Сравнительный анализ алгоритмов оценивания параметров многомерной линейной регрессии на модельных и эконометрических примерах

Кондратьев Роман Сергеевич, 522-я группа

Санкт-Петербургский Государственный Университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель — к.ф.-м.н., доцент **А.Н. Пепелышев**
Рецензент — к.ф.-м.н., доцент **Н.П. Алексеева**

Санкт-Петербург
2007г.

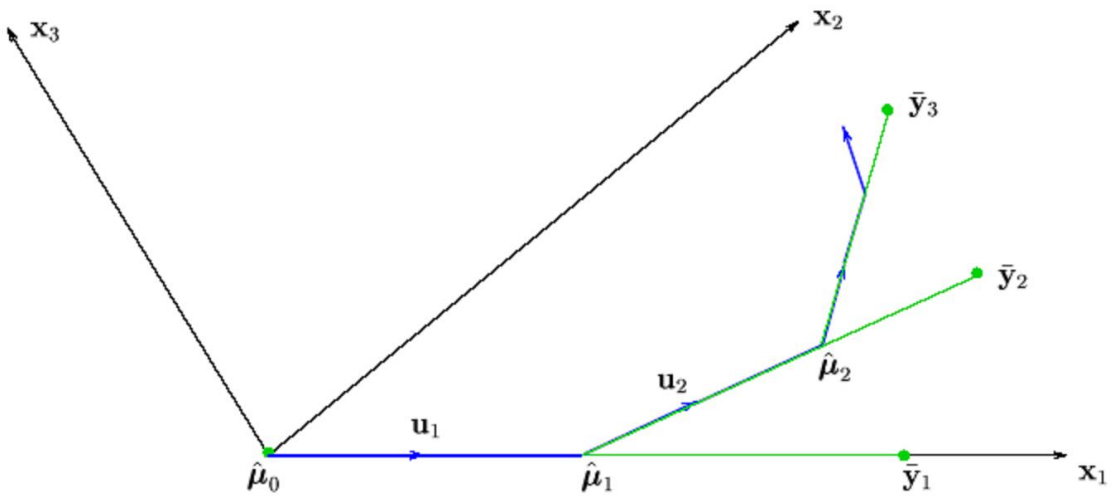
Постановка задачи

- Выборка (Y, X)
  Y — вектор значений зависимого признака
  X — матрица значений независимых признаков

- Уравнение многомерной линейной регрессии
  - Векторная запись
    $$Y = X_0\beta + \varepsilon$$
  - Покомпонентная запись
    $$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_s x_{si} + \varepsilon_i, \qquad i = 1, \ldots, m$$
  - Оценка параметров
    $$S(\beta) = \| Y - \mu \|^2 = \sum_{i=1}^{m} (y_i - \mu_i)^2 \to \min_{\beta}, \ \mu = X_0\beta, \ X_0 = (\mathbf{1}_{m\times 1} \vdots X)$$

- Ограничение
  $$\sum_{i=1}^{s} |\beta_i| < t$$

- Обычная оценка МНК
- LARS (Efron, Hastie, Tibshirani, 2002)
  1. Положить β=0
  2. Найти регрессор x_j имеющий наибольшую корреляцию с y
  3. Увеличить коэффициент β_j в направление знака корреляции с y
  4. Взять остатки $r = y - \widehat{y}$
  5. Остановиться, если какой-нибудь другой признак x_k имеет такую же корреляцию с r, как и x_j
  6. Увеличивать (β_j, β_k) в их совместном направлении, пока еще какой-нибудь признак x_l не станет так же коррелировать с r
  7. Делать шаги 2-6 пока все признаки не включены в модель

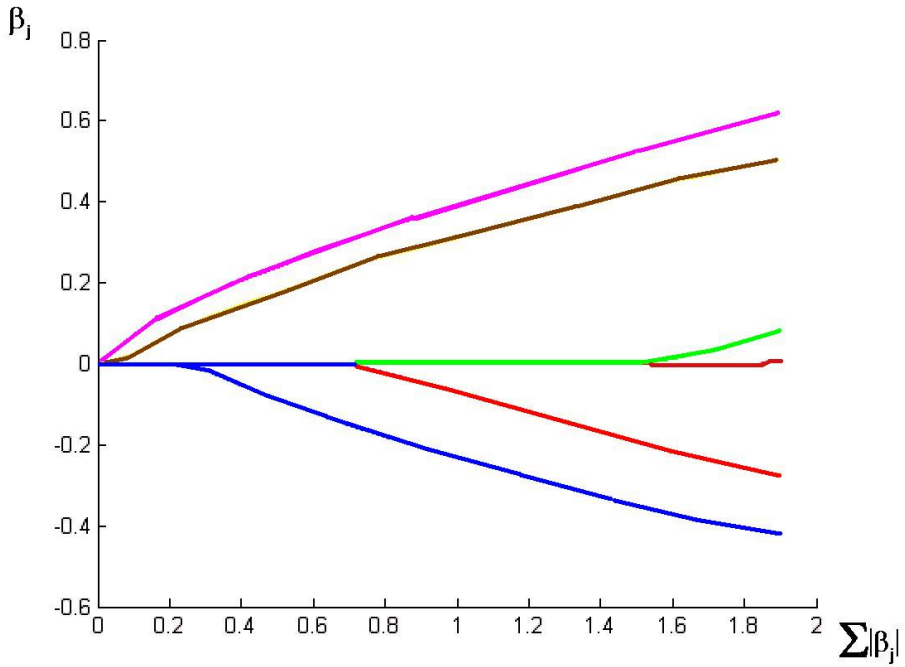


Дополнительные алгоритмы

- LinProg — обычный МНК с ограничением
- StageWise — ε шаги в направлении, как у LARS
- Lasso — шаги от обычного МНК к нулю

- DLARS
- StageWise-LARS
- Lasso-LARS

Рост оценок параметров по алгоритму Lasso



- Нахождение минимума по t средне-квадратичной ошибки ME для алгоритмов, зависящих от параметра t

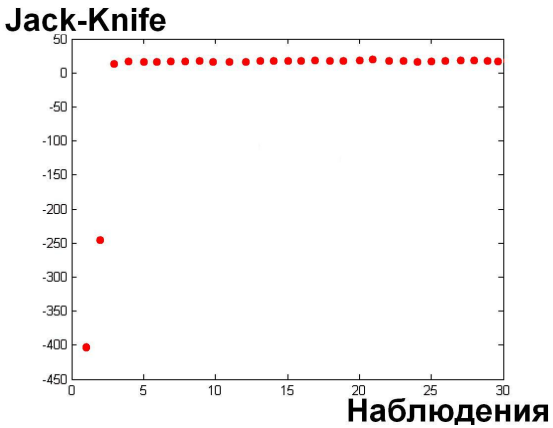
$$\mathrm{ME} = \mathrm{E}\{\widehat{\eta}(X) - \eta(X)\}^2$$

- Нахождение минимума по k статистики риска выбора модели C_p для пошаговых алгоритмов, где k — номер шага

$$C_p(\widehat{\mu}) = \frac{\|y - \widehat{\mu}\|^2}{\sigma^2} - s + 2k$$

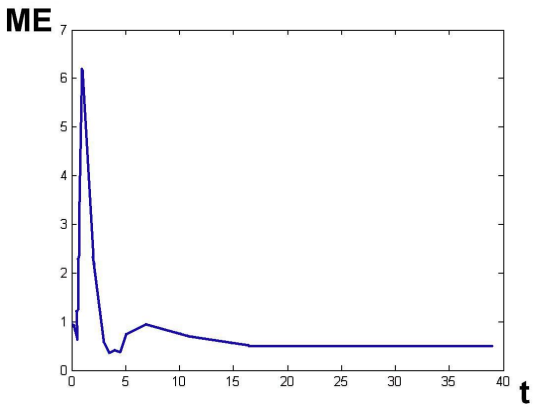
Метод вычисления оптимального параметра алгоритма

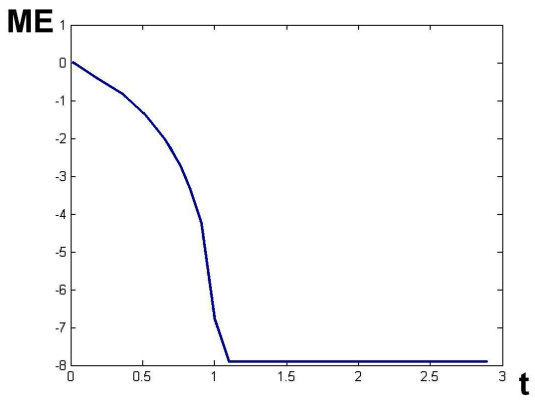
- Предварительное удаление выбросов посредством вычисления средне-квадратичной ошибки ME методом Jack-Knife



- Вычисление $ME = ME(t)$ методом Cross-Validation

$$ME = \mathsf{E}\{\widehat{\eta}(X) - \eta(X)\}^2$$





Пример 1. Модельные данные

- Истинное значение β

$$\beta = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10),\ \sigma^2 = 0.25$$

- Корреляционная матрица признаков

$$\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{|s-1|} \\ \rho & 1 & \rho & \dots & \rho^{|s-2|} \\ \rho^2 & \rho & 1 & \dots & \rho^{|s-3|} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho^{|s-1|} & \rho^{|s-2|} & \rho^{|s-3|} & \dots & 1 \end{pmatrix}$$

- Полученные Оценки
  - Оценка методом МНК, $\rho = 0.99$

$$\widehat{\beta}_{OLS} = (1.47\ \ -0.30\ \ 3.61\ \ 4.69\ \ 6.71\ \ 7.96\ \ 8.96\ \ 8.24\ \ 9.56\ \ 11.46)$$

  - Оценка методом $DLARS$, $\rho = 0.99$

$$\widehat{\beta}_{DLARS} = (0\ \ 0\ \ 0\ \ 4.24\ \ 6.43\ \ 4.54\ \ 7.62\ \ 6.61\ \ 9.88\ \ 8.21)$$
$$ME_{DLARS} = 0.0180,\ ME_{Lasso} = 0.0184$$

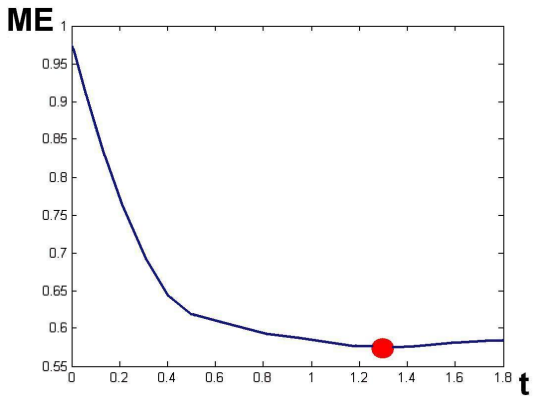
  - Оценка методом $DLARS$, $\rho = 0$

$$\widehat{\beta}_{DLARS} = (0\ \ 1.15\ \ 2.50\ \ 3.66\ \ 4.34\ \ 5.50\ \ 6.66\ \ 7.71\ \ 8.94\ \ 9.85)$$
$$ME_{DLARS} = 0.0169,\ ME_{Lasso} = 0.0031$$

- Результаты алгоритмов
  - не сильно зависят от корреляции
  - сильно зависит от выбросов

Пример 2. Реальные данные satisfaction-ratings.sta

- Семь признаков: производительность ПК, надежность, набор опций, выбор ПО, поддержка, гарантия, и цена
- Результаты
  - Кривая ME



  - Оценка методом Lasso

$$\widehat{\beta}_{\text{Lasso}} = (0 \;\; -0.1551 \;\; 0.7136 \;\; -0.0837 \;\; 0 \;\; -0.0966)$$

  - Оценка методом DLARS

$$\widehat{\beta}_{\text{LARS}} = (0 \;\; -0.2205 \;\; 0.7808 \;\; -0.1059 \;\; 0 \;\; -0.0846)$$

Пример 3. Эластичность

- Эластичность

$$E_D = \left| \frac{\%\text{Изменение количества спроса на продукт X}}{\%\text{Изменение цены на продукт X}} \right| = \frac{\frac{\Delta Q_D}{Q_D}}{\frac{\Delta P_D}{P_D}}$$

$$E_D = \frac{P}{Q} \times \frac{\partial Q}{\partial P}$$

- Модель эластичности для группы товаров

$$\Delta Q_i^j = \alpha_1 \Delta P_i^1 + \alpha_2 \Delta P_i^2 + \ldots + \alpha_s \Delta P_i^s$$

$$\Delta P_i^j = P_{i+1}^j - P_i^j, \; \Delta Q_i^j = Q_{i+1}^j - Q_i^j$$

- Модель

$$Q_i^j = Q_{i-1}^j + \beta^j (P_i^j - P_{i-1}^j) + \sum_{k \neq j} (P_i^k - P_{i-1}^k) + \varepsilon^j$$

$$j = 1, \ldots, s,\ i = 2, \ldots, m,\ Q_1^j = const,\ P_1^j = const$$

- Истинное значение β^j

- Таблица оценок методом Lasso для всех 6 признаков

	β_1	β_2	β_3	β_4	β_5	β_6
Q_1	-4.46	1.12	0.66	-0.30	1.56	1.37
Q_2	1.99	-5.22	-0.62	0.58	1.24	1.83
Q_3	0.83	1.06	-5.23	1.11	0.78	1.12
Q_4	-0.03	1.13	0	-6.3	-2.87	0
Q_5	0.47	0.88	0.72	1.17	-4.66	1.05
Q_6	0.78	1.29	0.25	1.8	1.18	-7.24

Пример 3. Эластичность. Реальные данные

Шесть наименований шампуней

Результаты

- Оценка МНК

$$\widehat{\beta}_{\text{OLS}} = (\text{-1.47} \;\; 0.12 \;\; 0.36 \;\; \text{-8.30} \;\; \text{-0.56} \;\; \text{-1.37})$$

- Метод Lasso

$$\widehat{\beta}_{\text{Lasso}} = (\text{-1.34} \;\; 0 \;\; 0.13 \;\; \text{-7.07} \;\; \text{-0.06} \;\; \text{-1.16})$$

- Метод DLARS

$$\widehat{\beta}_{\text{DLARS}} = (\text{-1.31} \;\; 0 \;\; 0 \;\; \text{-6.45} \;\; 0 \;\; \text{-1})$$

Заключение

- Изучены алгоритмы оценки параметров, дающие более удовлетворительные результаты в присутствии коррелированности признаков, чем обычный МНК
- Выявлена слабая зависимость от коррелированности данных и сильная зависимость от выбросов
- Исследована модель эластичности спроса по цене
- Разработаны программы в среде MATLAB