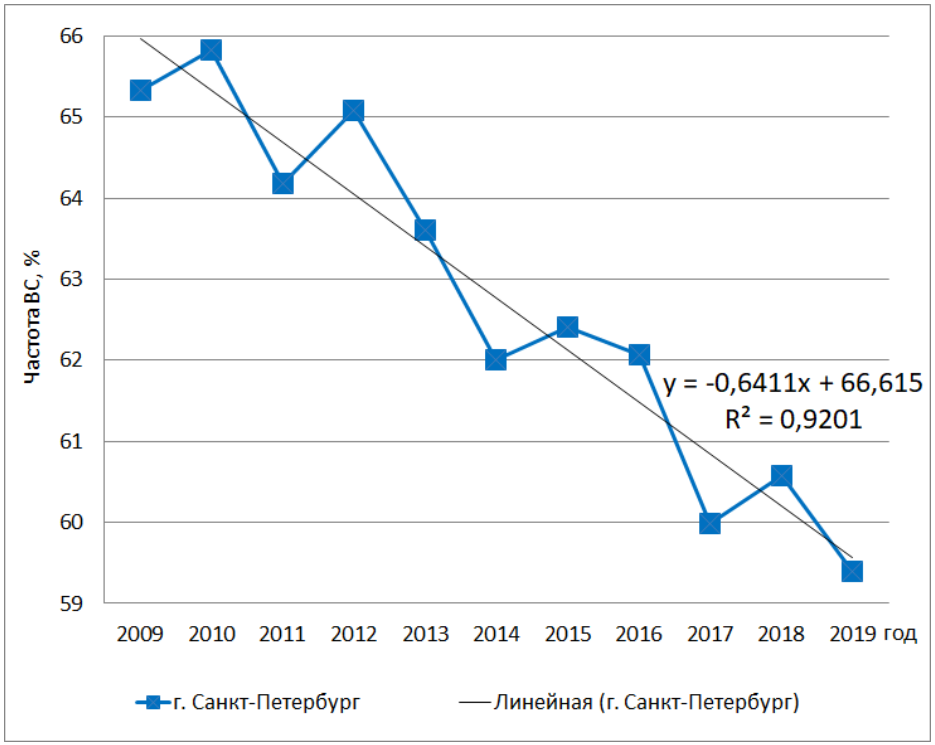
Пример парной линейной регрессии



$$y_t = a + bx_t + \varepsilon_t$$

$$\begin{pmatrix} n & \sum_{t=1}^{n} x_t \\ \sum_{t=1}^{n} x_t & \sum_{t=1}^{n} x_t^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} x_t y_t \end{pmatrix}$$

$$\begin{cases} \hat{b} = \dfrac{n\sum_{t=1}^{n} x_t y_t - (\sum_{t=1}^{n} x_t)(\sum_{t=1}^{n} y_t)}{n\sum_{t=1}^{n} x_t^2 - (\sum_{t=1}^{n} x_t)^2} \\ \hat{a} = \dfrac{\sum_{t=1}^{n} y_t - \hat{b}\sum_{t=1}^{n} x_t}{n} \end{cases}$$



Спасибо за внимание