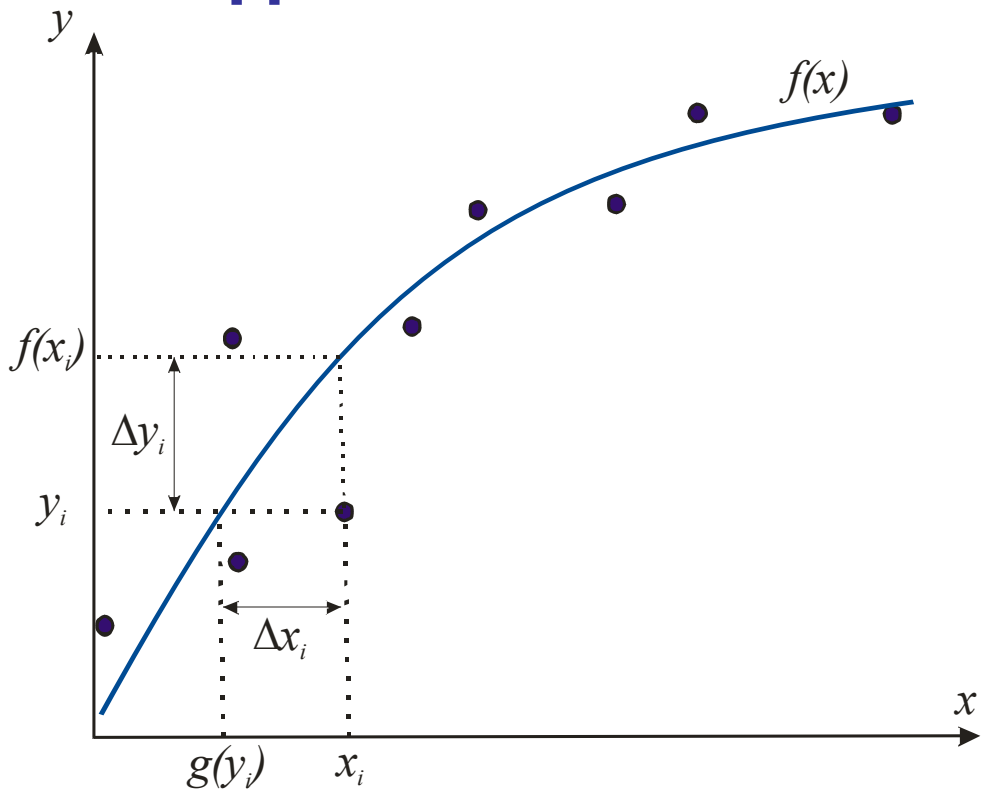
Аппроксимация методом наименьших квадратов

Линейная регрессия

Постановка задачи

на отрезке [a,b] заданы точки x_i,
(i = 0,...N, $x_0 = a$, $x_N = b$),
и значения функции $y(x_0)=y_0, \ldots, y(x_N)=y_N$.

является ли y некоторой функцией x:
$y=f(x)$??? **(1)**



Приближенная замена **неизвестной** функции более **простой**, которую нетрудно вычислить при любом значении аргумента в интервале его изменения. Такое приближение называют ***аппроксимацией***

$f(x)$ – регрессионная функция,
график - **линия регрессии**, в каком-то смысле "хорошо" расположенной относительно множества экспериментальных точек.

Критерий понятия «хорошо»- ???

Критерий: (Гаусс, 1821) мера, насколько "хорошо" лежат экспериментальные данные относительно линии регрессии - *сумма квадратов отклонений* Φ_Δ экспериментальных точек y_i от аппроксимирующей функции $f_i = f(x_i)$:

$$\Phi_{\Delta y} = \sum_{i=1}^{N} (\Delta y_i)^2 = \sum_{i=1}^{N} (y_i - f_i)^2 \qquad \textbf{(2)}$$

Наилучшая оценка положения теоретической линии - **минимальное** значение Φ_Δ (т.е. в этом случае погрешности минимальны).
Эта оценка *не зависит* от вида распределений величин x_i и y_i и их отклонений.

Задача приближения (аппроксимации) функции МНК: для экспериментальной функции $y(x)$, заданной таблично, найти функцию $f(x)$, *определенного* вида (его надо *выбрать*) так, чтобы сумма квадратов отклонений

$$\Phi_\Delta \rightarrow \min.$$

Тогда задача сводится к определению *параметров* A, B, C,... выбранного вида зависимости $f(x, A, B, C, \ldots)$.

Обычно используют простые по виду аналитические функции *f(x)*: линейная, полиномиальная, степенная, экспоненциальная, логарифмическая, обратная, ...

Линейная регрессия

Считаем, что переменная x является *независимой* $y=f(x)$.

$$f(x) = \mathrm{A}\cdot x + \mathrm{B}, \qquad (3)$$

Тогда: $\Delta y_i = y_i - (\mathrm{A}\cdot x_i + \mathrm{B}).$

$$\Phi_{\Delta y} = \sum_{i=1}^{\mathrm{N}} (\Delta y_i)^2 = \sum_{i=1}^{\mathrm{N}} (y_i - (\mathrm{A}\cdot x_i + \mathrm{B}))^2 \qquad (4)$$

$\Phi_{\Delta y}$ является функцией только A и B, а данные x_i, y_i *неизменны и не зависят* от A и B.
Для нахождения минимума функции $\Phi_{\Delta y}$= Φ(A,B):

$$\frac{\partial \Phi_{\Delta y}}{\partial \mathrm{A}} = \sum_{i=1}^{\mathrm{N}} 2(y_i - (\mathrm{A}\cdot x_i + \mathrm{B}))(-x_i)$$

$$\frac{\partial \Phi_{\Delta y}}{\partial \mathrm{B}} = \sum_{i=1}^{\mathrm{N}} 2(y_i - (\mathrm{A}\cdot x_i + \mathrm{B}))(-1)$$

Min($\Phi_{\Delta y}$): $\frac{\partial \Phi_{\Delta y}}{\partial \mathrm{A}} = 0; \quad \frac{\partial \Phi_{\Delta y}}{\partial \mathrm{B}} = 0$ **(5)**

Или, учитывая, что $\sum_{i=1}^{N} B = N \cdot B$

:

$$\begin{cases} \sum_{i=1}^{N} y_i x_i - A \cdot \sum_{i=1}^{N} x_i^{\,2} - B \cdot \sum_{i=1}^{N} x_i = 0 \\ \sum_{i=1}^{N} y_i - A \cdot \sum_{i=1}^{N} x_i - N \cdot B = 0 \end{cases} \quad \textbf{(5*)}$$

Из второго уравнения:

$$B = \frac{\sum_{i=1}^{N} y_i}{N} - A \cdot \frac{\sum_{i=1}^{N} x_i}{N} \quad \textbf{(6)}$$

или $\bar{y} = A\bar{x} + B$ Линия регрессии проходит через *центр тяжести* облака точек с координатами

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Подставляя выражение (6) в первое уравнение (5*):

$$A = \frac{N \cdot \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{N \cdot \sum_{i=1}^{N} x_i^{\,2} - \left(\sum_{i=1}^{N} x_i\right)^2} \quad \textbf{(7)}$$

Подставляя (5) в (6), найдем значение B.

Эти величины определяются только всеми значениями x_i и y_i и их количеством N.

Считаем, что переменная y является *независимой* $x=g(y).$

Тогда уравнение прямой линии запишем в виде:

$$x=\mathrm{C}\cdot y+\mathrm{D}$$

где C и D - неизвестные параметры.

$$\Delta x_i = x_i - (\mathrm{C}\cdot y_i + \mathrm{D}).$$

$$\Phi_{\Delta x} = \sum_{i=1}^{\mathrm{N}} (\Delta x_i)^2 = \sum_{i=1}^{\mathrm{N}} (x_i - (\mathrm{C}\cdot y_i + \mathrm{D}))^2 \quad . \qquad \textbf{(8)}$$

Для нахождения значений параметров C и D, при котором достигается минимум функции $\Phi_{\Delta x} = \Phi(\mathrm{C},\mathrm{D})$, дифференцируем её по C и D, и приравняем производные нулю. Действуя аналогично (5)–(7):

$$\mathrm{D} = \frac{\sum_{i=1}^{\mathrm{N}} x_i}{\mathrm{N}} - \mathrm{C}\cdot\frac{\sum_{i=1}^{\mathrm{N}} y_i}{\mathrm{N}} \quad . \textbf{(9)}$$

или $\overline{x} = C\overline{y} + D$ Эта линия регрессии **также** проходит через *центр тяжести* облака точек

$$\mathrm{C} = \frac{\mathrm{N}\cdot\sum_{i=1}^{\mathrm{N}} x_i y_i - \sum_{i=1}^{\mathrm{N}} x_i \cdot \sum_{i=1}^{\mathrm{N}} y_i}{\mathrm{N}\cdot\sum_{i=1}^{\mathrm{N}} {y_i}^2 - \left(\sum_{i=1}^{\mathrm{N}} y_i\right)^2} \qquad \textbf{(10)}$$

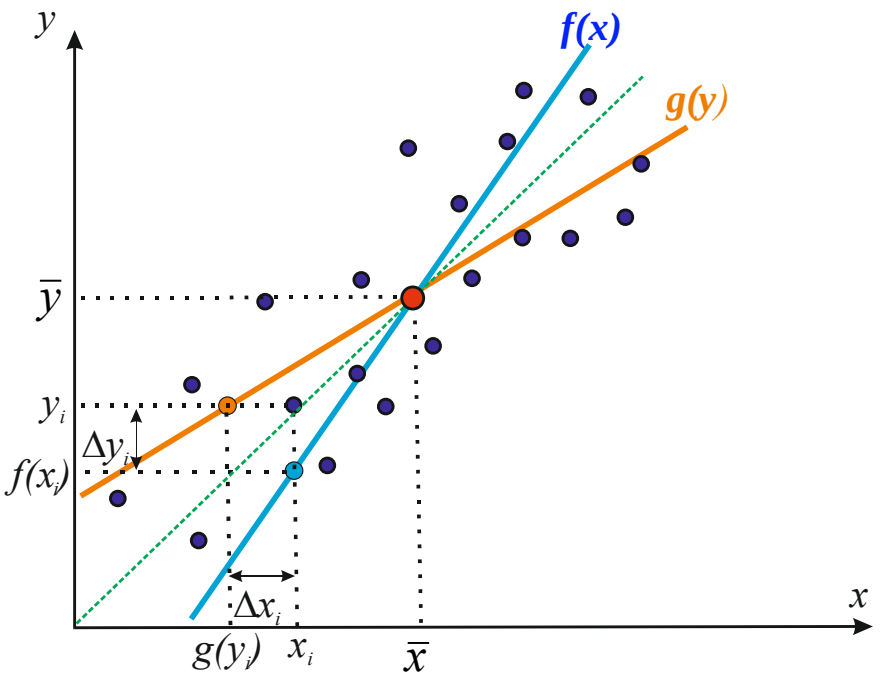
Оценка качества аппроксимации

коэффициент корреляции Пирсона величин x_i и y_i:

$$R = \sqrt{|\mathrm{A} \cdot \mathrm{C}|}, \quad \textbf{(11)}$$

или

$$R = \frac{\left(\mathrm{N} \cdot \sum_{i=1}^{\mathrm{N}} x_i y_i - \sum_{i=1}^{\mathrm{N}} x_i \cdot \sum_{i=1}^{\mathrm{N}} y_i \right)}{\sqrt{\left[\mathrm{N} \cdot \sum_{i=1}^{\mathrm{N}} x_i^{\,2} - \left(\sum_{i=1}^{\mathrm{N}} x_i \right)^2 \right] \cdot \left[\mathrm{N} \cdot \sum_{i=1}^{\mathrm{N}} y_i^{\,2} - \left(\sum_{i=1}^{\mathrm{N}} y_i \right)^2 \right]}}$$



$$0 \le R \le 1$$

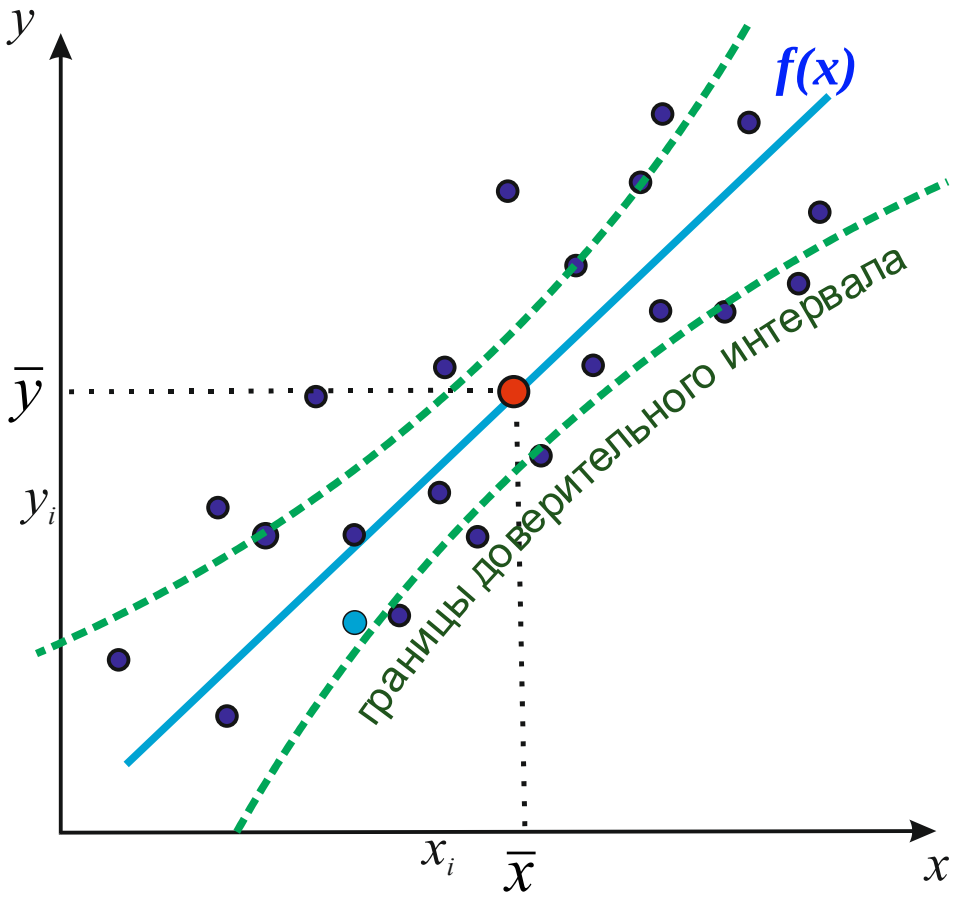
R = 1: C = 1/A, D = – B/A (линии $f(x)$ и $g(y)$ сливаются)

R ≈ 0 функциональной зависимости нет: A = 0, C = ∝ (линии $f(x)$ и $g(y)$ ортогональны).

Среднеквадратичная ошибка коэффициентов A и B:

$$s_A = \sqrt{\frac{\sum_{i=1}^{N} y_i^{\,2} - A\sum_{i=1}^{N} x_i y_i}{(N-2)\left[\sum_{i=1}^{N} x_i^{\,2} - \left(\sum_{i=1}^{N} \frac{x_i}{N}\right)^2\right]}}$$

$$s_B = s_A \cdot \sqrt{\bar{x}^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$



$$y = f(x) = (A \pm S_A)x + (B \pm S_B)$$

функции Excell:

Добавить линию тренда (на графике)

ЛИНЕЙН()

Использование критерия Стьюдента для проверки значимости параметров регрессионной модели

Проверка статистической значимости **параметров** регрессионного уравнения (коэффициентов регрессии) выполняется по *t*-критерию Стьюдента, которй рассчитывается по формуле:

$$t = \frac{|A|}{s_A}$$

где A - значение параметра;

s_A - стандартное отклонение параметра.

Рассчитанное значение критерия Стьюдента сравнивают с его табличным значением при выбранной доверительной вероятности (как правило, α=0.05) и числе степеней свободы $f=N-k-1$,
где N - число точек, k-число переменных в регрессионном уравнении (например, для линейной модели $y=Ax+B$ k=1 и $f=N-2$).

Если вычисленное значение *t* больше, чем критическое (табличное или вычисленное),

$t > t_{кр}$ (= СТЬЮДЕНТ.ОБР.2Х(α; $N-k-1$)
или T.INV.2T)

то коэффициент регрессии (A) является **значимым** с заданной доверительной вероятностью.

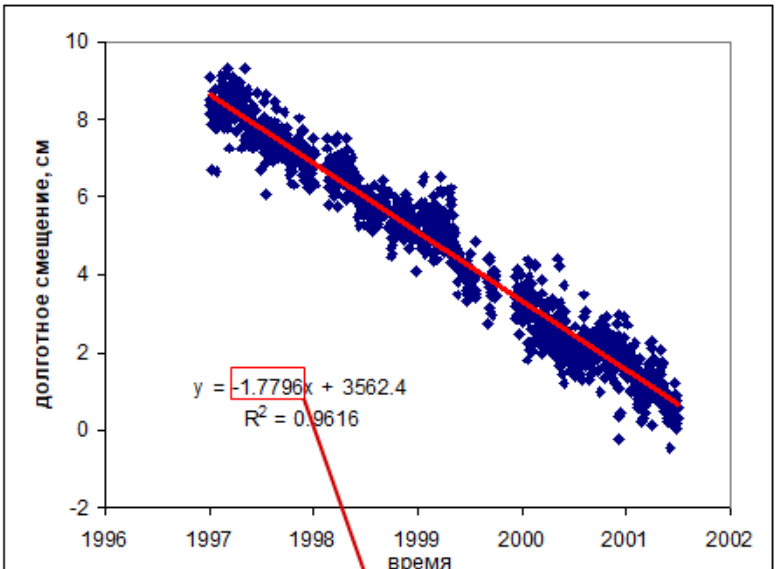
Аналогично можно оценивать значимость параметра B

Критические уровни распределения Стьюдента

α	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df (n-1)											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300

Применения

Вычисление скорости движения пункта GPS



= скорость движения по данной координате, в см/год