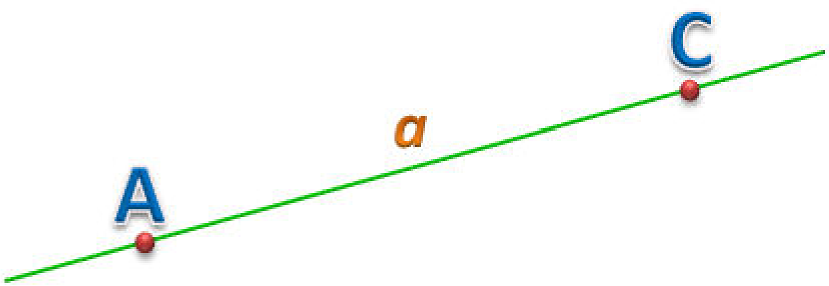
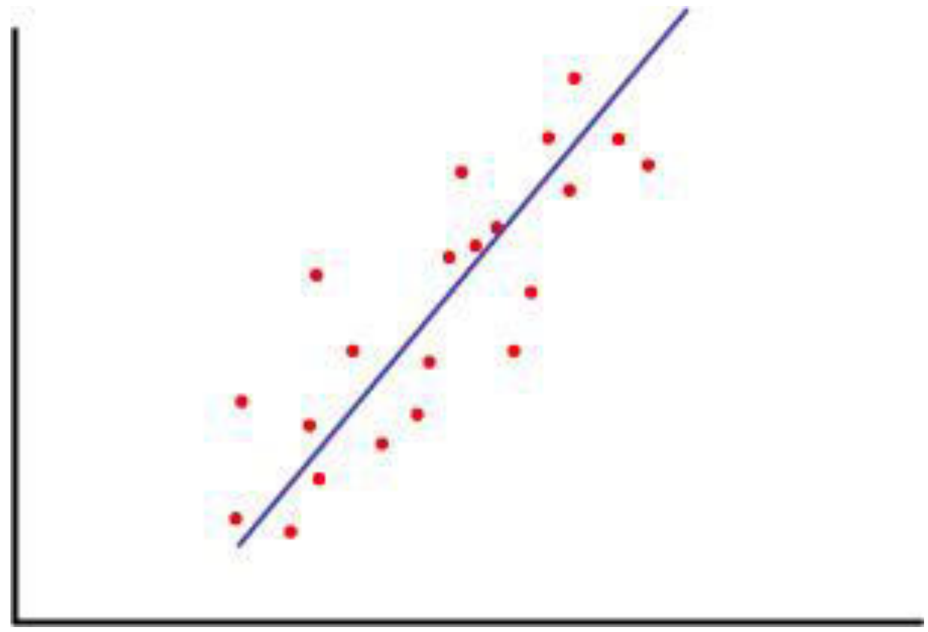
Занятие 3. Метод наименьших квадратов
Линейная регрессия

Через две точки на плоскости можно провести прямую и только одну



А если точек на плоскости – три и более?



Часть 1. Одномерная линейная регрессия

Метод наименьших квадратов

<u>Дано:</u>

1. **Набор экспериментальных точек** $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$
2. **Линейная модель** $y = a + bx$

Найти коэффициенты a **и** b

Переопределённая система уравнений

$$\begin{cases} a + bx_1 = y_1 \\ a + bx_2 = y_2 \\ \ldots \\ a + bx_n = y_n \end{cases}$$

В общем случае решения не имеет (т.к. экспериментальные точки обычно не ложатся в точности на одну прямую)

Необходимость в приближенных методах

Метод наименьших квадратов (МНК)

Минимизация суммы квадратов отклонений RSS (Resudiual Sum of Squares)

$$RSS = \sum_i (y_i - (a + bx_i))^2$$

Линейная регрессия: коэффициенты

Минимизируемая функция

$$RSS = \sum_i (y_i - (a + bx_i))^2$$

Результат расчёта

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} \end{cases}$$

Поиск стационарных точек для RSS

$$\begin{cases} \dfrac{\partial RSS}{\partial a} = \sum_i 2(y_i - a - bx_i) = 0 \\ \dfrac{\partial RSS}{\partial b} = \sum_i 2(y_i - a - bx_i)x_i = 0 \end{cases}$$

$$\begin{cases} \sum_i y_i - na - b\sum_i x_i = 0 \\ \sum_i x_i y_i - a\sum_i x_i - b\sum_i x_i^2 = 0 \end{cases}$$

$$\begin{cases} \bar{y} - a - b\bar{x} = 0 \\ \overline{xy} - a\bar{x} - b\overline{x^2} = 0 \end{cases}$$

$$\begin{cases} a = \bar{y} - b\bar{x} \\ \overline{xy} - (\bar{y} - b\bar{x})\bar{x} - b\overline{x^2} = 0 \end{cases}$$

$$\begin{cases} a = \bar{y} - b\bar{x} \\ \overline{xy} - \bar{x}\bar{y} + b\left[(\bar{x})^2 - \overline{x^2}\right] = 0 \end{cases}$$

Линейная регрессия: коэффициенты *r* и *R*²

Коэффициент детерминации R^2

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

$$RSS + ESS = TSS$$

Коэффициент корреляции Пирсона $r_{y,\hat{y}}$

$$r_{y,\hat{y}} = \frac{\sum_i (y_i - \bar{y})(\hat{y} - \bar{y})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{y})^2}} = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{TSS \cdot ESS}}$$

$$\hat{y}_i = a + bx_i$$

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

residual sum of squares (сумма квадратов отклонений)

$$TSS = \sum_i (y_i - \bar{y})^2$$

total sum of squares (общая сумма квадратов)

$$ESS = \sum_i (\hat{y}_i - \bar{y})^2$$

explained sum of squares (объяснённая сумма квадратов)

Связь между R^2 и $r_{y,\hat{y}}$

0 (т.к. МНК)

$$r_{y,\hat{y}} = \frac{\sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{TSS \cdot ESS}} = \frac{\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_i (\hat{y}_i - \bar{y})^2}{\sqrt{TSS \cdot ESS}} = \sqrt{\frac{ESS}{TSS}} = \sqrt{R^2}$$

Линейная регрессия: критерий Фишера (F-тест)

Шаг 1. Найти $F_{эмп}$

$$F_{emp} = \frac{R^2}{1 - R^2} \cdot \frac{f_2}{f_1}$$

f_2 - число степеней свободы для данных ($N - 2$ для $y = a + bx$)

f_1 - число независимых коэффициентов (1 для $y = a + bx$)

Откуда взята формула?

$$\frac{ESS/f_1}{RSS/f_2} = \frac{ESS/TSS}{RSS/TSS} \cdot \frac{f_2}{f_1} = \frac{R^2}{1 - R^2} \cdot \frac{f_2}{f_1}$$

Шаг 2. Сравнить с табличным значением квантиля

Если $F_{emp} \sim F(f_1; f_2)$, то зависимость статистически незначима

На практике:

Если $F_{emp} < F(\alpha; f_1; f_2)$

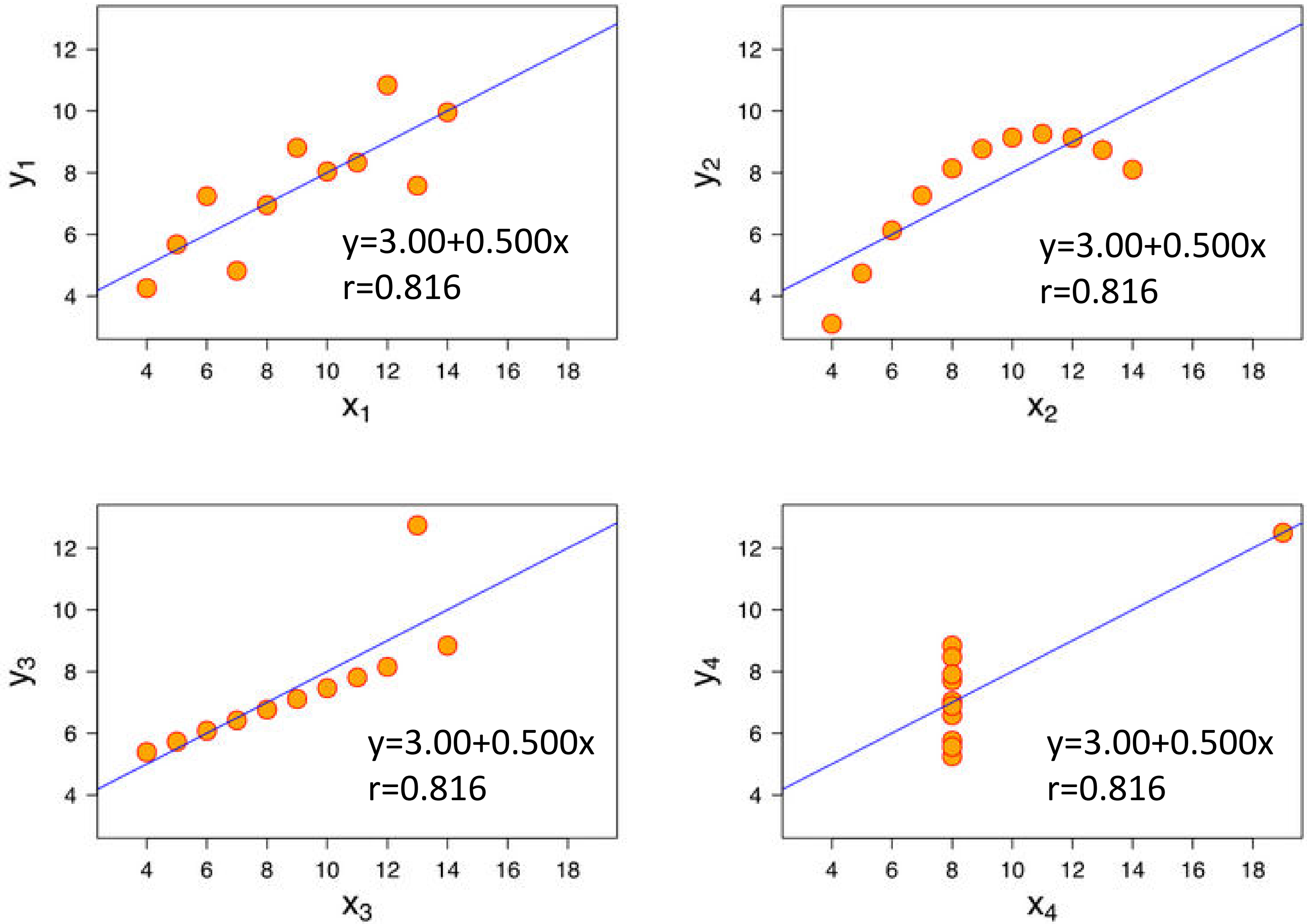
(т.е. сравнивают с табличным значением квантиля)

Пример: R^2=0.667, N=11, аппроксимация $y = a + bx$

$$F_{emp} = \frac{0.667}{0.333} \cdot \frac{9}{1} = 18 \qquad F(0.95,1,9) = 5.12$$

$\Rightarrow$ **Регрессия значима**

Квартет Энскомба (Anscombe's quartet)



Линейная регрессия в MS Excel

Способ 1. Линия тренда на графике

1. Построить точечный график по имеющимся данным вида (y_i, x_i)
2. Щелкнуть правой кнопкой мыши на серии точек и выбрать «добавить линию тренда»
3. Отметить флажками нужные опции (вид аппроксимирующей функции, показывать ли уравнение на диаграмме, показывать ли R^2 и т.п.)

Нагляден, но не проводятся F-тест и оценка доверительных интервалов коэффициентов регрессии

Способ 2. Использование пакета анализа данных

1. Выбрать вкладку данные, щелкнуть по пункту меню «анализ данных»
2. Из предлагаемых опций выбрать регрессию
3. Указать входные данные и изучить результат

Нагляден, содержит F-тест и оценку доверительных интервалов коэффициентов регрессии

Способ 3. Вручную

Использовать функции MS Excel вроде КОВАР, ДИСП, СРЗНАЧ, СУММ, КОРРЕЛ и т.п.

На практике способ не удобен, но полезен для понимания сути происходящего

Линейная регрессия: линеаризация

Если данные описываются нелинейной зависимостью, то в некоторых случаях её можно линеаризовать

Пример 1: $k_1 = k_0 exp\left(-\frac{E_a}{RT}\right)$ (уравнение Аррениуса)

Решение: $\ln k_1 = \ln k_0 - \frac{E_a}{RT}$ (т.е. вместо (k; T) – ($\ln k$; $1/T$))

Пример 2: $\Delta_{mix}H = x(1-x)(A+Bx)$ (энтальпия смешения)

Решение: $\frac{\Delta_{mix}H}{x(1-x)} = A + Bx$

Пример 3: $v = \frac{v_m S}{S+K_m}$ (схема Михаэлиса-Ментен)

Решение: $\frac{1}{v} = \frac{1}{v_m} + \frac{K_m}{v_m S}$

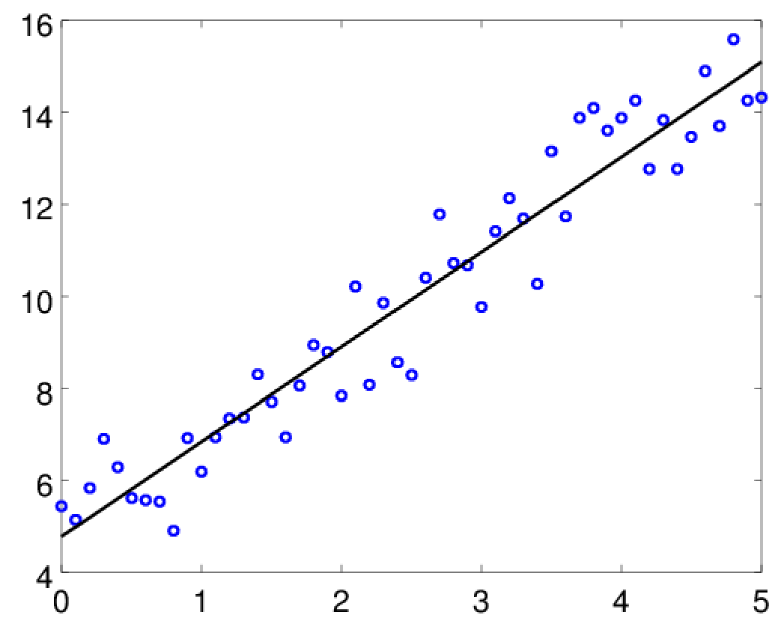
Простая линейная регрессия в GNU Octave

$$\begin{cases} ax_1 + b = y_1 \\ \dots \\ ax_n + b = y_n \end{cases} \Leftrightarrow \begin{pmatrix} x_1 & 1 \\ \dots & \dots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} \Leftrightarrow X\beta = y$$

Т.к. матрица X – не квадратная, то записать $\beta = X^{-1}y$ нельзя

Но Octave/MATLAB решит эту систему уравнений, если написать `b=X\y`

```
>> x = (0:0.1:5)';
>> y = 2*x + 5 + randn(size(x));
>> X = [x ones(size(x))];
>> beta = X\y
beta =
   2.0653
   4.7701
>> close all;
>> plot(x,y,'bo','LineWidth',2);
>> hold on;
>> yfunc = @(x)beta(1)*x+beta(2);
>> plot(x,yfunc(x),'k-','LineWidth',2);
>> hold off;
>> print(gcf,'graph','-dpng','-r75');
```



Часть 2. Многомерная линейная регрессия

Постановка задачи

Исходные данные
Точки в (k+1) – мерном пространстве
$(y_i, x_{i1}, \dots, x_{ik})$

+

Аппроксимирующая функция
$y = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$
β – параметры модели

Система уравнений (переопределённая):

$$\begin{cases} \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_k x_{1k} = y_1 \\ \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_k x_{2k} = y_2 \\ \dots \\ \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_k x_{nk} = y_n \end{cases}$$

Система в матричном виде:
$\boldsymbol{X\beta = y}$

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}; \; y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix};$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

Некоторые свойства матриц

Умножение и транспонирование

1. $(A+B)^\top = A^\top + B^\top$
2. $(AB)^\top = B^\top A^\top$
3. $(A^{-1})^\top = (A^\top)^{-1}$
4. $(AB)C = A(BC)$
5. $A(B+C) = AB + AC$
6. $(A+B)C = AC + BC$

След матрицы

След матрицы – сумма элементов её главной диагонали

$$\mathrm{tr}(A) = \sum_i a_{ii}$$

1. $\mathrm{tr}(\alpha A + \beta B) = \alpha \mathrm{tr}(A) + \beta \mathrm{tr}(B)$
2. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
3. $\mathrm{tr}(A^\top) = \mathrm{tr}(A)$
4. $\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$

Метод наименьших квадратов

Сумма квадратов отклонений

$$RSS = \sum_i e_i^2 = e^\top e =$$
$$= (y - X\beta)^\top (y - X\beta) =$$
$$= y^\top y - 2y^\top X\beta + \beta^\top X^\top X\beta$$

Поиск минимума

$$\frac{\partial RSS}{\partial \beta} = -2\frac{\partial (y^\top X\beta)}{\partial \beta} + \frac{\partial (\beta^\top X^\top X\beta)}{\partial \beta} = 0$$
$$\frac{\partial RSS}{\partial \beta} = -2X^\top y + 2X^\top X\beta = 0$$
$$X^\top X\beta = X^\top y \Leftrightarrow \hat{\beta} = (X^\top X)^{-1} X^\top y$$

Дифференцирование

$$\frac{\partial (y^\top X\beta)}{\partial \beta_k} = \frac{\partial (\sum_i P_i \beta_i)}{\partial \beta_k} = P_k \Rightarrow \frac{\partial (y^\top X\beta)}{\partial \beta} = (y^\top X)^\top = X^\top y, \text{ где } P_i = (y^\top X)_i$$

$$\beta^\top X^\top X\beta = (X\beta)^\top (X\beta) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} x_{ij}\beta_j \right)^2 \Rightarrow \frac{\partial (\beta^\top X^\top X\beta)}{\partial \beta_k} = 2\sum_{i=1}^{n} x_{ik} \sum_{j=1}^{m} x_{ij}\beta_j$$

$$2\sum_{i=1}^{n} x_{ik} \sum_{j=1}^{m} x_{ij}\beta_j = 2\big((X\beta)^\top X\big)^\top = 2X^\top X\beta$$

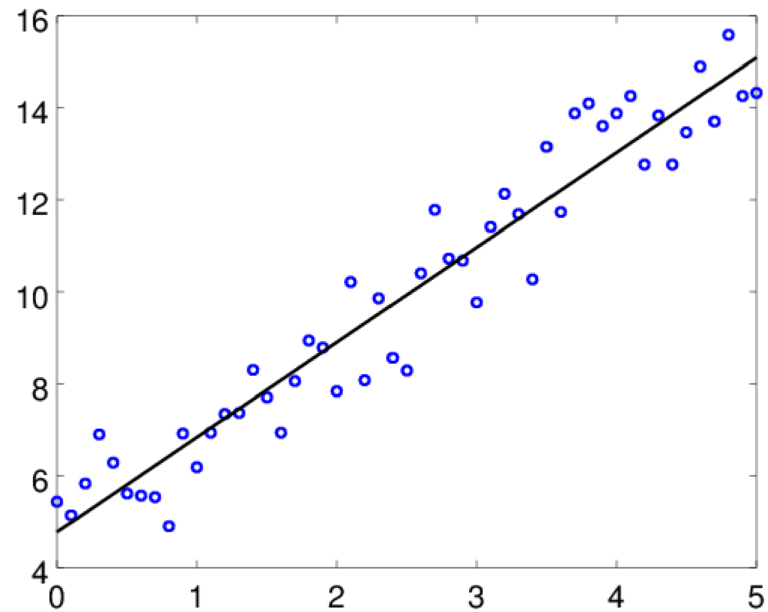
Метод наименьших квадратов: одномерный случай

$$\begin{cases} ax_1 + b = y_1 \\ \dots \\ ax_n + b = y_n \end{cases} \Leftrightarrow \begin{pmatrix} x_1 & 1 \\ \dots & \dots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} \Leftrightarrow X\beta = y$$

Т.к. матрица X – не квадратная, то записать $\beta = X^{-1}y$ нельзя

Но Octave/MATLAB решит эту систему уравнений, если написать `b=X\y`

```
>> x = (0:0.1:5)';
>> y = 2*x + 5 + randn(size(x));
>> X = [x ones(size(x))];
>> beta = X\y
beta =
   2.0653
   4.7701
>> close all;
>> plot(x,y,'bo','LineWidth',2);
>> hold on;
>> yfunc = @(x)beta(1)*x+beta(2);
>> plot(x,yfunc(x),'k-','LineWidth',2);
>> hold off;
>> print(gcf,'graph','-dpng','-r75');
```



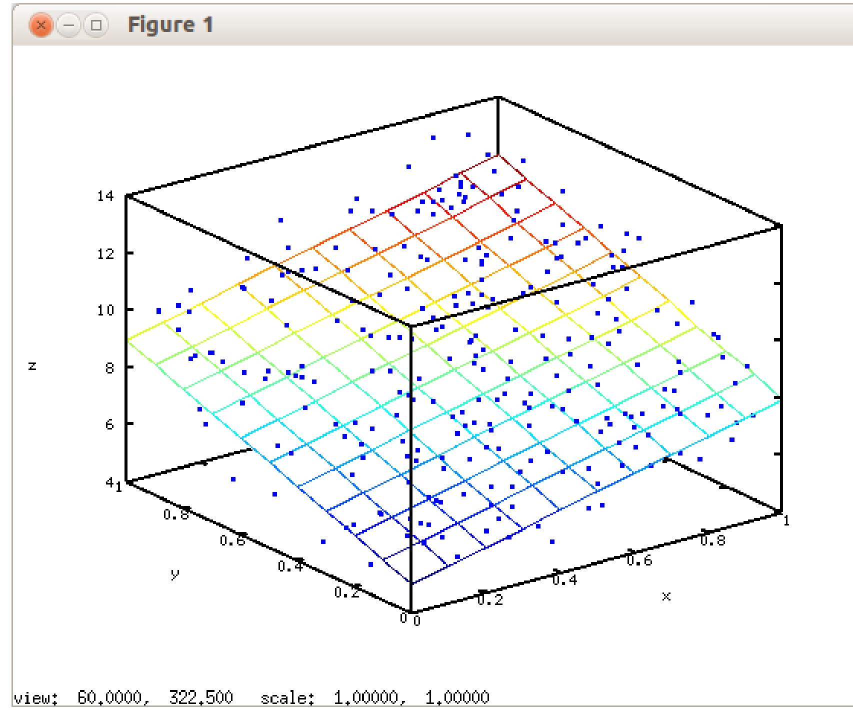
Задача: нахождение коэффициентов регрессии

Шаг 1. Создать выборку точек

```
x = rand(500, 1);
y = rand(500, 1);
z = 3*x+4*y+5+randn(size(x));
plot3(x,y,z,'bo');
```

Шаг 2. Записать и решить систему уравнений

```
X = [x y ones(size(x))];
b = (X'*X)\(X'*z);
format long;
disp(b);
xv = 0:0.1:1;
[Xm,Ym]=meshgrid(xv,xv);
Zm = b(1)*Xm + b(2)*Ym + b(3);
hold on;
mesh(Xm,Ym,Zm); hold off;
```




```
octave:5> X = [x y ones(size(x))];
octave:6> b = (X'*X)\(X'*z);
octave:7> format long;
octave:8> disp(b);
   2.97216535047754
   3.96977787608039
   4.99030081951429
octave:9> xv = 0:0.1:1;
```

Теорема Гаусса-Маркова

Пусть выполняются следующие условия:

1. Модель правильно специфицирована
2. $rang(X) = m$, где m – число коэффициентов регрессии
3. $E[\varepsilon_i] = 0$ (нулевое матожидание ошибок регрессии)
4. $E\left[\varepsilon_i \varepsilon_j\right] = E[\varepsilon_i] E\left[\varepsilon_j\right] = 0$ (независимость ошибок друг от друга)
5. $Var[\varepsilon_i] = E[\varepsilon_i \varepsilon_i] = \sigma^2$ (гомоскедастичность ошибок регрессии)

Тогда оценки параметров регрессии методом наименьших квадратов являются наилучшими в классе линейных несмещённых оценок (англ. Best Linear Unbiased Estimator, BLUE).

Ковариационная матрица

Несмещённость оценок параметров регрессии

$$E\left[\hat{\beta}\right] = E[(X^\top X)^{-1}X^\top(XB+\varepsilon)] = E[B] + (X^\top X)^{-1}X^\top E[\varepsilon] = E[B]$$

B – истинное значение параметров регрессии, ε – вектор ошибок

Ковариационная матрица

$$\mathrm{cov}\left(\hat{\beta},\hat{\beta}\right) = E\left[\left(\hat{\beta}-B\right)\left(\hat{\beta}-B\right)^\top\right] = E[(X^\top X)^{-1}X^\top\varepsilon\varepsilon^\top X(X^\top X)^{-1}] =$$
$$= (X^\top X)^{-1}X^\top E[\varepsilon\varepsilon^\top]X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}$$

B – истинное значение параметров регрессии, ε – вектор ошибок

Вид ковариационной матрицы

$$\mathrm{cov}\left(\hat{\beta},\hat{\beta}\right) = \begin{pmatrix} \mathrm{Var}[\hat{\beta}_1] & \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_2) & \cdots & \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_m) \\ \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_2) & \mathrm{Var}[\hat{\beta}_2] & \cdots & \mathrm{cov}(\hat{\beta}_2,\hat{\beta}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\hat{\beta}_m,\hat{\beta}_1) & \mathrm{cov}(\hat{\beta}_m,\hat{\beta}_2) & \cdots & \mathrm{Var}[\hat{\beta}_m] \end{pmatrix}$$

Доверительные интервалы

$$s^2_{\beta_i} = \mathrm{Var}[\hat{\beta}]$$
$$\Delta\hat{\beta}_i = s_{\beta_i}\cdot t(\alpha, f)$$

t – двухсторонний квантиль t-распределения;

α – вероятность, $f = n - m$ – число степеней свободы

Оценка ошибки регрессии

Несмещённая оценка ошибки регрессии

$$\hat{\sigma}^2 = \frac{1}{n-m}\sum_i e_i^2 = \frac{e^\top e}{n-m}$$

$$e = y - \hat{y}$$

n – число точек, m – число коэффициентов регрессии

Проекционная матрица

$$\hat{y} = Hy; H = X(X^\top X)^{-1}X^\top$$

Свойства

1. $H^\top = H$ (симметричность)
2. $H^2 = H$ (идемпотентность)
3. $HX = X$

Связь погрешности с проекционной матрицей

$$e = y - \hat{y} = (I - H)y = M(XB + \varepsilon) = M\varepsilon; RSS = e^\top e = (M\varepsilon)^\top(M\varepsilon) = \varepsilon^\top M\varepsilon$$

$$e^\top e = \mathrm{tr}(e^\top e) = \mathrm{tr}(\varepsilon^\top M\varepsilon) = \mathrm{tr}(M\varepsilon\varepsilon^\top) \Rightarrow E[RSS] = tr(ME[\varepsilon\varepsilon^\top]) = \sigma^2\mathrm{tr}(M)$$

Вычисление следа проекционной матрицы

$$tr(M) = tr(I_n) - tr[X(X^\top X)^{-1}X^\top] = n - tr[(X^\top X)(X^\top X)^{-1}] = n - m$$

Задача: доверительные интервалы значений $\widehat{\beta}$

Шаг 1. Ошибка регрессии

```
>> res = z-(b(1)*x+b(2)*y+b(3));
>> f = numel(res) - numel(b);
>> sigma2 = res'*res/f
  sigma2 =  0.80841663065686
```

Шаг 2. Ковариационная матрица

```
>> format short;
>> C = sigma2 * inv(X'*X)
C =
   0.0204893 -0.0013650 -0.0096530
  -0.0013650  0.0188808 -0.0085331
  -0.0096530 -0.0085331  0.0106459
```

Шаг 3. Ошибки и доверительные интервалы коэффициентов

```
>> sb = sqrt(diag(C));disp(sb');
   0.14314   0.13741   0.10318
>> db = sb * tinv(1-0.05/2,f);
>> disp(db');
   0.28124   0.26997   0.20272
```

Шаг 4. Корреляционная матрица

```
>> sbm = [sb sb sb];
>> r = C./(sbm.*sbm')
r =
  1.000000 -0.069400 -0.653593
 -0.069400  1.000000 -0.601874
 -0.653593 -0.601874  1.000000
```

Внимание! $r\left(\widehat{\beta}_i, \widehat{\beta}_j\right) \neq 0$

Оставляйте «запасные» знаки при округлении $\widehat{\beta}$!

Шаг 4. R^2 и F-критерий

```
>> TSS = sum((z-mean(z)).^2)
TSS =  1497.7
>> RSS = res'*res;
RSS =  401.78
>> R2 = 1 - RSS/TSS;
R2 =  0.73173
>> F = R2/(1-R2)*f/2
F =  677.80
>> finv(0.95,2,f)
ans =  3.0139
```

Линеаризация многомерной нелинейной регрессии

1. Нелинейная зависимость

$$c_p(T) = a + bT + \frac{c}{T}$$

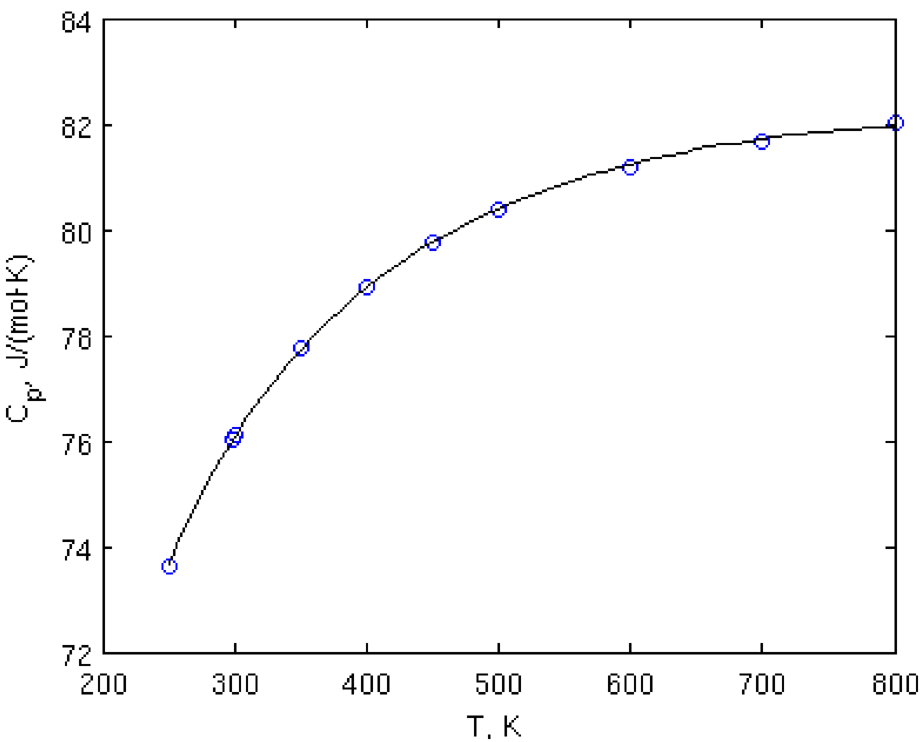
2. Линеаризация

$$c_p(T) = ax_1 + bx_2 + cx_3;$$
$$x_1 = 1;\ x_2 = T;\ x_3 = T^{-1}$$

3. Система уравнений

$$X\beta = y$$

$$X = \begin{pmatrix} 1 & T_1 & 1/T_1 \\ \vdots & \ddots & \vdots \\ 1 & T_n & 1/T_n \end{pmatrix}; \beta = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$



Доверительный интервал и интервал предсказания

Доверительный интервал $\hat{y}$ (confidence interval)

$$\hat{y} \pm \hat{\sigma} t_{\alpha,n-m} \sqrt{x^\top (X^\top X)^{-1} x}$$

Исходная функция с вероятностью 95% проходит через этот интервал

Интервал предсказания (prediction interval)

$$\hat{y} \pm \hat{\sigma} t_{\alpha,n-m} \sqrt{1 + x^\top (X^\top X)^{-1} x}$$

Новая точка попадёт в этот интервал с вероятностью 95%