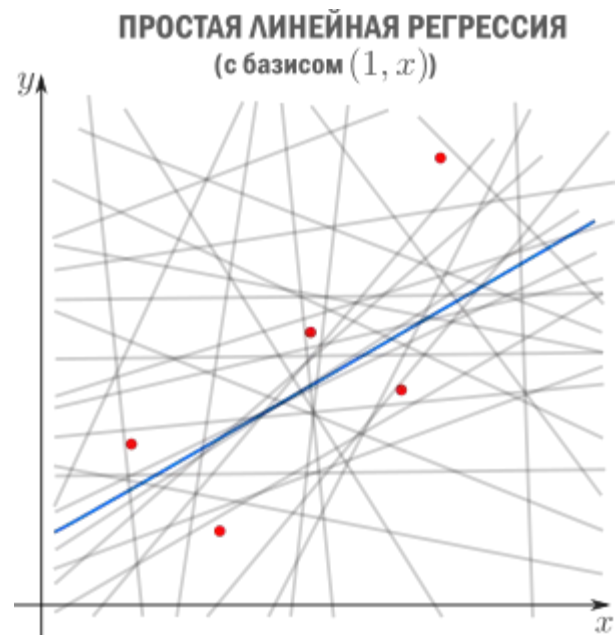


Линейная регрессия

Регрессия — способ выбрать из семейства функций ту, которая минимизирует функцию потерь. Последняя характеризует насколько сильно пробная функция отклоняется от значений в заданных точках. Если точки получены в эксперименте, они неизбежно содержат ошибку измерений, шум, поэтому разумнее требовать, чтобы функция передавала общую тенденцию, а не точно проходила через все точки. В каком-то смысле регрессия — это «интерполирующая аппроксимация»: мы хотим провести кривую как можно ближе к точкам и при этом сохранить ее максимально простой чтобы уловить общую тенденцию. За баланс между этими противоречивыми желаниями как-раз отвечает функция потерь (в английской литературе «loss function» или «cost function»).



Цель регрессии — найти коэффициенты этой линейной комбинации, и тем самым определить регрессионную функцию (которую также называют *моделью*). Отмечу, что линейную регрессию называют линейной именно из-за линейной комбинации базисных функций — это не связано с самими базисными функциями (они могут быть линейными или нет).

$$f = \sum_i w_i f_i.$$

Метод наименьших квадратов

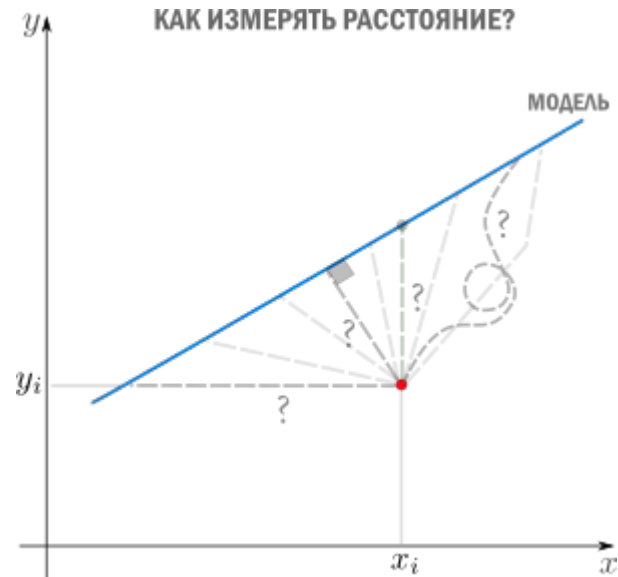
Начнём с простейшего двумерного случая. Пусть нам даны точки на плоскости $\{(x_1, y_1), \dots, (x_N, y_N)\}$

и мы ищем такую аффинную функцию

$$f(x) = a + b \cdot x,$$

чтобы ее график ближе всего находился к точкам. Таким образом, наш базис состоит из константной функции и линейной .

Как видно из иллюстрации, расстояние от точки до прямой можно понимать по-разному, например геометрически — это длина перпендикуляра. Однако в контексте нашей задачи нам нужно функциональное расстояние, а не геометрическое. Нас интересует разница между экспериментальным значением и предсказанием модели для каждого — поэтому измерять нужно вдоль оси .

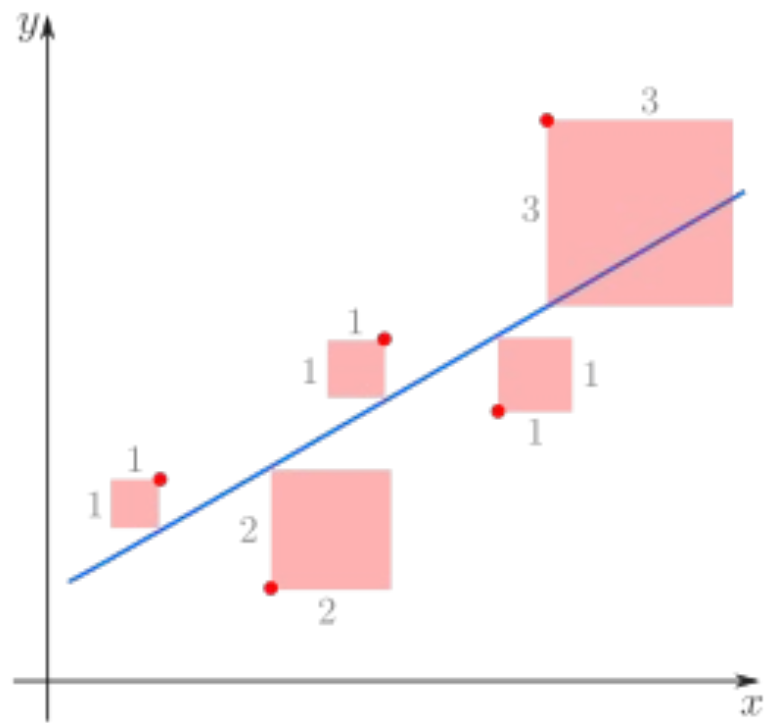
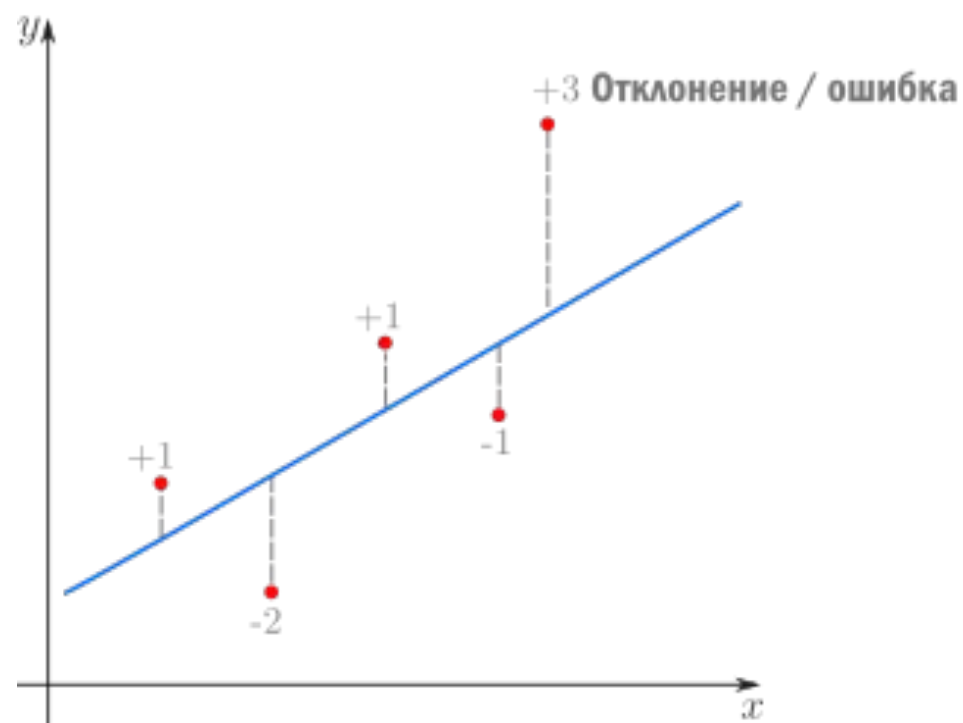


Первое, что приходит в голову, в качестве функции потерь попробовать выражение, зависящее от абсолютных значений разниц $|f(x_i) - y_i|$. Простейший вариант — сумма модулей отклонений $\sum_i |f(x_i) - y_i|$ приводит к Least Absolute Distance (LAD) регрессии.

Впрочем, более популярная функция потерь — сумма квадратов отклонений регрессанта от модели. В англоязычной литературе она носит название Sum of Squared Errors (SSE)

$$\text{SSE}(a, b) = \text{SS}_{\text{residuals}} = \sum_{i=1}^N \text{отклонение}_i^2 = \sum_{i=1}^N (y_i - f(x_i))^2 = \sum_{i=1}^N (y_i - a - b \cdot x_i)^2,$$

Метод наименьших квадратов (по англ. OLS) — линейная регрессия с в качестве функции потерь.



Наша задача — найти параметры \hat{a} и \hat{b} , минимизирующие $SSE(a,b)$. Эту функцию иногда называют функцией ошибок, функцией соответствия или функцией потерь.

$$(\hat{a}, \hat{b}) = \operatorname{argmin}_{a,b} SSE(a,b).$$

Простейший способ найти — вычислить частные производные по a и b , приравнять их нулю и решить систему линейных уравнений

$$\frac{\partial}{\partial a} SSE(a,b) = -2 \sum_{i=1}^N (y_i - a - bx_i),$$

$$\frac{\partial}{\partial b} SSE(a,b) = -2 \sum_{i=1}^N (y_i - a - bx_i)x_i.$$

Значения параметров, минимизирующие функцию потерь, удовлетворяют уравнениям

$$0 = -2 \sum_{i=1}^N (y_i - \hat{a} - \hat{b}x_i),$$

$$0 = -2 \sum_{i=1}^N (y_i - \hat{a} - \hat{b}x_i)x_i,$$

$$\hat{a} = \frac{\sum_i y_i}{N} - \hat{b} \frac{\sum_i x_i}{N},$$

$$\hat{b} = \frac{\frac{\sum_i x_i y_i}{N} - \frac{\sum_i x_i}{N} \frac{\sum_i y_i}{N}}{\frac{\sum_i x_i^2}{N} - \left(\frac{\sum_i x_i}{N}\right)^2}.$$

Полученные формулы можно компактно записать с помощью статистических эstimаторов: среднего $\langle \cdot \rangle$, вариации σ . (стандартного отклонения), ковариации $\sigma(\cdot, \cdot)$ и корреляции $\rho(\cdot, \cdot)$

$$\hat{a} = \langle y \rangle - \hat{b} \langle x \rangle,$$

$$\hat{b} = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\langle x^2 \rangle - \langle x \rangle^2}.$$

Перепишем \hat{b} как

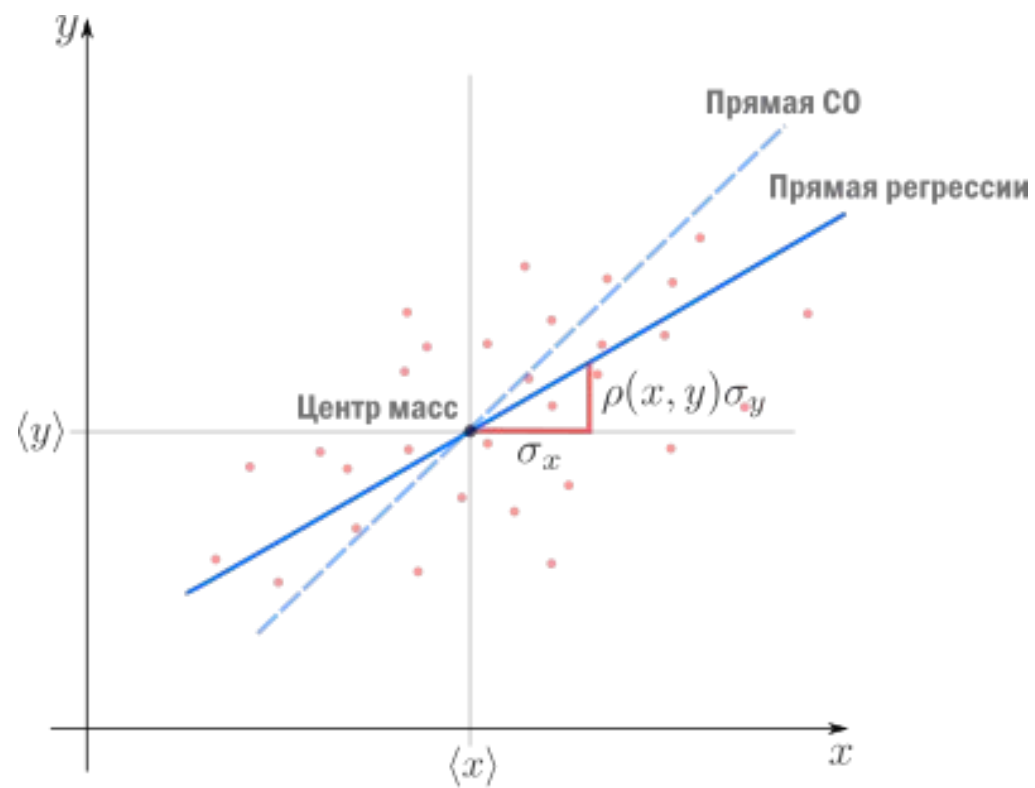
$$\hat{b} = \frac{\sigma(x, y)}{\sigma_x^2},$$

где σ_x это нескорректированное (смещенное) стандартное выборочное отклонение, а $\sigma(x, y)$ — ковариация. Теперь вспомним, что коэффициент корреляции (коэффициент корреляции Пирсона)

$$\rho(x, y) = \frac{\sigma(x, y)}{\sigma_x \sigma_y}$$

и запишем

$$\hat{b} = \rho(x, y) \frac{\sigma_y}{\sigma_x}.$$



Теперь мы можем оценить все изящество дескриптивной статистики, записав уравнение регрессионной прямой так

$$y - \langle y \rangle = \rho(x, y) \frac{\sigma_y}{\sigma_x} (x - \langle x \rangle).$$

Во-первых, это уравнение сразу указывает на два свойства регрессионной прямой:

- прямая проходит через центр масс $(\langle x \rangle, \langle y \rangle)$;
- если по оси x за единицу длины выбрать σ_x , а по оси y — σ_y , то угол наклона прямой будет от -45° до 45° . Это связано с тем, что $-1 \leq \rho(x, y) \leq 1$.

Во-вторых, теперь становится понятно, почему метод регрессии называется именно так. В единицах стандартного отклонения отклоняется от своего среднего значения меньше чем , потому что:

$$|\rho(x, y)| \leq 1$$

Возведя коэффициент корреляции в квадрат, получим коэффициент детерминации $R = \rho^2$. Квадрат этой статистической меры показывает насколько хорошо регрессионная модель описывает данные. R^2 , равный 1, означает что функция идеально ложится на все точки — данные идеально скоррелированы. Можно доказать, что R^2 показывает какая доля вариативности в данных объясняется лучшей из линейных моделей. Чтобы понять, что это значит, введем определения

$$\text{Var}_{data} = \frac{1}{N} \sum_i (y_i - \langle y \rangle)^2,$$

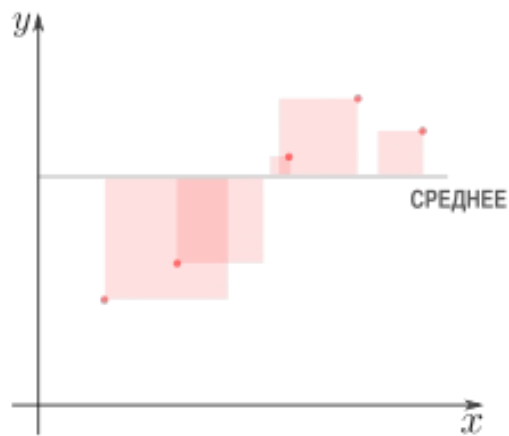
$$\text{Var}_{res} = \frac{1}{N} \sum_i (y_i - \text{модель}(x_i))^2,$$

$$\text{Var}_{reg} = \frac{1}{N} \sum_i (\text{модель}(x_i) - \langle y \rangle)^2.$$

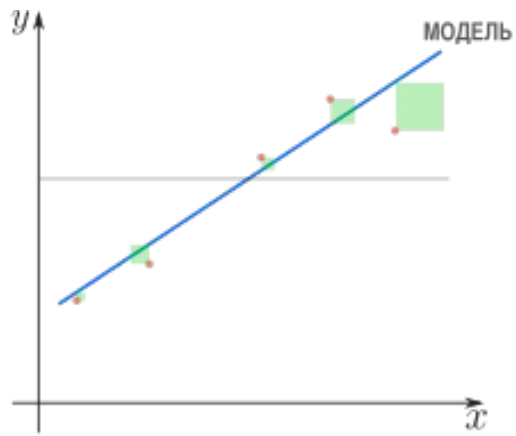
Var_{data} — вариация исходных данных (вариация точек y_i).

Var_{res} — вариация остатков, то есть вариация отклонений от регрессионной модели — от y_i нужно отнять предсказание модели и найти вариацию.

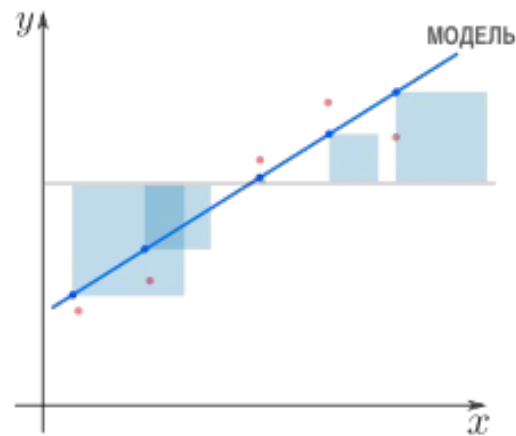
Var_{reg} — вариация регрессии, то есть вариация предсказаний регрессионной модели в точках x_i (обратите внимание, что среднее предсказаний модели совпадает с $\langle y \rangle$).



$$\text{Var}_{data} = \text{[red squares]} + \text{[red squares]} + \dots + \text{[red squares]} + \text{[red squares]}$$



$$\text{Var}_{res} = \text{[green squares]} + \text{[green squares]} + \dots + \text{[green squares]} + \text{[green squares]}$$



$$\text{Var}_{reg} = \text{[blue squares]} + \text{[blue squares]} + \dots + \text{[blue squares]} + \text{[blue squares]}$$

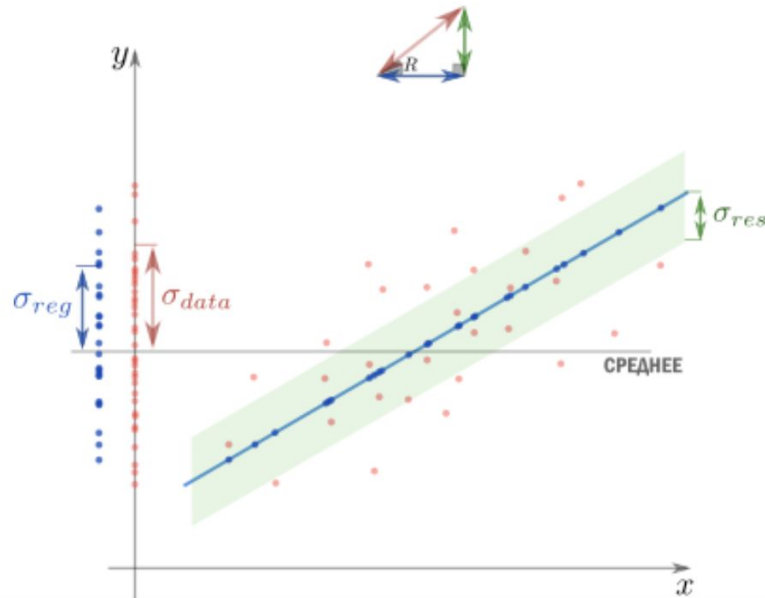
Дело в том, что вариация исходных данных разлагается в сумму двух других вариаций: вариации случайного шума (остатков) и вариации, которая объясняется моделью (регрессии)

$$\text{Var}_{data} = \text{Var}_{res} + \text{Var}_{reg}.$$

или

$$\sigma_{data}^2 = \sigma_{res}^2 + \sigma_{reg}^2.$$

Как видим, стандартные отклонения образуют прямоугольный треугольник.



Проблема выбора размерности

На практике часто приходится самостоятельно строить модель явления, то есть определяться сколько и каких нужно взять базисных функций. Первый порыв «набрать побольше» может сыграть злую шутку: модель окажется слишком чувствительной к шумам в данных (переобучение). С другой стороны, если излишне ограничить модель, она будет слишком грубой (недообучение).

Есть два способа выйти из ситуации. Первый: последовательно наращивать количество базисных функций, проверять качество регрессии и вовремя остановиться. Или же второй: выбрать функцию потерь, которая определит число степеней свободы автоматически. В качестве критерия успешности регрессии можно использовать коэффициент детерминации, о котором уже упоминалось выше, однако, проблема в том, что R^2 монотонно растет с ростом размерности базиса. Поэтому вводят скорректированный коэффициент

$$\bar{R}^2 = 1 - (1 - R^2) \left[\frac{N - 1}{N - (n + 1)} \right],$$

где N — размер выборки, n — количество независимых переменных. Следя за \bar{R}^2 , мы можем вовремя остановиться и перестать добавлять дополнительные степени свободы.

Вторая группа подходов — регуляризации, самые известные из которых Ridge (L_2 /гребневая/Тихоновская регуляризация), Lasso (L_1 регуляризация) и Elastic Net (Ridge+Lasso). Главная идея этих методов: модифицировать функцию потерь дополнительными слагаемыми, которые не позволят вектору коэффициентов \mathbf{w} неограниченно расти и тем самым воспрепятствуют переобучению

$$E_{\text{Ridge}}(\mathbf{w}) = \text{SSE}(\mathbf{w}) + \alpha \sum_i |w_i|^2 = \text{SSE}(\mathbf{w}) + \alpha \|\mathbf{w}\|_{L_2}^2,$$

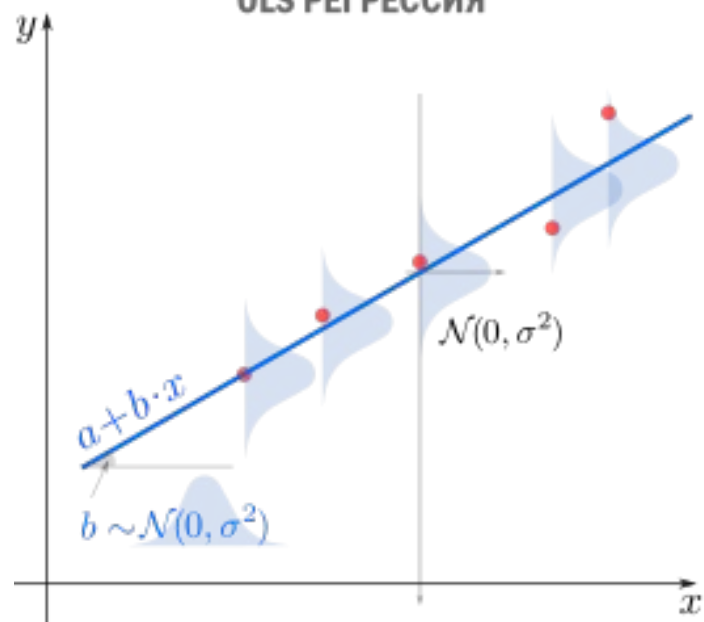
$$E_{\text{Lasso}}(\mathbf{w}) = \text{SSE}(\mathbf{w}) + \beta \sum_i |w_i| = \text{SSE}(\mathbf{w}) + \beta \|\mathbf{w}\|_{L_1},$$

$$E_{\text{EN}}(\mathbf{w}) = \text{SSE}(\mathbf{w}) + \alpha \|\mathbf{w}\|_{L_2}^2 + \beta \|\mathbf{w}\|_{L_1},$$

где α и β — параметры, которые регулируют «силу» регуляризации. Это обширная тема с красивой геометрией, которая заслуживает отдельного рассмотрения. Упомяну кстати, что для случая двух переменных при помощи вероятностной интерпретации можно получить Ridge и Lasso регрессии, удачно выбрав априорное распределения для коэффициента b

$$y = a + bx + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad \begin{cases} b \sim \mathcal{N}(0, \tau^2) & \leftarrow \text{Ridge}, \\ b \sim \text{Laplace}(0, \alpha) & \leftarrow \text{Lasso}. \end{cases}$$

OLS РЕГРЕССИЯ



LAD РЕГРЕССИЯ

