Лекция 2. Метод наименьших квадратов

1. Начать с примера из жизни – аппроксимация напорных QH-характеристик насосных агрегатов. В модели движения на работу была детерминированная составляющая – время собственно движения, и случайная составляющая – время ожидания. При этом детерминированная составляющая от эксперимента к эксперименту постоянна.
2. Если посмотреть на реальные точки, то видна их случайная и детерминированная составляющая. Для одного и того же расхода дифнапор от эксперимента к эксперименту скачет. (Буквально таких двух точек с одинаковым расходом нет, но если бы были, то вряд ли дифнапор был бы один и тот же). Несмотря на наличие случайности, дифнапор, однако, имеет тенденцию изменяться при изменении расхода.
3. Поэтому естественно записать следующую вероятностную модель, называемую регрессией.
$$H = \underbrace{\sum_{j=0}^{k-1} \beta_j \cdot Q^j}_{\substack{\text{детерминированная} \\ \text{составляющая}}} + \underbrace{\varepsilon}_{\substack{\text{случайная} \\ \text{составляюшая,} \\ \text{шум}}}$$
$$\varepsilon \sim N(0, \underbrace{D_\varepsilon}_{\substack{\text{дисперсия} \\ \text{шума}}})$$
4. Степени расхода $Q^0, Q^1, \dots, Q^{k-1}$ называются факторами. Перейдем к векторной записи. Введем вектор факторов *x* размерности k+1 и вектор коэффициентов.
$$x = \underbrace{\begin{pmatrix} Q^0 \equiv 1 \\ Q^1 \\ \dots \\ Q^{k-1} \end{pmatrix}}_{[k\ x\ 1]} \quad \beta_{\text{ист}} = \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{k-1} \end{pmatrix}}_{[k\ x\ 1]}$$
$$y = x^T \cdot \beta_{\text{ист}} + \varepsilon$$
5. Если $k = 0$, то модель становится похожа на модель разницы показаний датчиков:
$$H = \beta_{\text{ист},0} + \varepsilon$$
дифнапор имеет, вообще говоря, ненулевое постоянное значение $\beta_{\text{ист},0}$, не зависящее от расхода, и зашумленное случайным шумом ε
6. Если $k > 0$, то детерминированная составляющая H полиномиально зависит от Q и также зашумлена шумом ε.
7. В учебнике Лурье по Трубопроводному транспорту задана зависимость особого рода
$$H = a - bQ^2$$
Какие для нее будут x и β?
Тот же вопрос для зависимостей
$H = aQ$ прямая пропорциональность
$H = a + bQ$ линейная зависимость
8. Поскольку коэффициенты $\beta_{\text{ист}}$ задают детерминированную составляющую отклика y, то они разумеется детерминированные. Однако, на практике они неизвестны, их нужно оценить по реальным данным. Еще раз, вместо истинных β у нас будут только их оценки. Рассмотрим МНК-оценивание коэффициентов $\beta_{\text{ист}}$. Дана выборка объема n, $\{H_i, Q_i\}$, снятая с реального насосного агрегата. Сформируем по ней:
   a. отклики $y_i = H_i$
   b. вектор**ы** факторов $x_i^T = (1, Q^1, Q^2, \dots, Q^{k-1})$.
9. Пусть как-то наугад выбрано значение коэффициентов, в общем случае, не равное истинному, $\beta \neq \beta_{\text{ист}}$. Как оценить, насколько, это значение является удачным или неудачным среди всех других возможных значений? Другими словами, как оценить качество модели $x^T \cdot \beta$? Посмотреть, насколько фактическое значение y_i отличается от прогнозного $x_i^T \cdot \beta$ во всех элементах выборки
$$e_i = y_i - x_i^T \cdot \beta$$

10. Величину e_i будем называть остатками. Обратите внимание, что остаток и шум после измерений – вещи разные. Сумма квадратов остатков характеризует качество модели на имеющейся выборке:

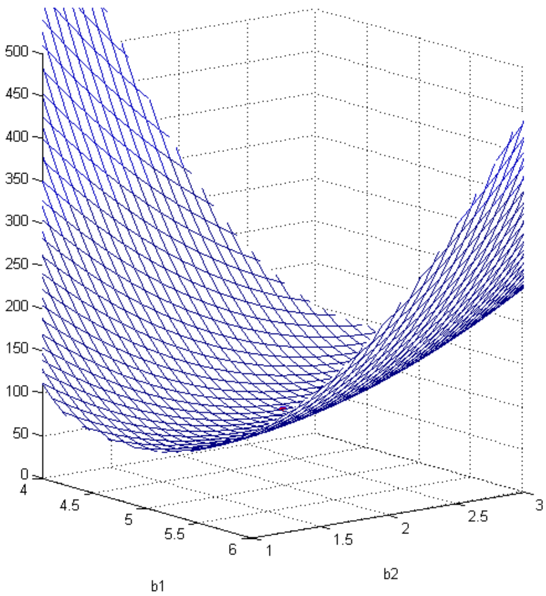
$$J(\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - x_i^T \cdot \beta\right)^2$$

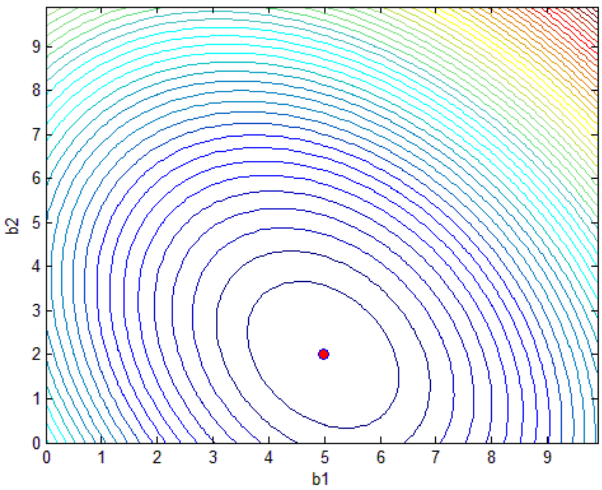
(функционал является скалярной величиной, хотя и зависит от векторного аргумента β)

11. Если варьировать β, то это будет соответствовать изменению вида графика $x^T \cdot \beta$ в осях Q-H. Рассмотрим пример для линейной модели.
12. В момент поиска β выборка фиксирована (после измерений), поэтому J зависит только компонентов вектора β. В частности для линейной аппроксимации нужны два коэффициента, функционал J будет функцией от двух аргументов. Для хороших значений β функционал $J(\beta)$ будет небольшим, для плохих – большим. Или по-другому, значение β тем лучше, чем меньше функционал $J(\beta)$.
13. Может ли $J(\beta)$ быть ниже нуля? Может ли $J(\beta)$ быть нулем?
14. Пример 1. Зависимость $H = aQ$,

$$J(\beta) = J(a) = \sum_{i=1}^{n} (H_i - Q_i \cdot a)^2$$

15. Как выбрать β? Если втупую – то перебрать много значений, построить поверхность в (k+1)-мерном пространстве и найти там минимум. Это очень тупо, зато понятно. Приведенный график имеет характер оврага. Но в середине оврага он еще глубже, там красная точка.





16. Как найти минимум по-умному. Нужно найти экстремум функции, т.е. точку, где все производные равны нулю.

$$\frac{\partial J(\beta_0, \beta_1, \dots, \beta_{k-1})}{\partial \beta_i} = 0, \qquad i = 0, \dots, k-1 \text{ (скалярная запись)}$$

17. Сделаем вывод в векторном виде. Для этого перепишем критерий МНК в векторном виде. Введем вектор фактических значений по всей выборке.

$$Y = \underbrace{\begin{pmatrix} y_1 \\ y_1 \\ \dots \\ y_n \end{pmatrix}}_{[n\ x\ 1]} \quad \beta_{\text{ист}} = \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{k-1} \end{pmatrix}}_{[k\ x\ 1]}$$

18. Введем матрицу всех факторов по всей выборке, матрицу X, называемую матрицей плана.

$$X = \underbrace{\begin{pmatrix} x_1^T \\ x_2^T \\ \dots \\ x_n^T \end{pmatrix}}_{[n\ x\ k]} = \begin{pmatrix} Q_1^0 & Q_1^1 & \dots & Q_1^{k-1} \\ Q_2^0 & Q_2^1 & \dots & Q_2^{k-1} \\ \dots & \dots & \dots & \dots \\ Q_n^0 & Q_n^1 & \dots & Q_n^{k-1} \end{pmatrix}$$

при произвольной β предсказания по всем элементам выборки будут:

$$X\beta = \underbrace{\begin{pmatrix} x_1^T\beta \\ x_2^T\beta \\ \dots \\ x_n^T\beta \end{pmatrix}}$$

19. Вектор всех остатков выборки можно рассчитать так:

$$\vec{e} = Y - X\beta$$

20. Сумма квадратов остатков, т.е. МНК-критерий примет вид:

$$J(\beta) = e^T e = (Y - X\beta)^T (Y - X\beta)$$

21. Требуется найти оптимальное значение функционала качества, доставляющее минимум критерию:

$$\beta$$

22. Условие экстремума состоит в том, что вектор градиент $grad\ J(\beta)$ равен нуль-вектору:

$$grad\ J(\beta) = \frac{\partial J(\beta)}{\partial \beta^T} = \underbrace{\left(\frac{\partial J}{\partial \beta_0}, \frac{\partial J}{\partial \beta_1}, \dots, \frac{\partial J}{\partial \beta_{k-1}}\right)}_{\text{вектор}-\text{градиент}} = 0 \text{ (векторная запись)}$$

23. В чем разница между $\beta_{\text{ист}}$, β и $\hat{\beta}$? Если все понимают, то делаем последний рывок.

24. Рассмотрим правила векторного дифференцирования

а. $\varphi(x) = a^T x$, если есть скаляр $\varphi(x)$, то при дифференцировании получаем

$$\frac{\partial \varphi(x)}{\partial x^T} = a^T$$

б. $\varphi(x) = x^T A x$

$$\frac{\partial \varphi}{\partial x^T} = \frac{\partial (x^T A x)}{\partial x^T} = x^T (A + A^T)$$

если матрица симметрическая $A = A^T$, то

$$\frac{\partial (x^T A x)}{\partial x^T} = 2x^T A$$

25. Раскроем скобки в сумме квадратов и посчитаем

$$J(\beta) = (Y - X\beta)^T (Y - X\beta) = Y^T Y - Y^T X\beta - X^T \beta^T Y + \beta^T X^T X\beta$$

26. Посчитаем производные каждого члена суммы

1) $\frac{\partial Q(y^T y)}{\partial \beta^T} = \vec{0}$

2) $\frac{\partial Q(y^T x\beta)}{\partial \beta^T} = y^T x$

3) $\frac{\partial Q(\beta^T x^T x\beta)}{\partial \beta^T} = \beta^T \left(x^T x + \left(x^T x \right)^T \right) = 2\beta^T x^T x$

$\left(x^T x\right)^T = x^T \left(x^T\right)^T = x^T x \Rightarrow x^T x -$ симетрическая

27. Итого,

$\frac{\partial Q(\beta)}{\partial \beta^T} = -2y^T x + 2\beta^T x^T x = \vec{0}$ (приравниваем производные к 0, чтобы найти min)

$\beta^T x^T x = y^T x$

$x^T x \beta = x^T y$

Размерность системы:

$$\underset{[k \times n]}{X^T}\underset{[n \times k]}{X}\underset{[k \times 1]}{\beta} = \underset{[k \times n]}{X^T}\underset{[n \times 1]}{Y}$$

28. Получили систему линейных уравнений относительно β. Ее решение доставит минимум функционалу качества $J(\beta)$. Эта система уравнений имеет специальное название система нормальных уравнений Умножим обе части слева на $(X^T X)^{-1}$, получим оптимальные оценки коэффициентов:

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