Учёт систематической погрешности при регрессионном анализе

Краткое содержание

1. Традиционный метод наименьших квадратов
2. Методы Ньютона, Гаусса-Ньютона, Левенберга-Марквардта
3. Учёт систематических ошибок и метод максимального правдоподобия
4. Практическая реализация метода наименьшего правдоподобия в среде MATLAB

Rudnyi E.B. Statistical model of systematic errors: linear error model // Chemometrics and Intelligent Laboratory Systems. 1996. V. 34. P. 41-54.

Метод наименьших квадратов

Метод наименьших квадратов: постановка задачи

Исходные данные

1. Набор точек $(x_{1i}, \dots, x_{mi}, y_i)$, где x – независимые величины, y – зависимая величина
2. Аппроксимирующая функция $\varphi(\vec{\beta}, \vec{x})$, где $\vec{\boldsymbol{\beta}}$ - параметры модели

Система уравнений (переопределённая)

$$\varphi(\vec{\beta}, \vec{x_i}) = y_i$$

или (с учётом ошибок)

$$\varphi(\vec{\beta}, \vec{x_i}) + \varepsilon_i = y_i$$

Минимизируемая сумма квадратов отклонений

$$F(\vec{\beta}) = \sum_i (\hat{y}_i - y_i)^2 = \sum_i \left(\varphi(\vec{\beta}, \vec{x}_i) - y_i\right)^2 = \sum_i e_i^2$$

Система уравнений для нахождения минимума целевой функции

$$\frac{\partial F(\vec{\beta})}{\partial \beta_i} = 0$$

МНК: поиск коэффициентов регрессии

Система уравнений для нахождения минимума целевой функции

$$\frac{\partial F(\vec{\beta})}{\partial \beta_i} = 0$$

Линейная регрессия

1. Сумма квадратов отклонений

$$RSS = \sum_i e_i^2 = e^\top e =$$

$$= (y - X\beta)^\top (y - X\beta) =$$

$$= y^\top y - 2y^\top X\beta + \beta^\top X^\top X\beta$$

2. Оценка параметров

$$X^\top X\beta = X^\top y \Leftrightarrow \hat{\beta} = (X^\top X)^{-1} X^\top y$$

3. Стандартная ошибка параметров

$$s_{\hat{\boldsymbol{\beta}}}^2 = \hat{\sigma}^2 (X^\top X)^{-1}$$

Нелинейная регрессия

1. Оценка параметров

Численные методы (Гаусса-Ньютона, Левенберга-Марквардта)

2. Доверительные интервалы

$$s_{\hat{\boldsymbol{\beta}}}^2 = \hat{\sigma}^2 (J^\top J)^{-1}$$

J – якобиан вида $J_{ij} = \frac{\partial \varphi(\vec{\beta}, \vec{x_i})}{\partial \beta_j}$

Якобиан и отклонения рассчитываются в точке $\hat{\boldsymbol{\beta}}$

Численные методы решения нелинейных уравнений

Метод Ньютона

Одномерный случай (нелинейное уравнения)

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) = 0$$

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Многомерный случай (система нелинейных уравнений)

$$F(\vec{\beta}) \approx F(\vec{\beta}_0) + \sum_i \frac{\partial F(\vec{\beta}_0)}{\partial \beta_i}\left(\beta_i - \beta_i^{(0)}\right) = 0 \Rightarrow F(\vec{\beta}_0) + \nabla F(\vec{\beta}_0)\vec{p} = 0$$

Особенности метода Ньютона

1. Быстро сходится (квадратичная сходимость)
2. В случае нескольких корней может быть найден любой из них (зависит от начального приближения)
3. Скорость сходимости (и сама сходимость) может зависеть от начального приближения

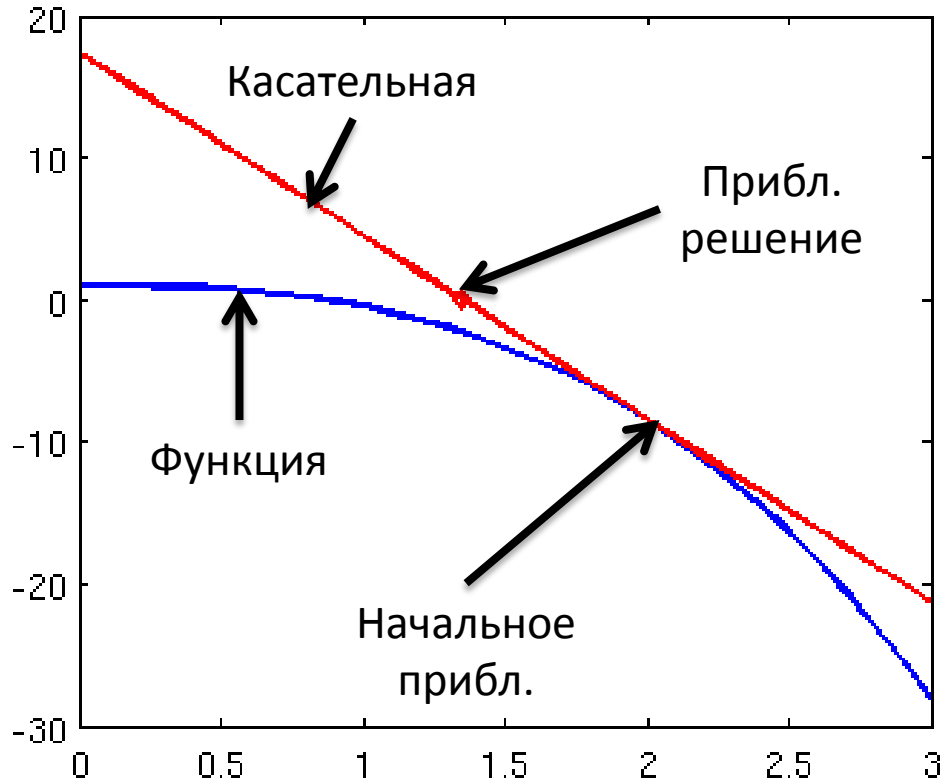
Метод Ньютона: практическая реализация

Уравнение

$$f(x) = \cos x - x^3 = 0$$
$$f'(x) = -\sin x - 3x^2 = 0$$

Программа

```
format long;
delta = 1e-15;
x0 = 0.5;
x = x0 + 2*delta;
while abs(x-x0) >= delta
    x0 = x;
    f = cos(x0)-x0^3;
    df = -sin(x0)-3*x0^2;
    x = x0 - f/df;
    disp(x);
end
```




```
2.000000000000000
1.348055393079852
1.001262924310922
0.880591138294078
0.865691456412747
0.865474078978736
0.865474033101617
0.865474033101614
```

МНК и метод Гаусса-Ньютона

Система уравнений и метод Ньютона

$$\begin{cases} \dfrac{\partial F(\vec{\beta})}{\partial \beta_1} = 0 \\ \cdots \\ \dfrac{\partial F(\vec{\beta})}{\partial \beta_m} = 0 \end{cases} \Rightarrow \frac{\partial F(\vec{\beta})}{\partial \beta_i} \approx \frac{\partial F(\vec{\beta}^{(0)})}{\partial \beta_i} + \sum_j \frac{\partial F(\vec{\beta}^{(0)})}{\partial \beta_i \partial \beta_j}\left(\beta_j - \beta_j^{(0)}\right) = 0$$

Матричная запись и метод Гаусса-Ньютона

$$\nabla F(\vec{\beta}^{(0)}) + H(\vec{\beta}^{(0)})\vec{p} = 0 \Rightarrow J^\top(\vec{\beta}^{(0)}) f(\vec{\beta}^{(0)}) + J^\top(\vec{\beta}^{(0)}) J(\vec{\beta}^{(0)})\vec{p} = 0$$

Градиент, якобиан и гессиан

$$(\nabla F)_i = \frac{\partial F}{\partial \beta_i} = \frac{\partial}{\partial \beta_i}\left(\sum_j f_j^2\right) = 2\sum_j f_j \frac{\partial f_j}{\partial \beta_i} = 2J^\top f; J_{ij} = \frac{\partial f_i}{\partial \beta_j} = \frac{\partial \varphi(\vec{\beta}, \vec{x_i})}{\partial \beta_j}$$

$$H_{ij} = \frac{\partial F(\beta)}{\partial \beta_i \partial \beta_j} = \frac{\partial}{\partial \beta_j}\left[2\sum_k f_k \frac{\partial f_k}{\partial \beta_i}\right] = 2\sum_k \left[\frac{\partial f_k}{\partial \beta_i}\frac{\partial f_k}{\partial \beta_j} + f_k \frac{\partial f_k}{\partial \beta_i \partial \beta_j}\right] \approx 2\sum_k \frac{\partial f_k}{\partial \beta_i}\frac{\partial f_k}{\partial \beta_j} = 2J^\top J$$

Метод Левенберга-Марквардта

$$(J^\top J + \lambda)\vec{p} = -\nabla F = -J^\top f$$

Каждая итерация включает в себя подбор шага λ:

1. Взять начальное (очень малое) значение λ
2. Выполнить итерацию, найдя $\vec{p}$
3. Если значение F возросло, то увеличить шаг λ и вернуться к п.1
4. Если значение F уменьшилось, то принять полученное значение $\vec{p}$ и перейти к следующей итерации

Сходится медленнее метода Гаусса-Ньютона, но менее требователен к начальному приближению. Широко применяется на практике.

Метод Л.-М. – комбинация методов Гаусса-Ньютона и градиентного спуска

а) метод градиентного спуска: $\vec{p} = -\lambda \nabla F\left(\vec{\beta}^{(0)}\right)$

б) метод Гаусса-Ньютона: $J^\top J\vec{p} = -\nabla F$

Обозначения: $\vec{p} = \vec{\beta} - \vec{\beta}^{(0)}$, $\vec{\beta}^{(0)}$ – начальное приближение; $\vec{\beta}$ – результат итерации,λ – шаг, F – минимизируемая функция

Систематические погрешности и метод максимального правдоподобия

Систематические погрешности: постановка задачи

Условия задачи

1. Имеются наборы экспериментальных точек вида (x_{ij}, y_{ij}) из нескольких разных источников (i - номер источника, j – номер точки)
2. Суммарная погрешность (ε_{ij}) для каждой точки складывается из случайной ($\varepsilon_{r,ij}$) и двух видов систематических: сдвига ($\varepsilon_{a,i}$) и наклона ($\varepsilon_{b,i}$), т.е. $\varepsilon_{ij} = \varepsilon_{r,ij} + \varepsilon_{a,i} + \varepsilon_{b,i}(x_{ij} - x_i)$, где $x_i = \sum_j \frac{x_{ij}}{N_i}$
3. Дисперсии для разных видов ошибок и разных выборок могут быть не равны между собой

Требуется найти

1. Параметры модели $\vec{\boldsymbol{\beta}}$
2. Дисперсии по выборкам, связанные со случайными ошибками ($\sigma_{r,i}^2$)
3. Дисперсии по выборкам, связанные со систематическими ошибками, т.е. $\sigma_{a,i}^2$ и $\sigma_{b,i}^2$ (в виде, нормированном на $\sigma_{r,i}^2$, т.е. $\gamma_{a,i} = \sigma_{a,i}^2/\sigma_{r,i}^2$; $\gamma_{b,i} = \sigma_{b,i}^2/\sigma_{r,i}^2$)

<u>Из-за наличия σ в параметрах классический МНК неприменим!</u>

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).

Метод максимального правдоподобия

Основан на поиске максимума функции вида

$$L = -\ln\{\det[D(\varepsilon)]\} - \varepsilon^{\top} D(\varepsilon)^{-1}\varepsilon$$

ε – вектор отклонений

$D(\varepsilon) = cov(\varepsilon, \varepsilon)$ – дисперсионная матрица

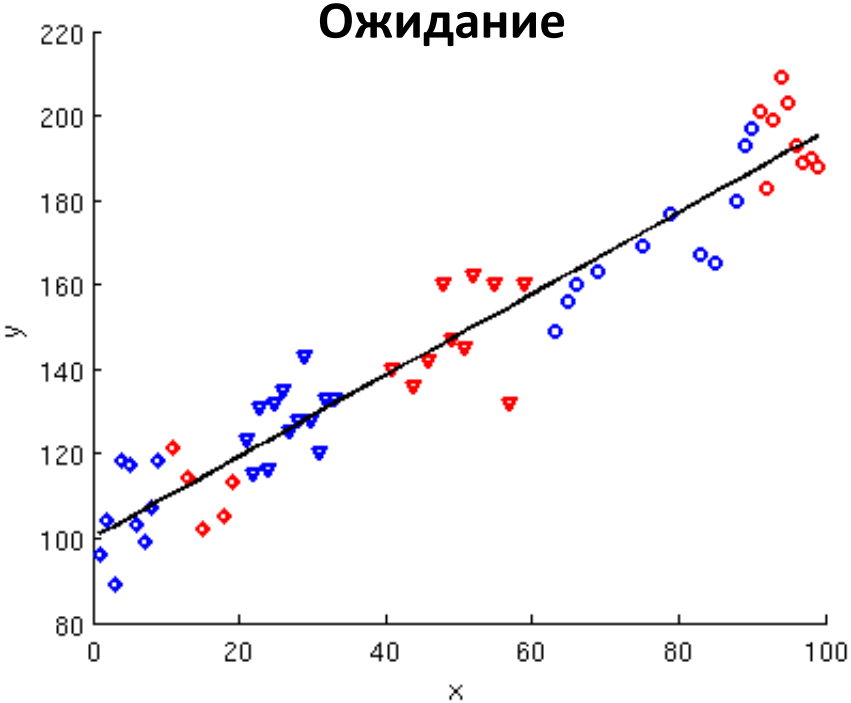
Доверительный интервал $\vec{\beta}$

$$\hat{\sigma}_{\beta}^{2} = [J^{\top} D(\varepsilon)^{-1} J]^{-1}$$

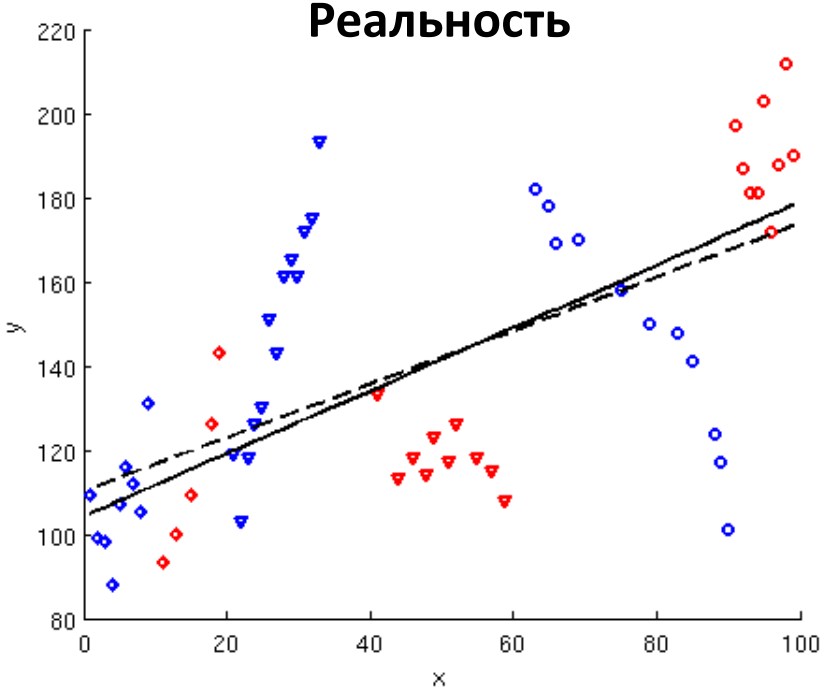
J – якобиан вида $J_{ij} = \dfrac{\partial\varphi(\vec{\beta}, \vec{x_i})}{\partial\beta_j}$

(каждая строка соотв. эксп. точке, столбец – параметру регрессии)

Ожидание



Реальность



М.м.п.: дисперсионная матрица

Общая погрешность для отдельной точки

$$\varepsilon_{ij} = \varepsilon_{r,ij} + \varepsilon_{a,i} + \varepsilon_{b,i}\left(x_{ij} - x_i\right) = \varepsilon_{r,ij} + \varepsilon_{a,i} + \varepsilon_{b,i}X_{ij}$$

Дисперсионная матрица: вывод

$$D(\varepsilon) = cov\left(\varepsilon_{ij}, \varepsilon_{kl}\right) = M\left[\left(\varepsilon_{r,ij} + \varepsilon_{a,i} + \varepsilon_{b,i}X_{ij}\right) \cdot \left(\varepsilon_{r,kl} + \varepsilon_{a,k} + \varepsilon_{b,k}X_{kl}\right)\right]$$

$$= M\left[\varepsilon_{r,ij}\varepsilon_{r,kl}\right] + \cancel{M\left[\varepsilon_{r,ij}\right]M\left[\varepsilon_{a,k}\right]} + \cancel{M\left[\varepsilon_{r,ij}\right]M\left[\varepsilon_{b,k}\right]}X_{kl} +$$

$$+\cancel{M\left[\varepsilon_{a,i}\right]M\left[\varepsilon_{r,kl}\right]} + M\left[\varepsilon_{a,i}\varepsilon_{a,k}\right] + \cancel{M\left[\varepsilon_{a,i}\right]M\left[\varepsilon_{b,k}\right]}X_{kl} +$$

$$+\cancel{M\left[\varepsilon_{b,i}\right]M\left[\varepsilon_{r,kl}\right]} + \cancel{M\left[\varepsilon_{b,i}\right]M\left[\varepsilon_{a,k}\right]} + M\left[\varepsilon_{b,i}\varepsilon_{b,k}\right]X_{ij}X_{kl} =$$

$$= M\left[\varepsilon_{r,ij}\varepsilon_{r,kl}\right] + M\left[\varepsilon_{a,i}\varepsilon_{a,k}\right] + M\left[\varepsilon_{b,i}\varepsilon_{b,k}\right]X_{ij}X_{kl}$$

M – математическое ожидание

cov - ковариация

М.м.п.: дисперсионная матрица
(блочно-диагональный вид)

$$D(\varepsilon) = \sigma_{r,i_i}^2 \delta_{ij,kl} + \delta_{ik}\left(\sigma_{a,i}^2 + \sigma_{b,i}^2 X_{ij} X_{kl}\right) \qquad D(\varepsilon) = diag\{V_i\}$$

$$V_i = \sigma_{r,i}^2 \boldsymbol{I_i} + \sigma_{a,i}^2 \mathbf{1}_i \mathbf{1}_i^\top + \sigma_{b,i}^2 \boldsymbol{X}_i \boldsymbol{X}_i^\top = \sigma_{r,i}^2 \left(\boldsymbol{I_i} + \gamma_{a,i} \mathbf{1}_i \mathbf{1}_i^\top + \gamma_{b,i} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$$

$$V_i = \sigma_{r,i}^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} + \sigma_{a,i}^2 \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \ddots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} + \sigma_{b,i}^2 \begin{pmatrix} X_{i1}X_{i1} & X_{i1}X_{i2} & \cdots & X_{i1}X_{iN_i} \\ X_{i1}X_{i2} & X_{i2}X_{i2} & \ddots & X_{i2}X_{iN_i} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{i2}X_{iN_i} & \cdots & X_{iN_i}X_{iN_i} \end{pmatrix}$$

Как в обычном МНК

Полезные выражения

$$D(\varepsilon)^{-1} = diag\{V_i^{-1}\}$$

$$V_i^{-1} = \sigma_{r,i}^{-2} \left\{ \boldsymbol{I}_i - \frac{\gamma_{a,i}}{1 + N_i \gamma_{a,i}} \mathbf{1}_i \mathbf{1}_i^\top - \frac{\gamma_{b,i}}{1 + N_i \gamma_{b,i}} \boldsymbol{X}_i \boldsymbol{X}_i^\top \right\}$$

$$\det V_i = \left(\sigma_{r,i}^2\right)^{N_i} \left(1 + N_i \gamma_{a,i}\right)\left(1 + P_i \gamma_{b,i}\right)$$

Части целевой функции

$$SS = \sum_i \varepsilon_i^\top V_i^{-1} \varepsilon_i$$

$$\ln\{\det D(\varepsilon)\} = \sum_i \ln\{\det V_i\}$$

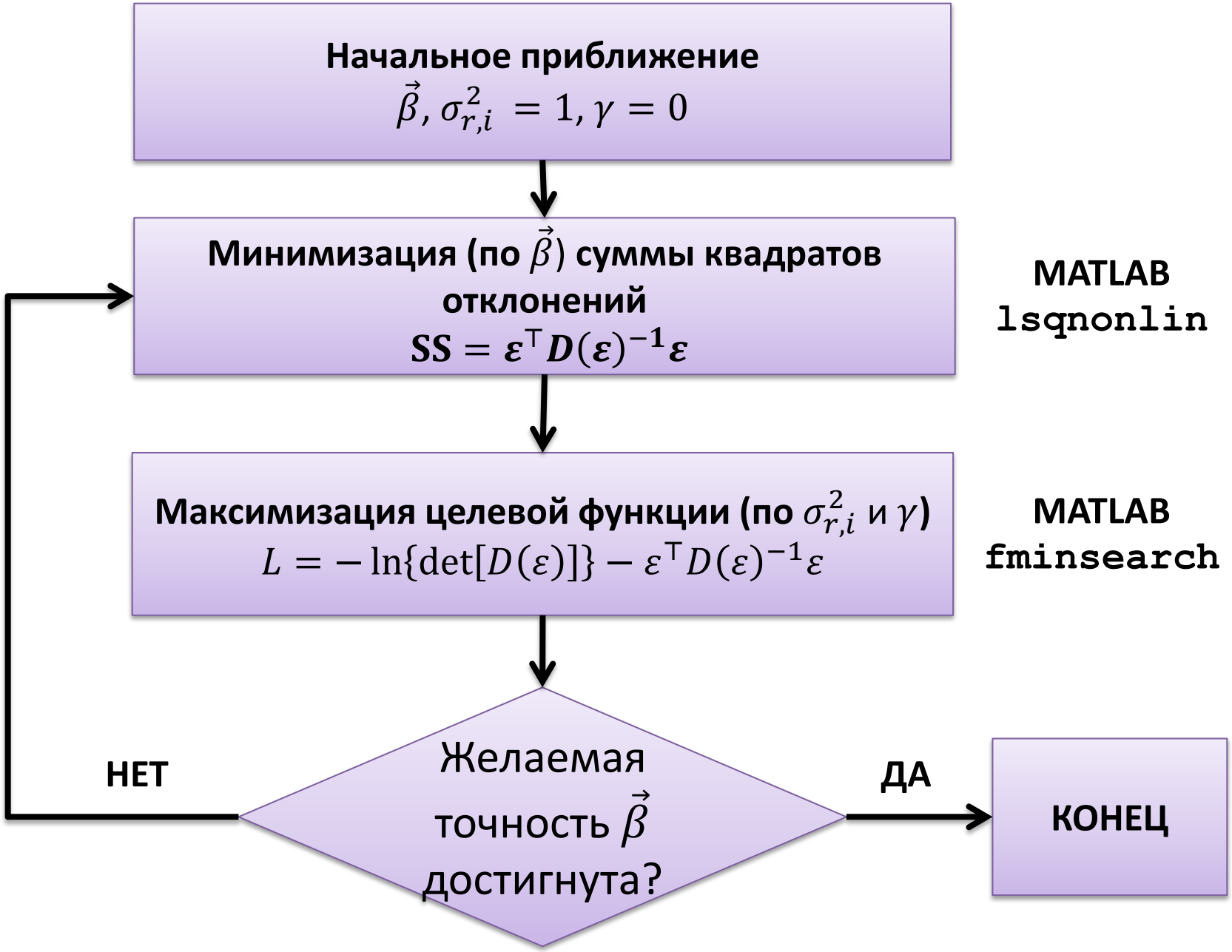
Практическая реализация метода максимального правдоподобия

Проблемы существующей реализации (Rudnyi-1996)

1. Отсутствует интерфейс пользователя, для задания новой модели требуется перекомпиляция программы
2. Зависимость от коммерческих библиотек
3. Не соответствует стандартам C++03 и C++11 (не компилируется современными компиляторами без модификации)

Rudnyi E.B. Statistical model of systematic errors: linear error model // Chemometrics and Intelligent Laboratory Systems. 1996. V. 34. P. 41-54.

Практическая реализация алгоритма



Тестовый пример

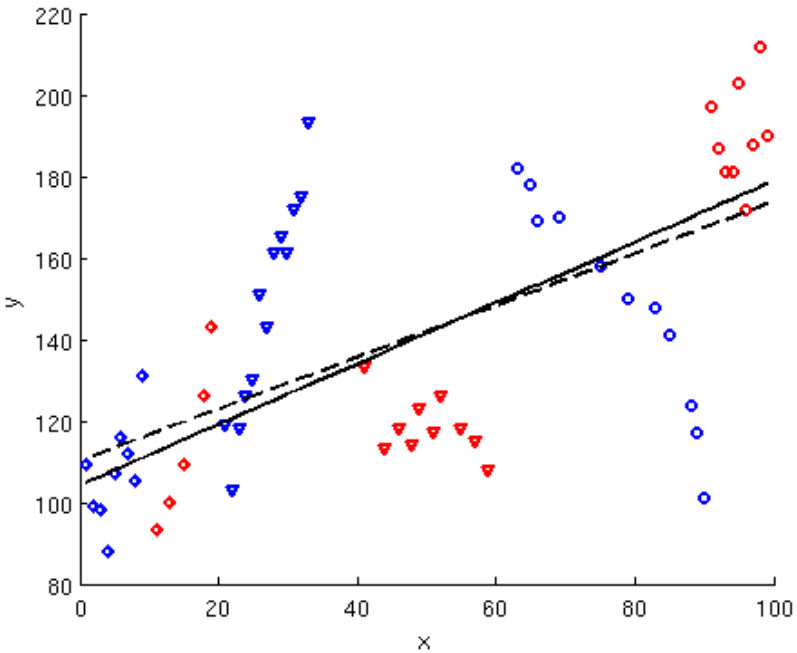
Наборы коэффициентов

Table 2
The initial parameter and variance values for the linear regression simulation and the results obtained by processing the pseudo-experimental values

	a	σ(a)	b	σ(b)	σ_r	$\sqrt{\gamma_a}$	$\sqrt{\gamma_b}$
a							
True	100.0	2.35	1.000	0.041	10.0	0.00	0.00
ML	100.0	2.09	0.963	0.037	8.9	0.00	0.00
OLS	100.0	2.09	0.963	0.037	8.9	0.00	0.00
b							
True	100.0	4.20	1.000	0.074	10.0	0.50	0.05
ML	102.6	2.67	0.959	0.047	9.3	0.23	0.04
OLS	102.5	2.30	0.965	0.040	9.8	0.00	0.00
c							
True	100.0	7.25	1.000	0.129	10.0	1.00	0.10
ML	107.4	5.73	0.873	0.102	9.4	0.80	0.18
OLS	104.3	3.172	0.920	0.056	13.5	0.00	0.00
d							
True	100.0	13.79	1.000	0.245	10.0	2.00	0.20
ML	104.2	10.92	0.747	0.196	8.9	1.73	0.39
OLS	110.2	6.06	0.635	0.106	25.8	0.00	0.00

+ электронное приложение к статье

Исходные данные



Взяты с сайта Е.Б. Рудного из программы `varcomp` (см. каталог `line/dat`): http://evgenii.rudnyi.ru/programming.html

Rudnyi E.B. Statistical model of systematic errors: linear error model // Chemometrics and Intelligent Laboratory Systems. 1996. V. 34. P. 41-54.

Тестовый пример

```
----- Maximum likehood method
ITERATION
beta = [110.23657 0.63540];
sigma_r=8.76383; sqrt(gamma_a)=1.07746; sqrt(gamma_b)=0.00000
ITERATION
beta = [121.02151 0.37552];
sigma_r=8.95076; sqrt(gamma_a)=2.13538; sqrt(gamma_b)=0.40311
ITERATION
beta = [103.70764 0.75683];
sigma_r=8.94423; sqrt(gamma_a)=1.73219; sqrt(gamma_b)=0.38604
ITERATION
beta = [104.15557 0.74705];
sigma_r=8.94436; sqrt(gamma_a)=1.72988; sqrt(gamma_b)=0.38642
ITERATION
beta = [104.16047 0.74694];
sigma_r=8.94436; sqrt(gamma_a)=1.72986; sqrt(gamma_b)=0.38643
ITERATION
beta = [104.16052 0.74694];
sigma_r=8.94436; sqrt(gamma_a)=1.72986; sqrt(gamma_b)=0.38643
ITERATION
beta = [104.16052 0.74694];
sigma_r=8.94436; sqrt(gamma_a)=1.72986; sqrt(gamma_b)=0.38643
>>>>> ML method final result:
>>>>> beta = [104.16052 0.74694]; sbeta = [10.92025 0.19552]
----- Classic LSQ method
beta = [110.23657 0.63540]; sbeta = [6.06017 0.10638]; sigma=25.814
```

Планы по развитию метода и его реализации

1. Написание универсальной реализации, пригодной для использования как в MATLAB, так и в GNU Octave
2. Исследовать возможность частичной или полной линеаризации проблемы максимизации целевой функции в отдельных случаях (линейная регрессия и т.п.)
3. Создание автономной программы с пользовательским интерфейсом (возможно, с использованием Ruby или Lua)