Прикладная статистика 9. Регрессионный анализ.

1 ноября 2013 г.

Variance-bias tradeoff

Модель:

$$y = F(x) + \varepsilon, \;\; \mathbb{E}\varepsilon = 0, \; \mathbb{D}\varepsilon = \sigma^2.$$

Ошибка предсказания y_0 по вектору x_0 с помощью модели $\hat{F}$:

$$PE\left(\hat{F}(x_0)\right) = \mathbb{E}\left(y_0 - \hat{F}(x_0)\right)^2 = \sigma^2 + MSE\left(\hat{F}(x_0)\right).$$

Среднеквадратичная ошибка оценки $\hat{F}$:

$$\begin{aligned} MSE\left(\hat{F}(x_0)\right) &= \mathbb{E}\left(F(x_0) - \hat{F}(x_0)\right)^2 = \\ &= \left(\mathbb{E}\left(F(x_0)\right) - F(x_0)\right)^2 + \mathbb{E}\left(\hat{F}(x_0) - \mathbb{E}\hat{F}(x_0)\right)^2 = \\ &= Bias^2\left(\hat{F}(x_0)\right) + Variance\left(\hat{F}(x_0)\right). \end{aligned}$$

Variance-bias tradeoff

В линейной регрессии:

$$y = x\theta + \varepsilon,$$
$$MSE\left(x_0\hat{\theta}\right) = Bias^2\left(x_0^T\hat{\theta}\right) + Variance\left(x_0^T\hat{\theta}\right).$$

МНК-оценка

$$\hat{\theta}^{OLS} = \left(X^TX\right)^{-1}X^Ty$$

является несмещённой $(Bias = 0)$ и имеет наименьшую дисперсию среди всех несмещённых оценок $(Variance = \sigma^2\left(X^TX\right)^{-1})$.

Если матрица X^TX плохо обусловлена, то:

- МНК-оценка имеет большую дисперсию;
- может возникать численная неустойчивость при обращении X^TX.

Гребневая регрессия

Для уменьшения дисперсии оценки и повышения вычислительной устойчивости добавим к $X^T X$ диагональную матрицу:

$$\hat{\theta}^{ridge} = \left(X^T X + \lambda I_n\right)^{-1} X^T y.$$

Такая оценка является решением регуляризованной задачи наименьших квадратов:

$$\hat{\theta}^{ridge} = \underset{\theta}{\text{argmin}} \left(\|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2\right).$$

λ — параметр регуляризации:

- при $\lambda = 0$ $\hat{\theta}^{ridge} = \hat{\theta}^{OLS}$, смещения нет;
- при $\lambda = \infty$ $\hat{\theta}^{ridge} = 0$, дисперсии нет;
- в промежутке — баланс между смещением и дисперсией.

Важные детали

- Коэффициент θ_0 не входит в регуляризатор:

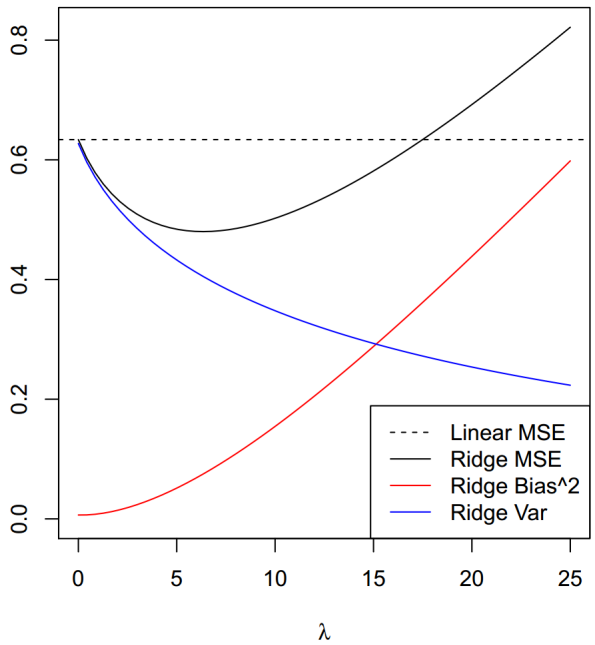
$$\left(\hat{\theta}_0, \hat{\theta}^{ridge}\right) = \underset{\theta_0 \in \mathbb{R}, \theta \in \mathbb{R}^k}{\text{argmin}} \left(\|y - \theta_0 I_n - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right).$$

  Если перед применением метода центрировать все признаки X (вычесть выборочное среднее), то можно положить $\hat{\theta}_0 = \bar{y}$ и исключить θ_0 из задачи минимизации.
  Если центрировать ещё и y, то $\hat{\theta}_0 = 0$.

- Штраф $\|\theta\|_2^2 = \sum_{j=1}^{k} \theta_j^2$ неравномерно распределяется между признаками, если они измерены в разных шкалах. Поэтому перед применением метода признаки X дополнительно стандартизируют, чтобы выборочная дисперсия каждого равнялась единице.
- После построения модели необходимо привести её к записи в терминах исходных признаков.

Пример 1

Пусть $n = 50,\ \ k = 30,\ \ X_{ij} \sim N(0,1),\ \ \sigma^2 = 1.$
Сгенерируем y согласно линейной модели с 10 большими коэффициентами (между 0.5 и 1) и 20 маленькими (между 0 и 0.3).



Линейная регрессия:

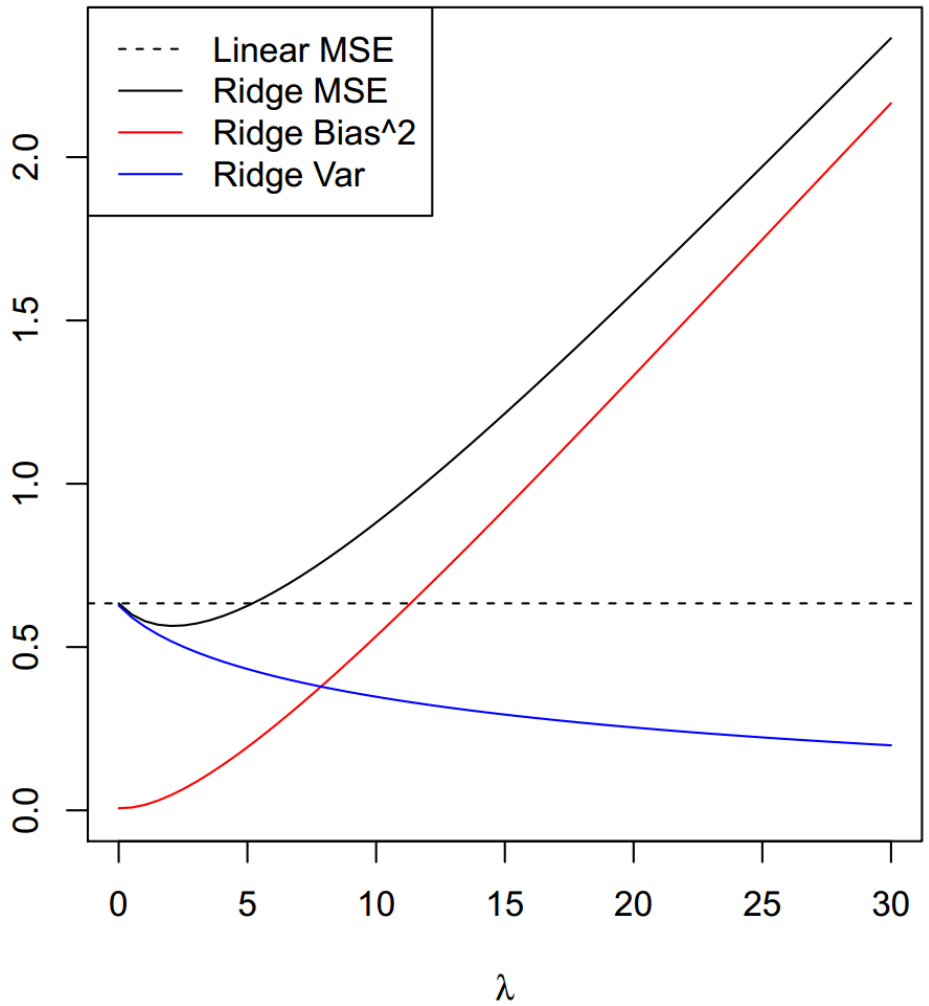
$$PE = 1 + MSE \approx 1 + 0.006 + 0.627 = 1.633.$$

Лучшая гребневая регрессия:

$$PE = 1 + MSE \approx 1 + 0.077 + 0.403 = 1.48.$$

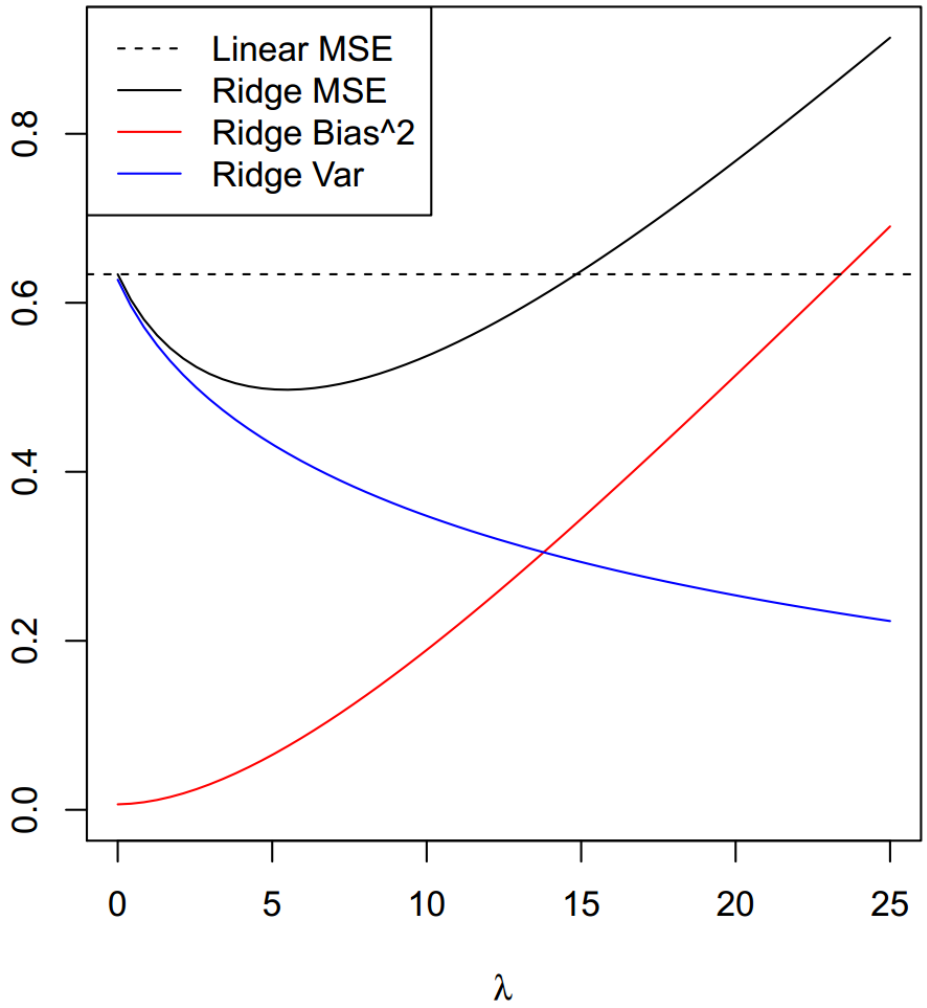
Пример 2

Сгенерируем y согласно линейной модели с 30 большими коэффициентами (между 0.5 и 1).

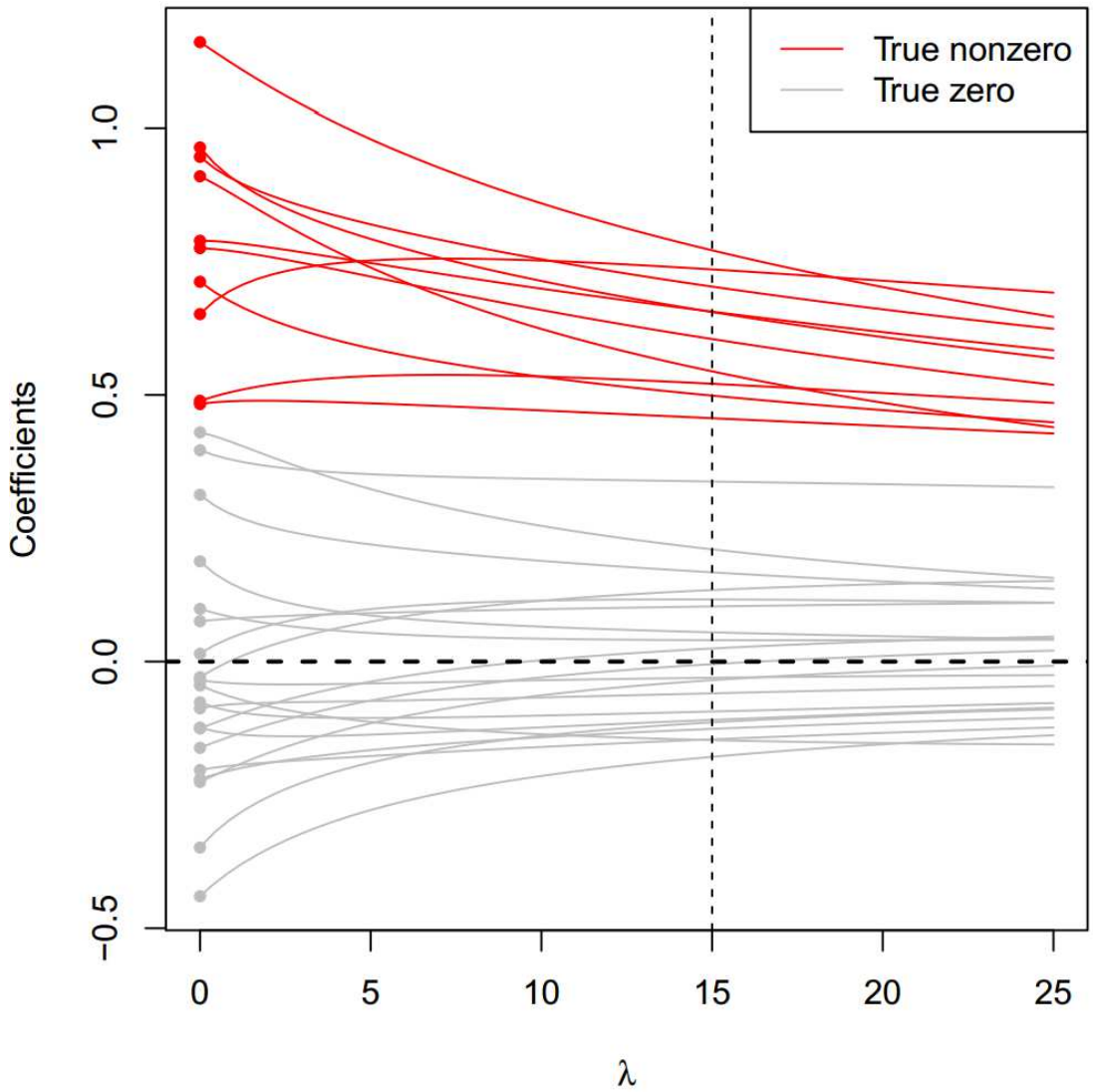


Пример 3

Сгенерируем y согласно линейной модели с 10 большими коэффициентами (между 0.5 и 1) и 20 нулевыми.



Пример 3



Оценки нулевых коэффициентов не становятся равны нулю, а только уменьшаются.

Выбор λ

Эмпирический способ: выбирается такое λ, начиная с которого коэффициенты модели меняются незначительно.

Кросс-валидация: выбирается λ, доставляющее минимум средней ошибке на контроле; модель затем настраивается по полным данным.

Лассо

Lasso (Least Absolute Selection and Shrinkage Operator):

$$\hat{\theta}^{lasso} = \underset{\theta}{\text{argmin}} \left(\|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right).$$

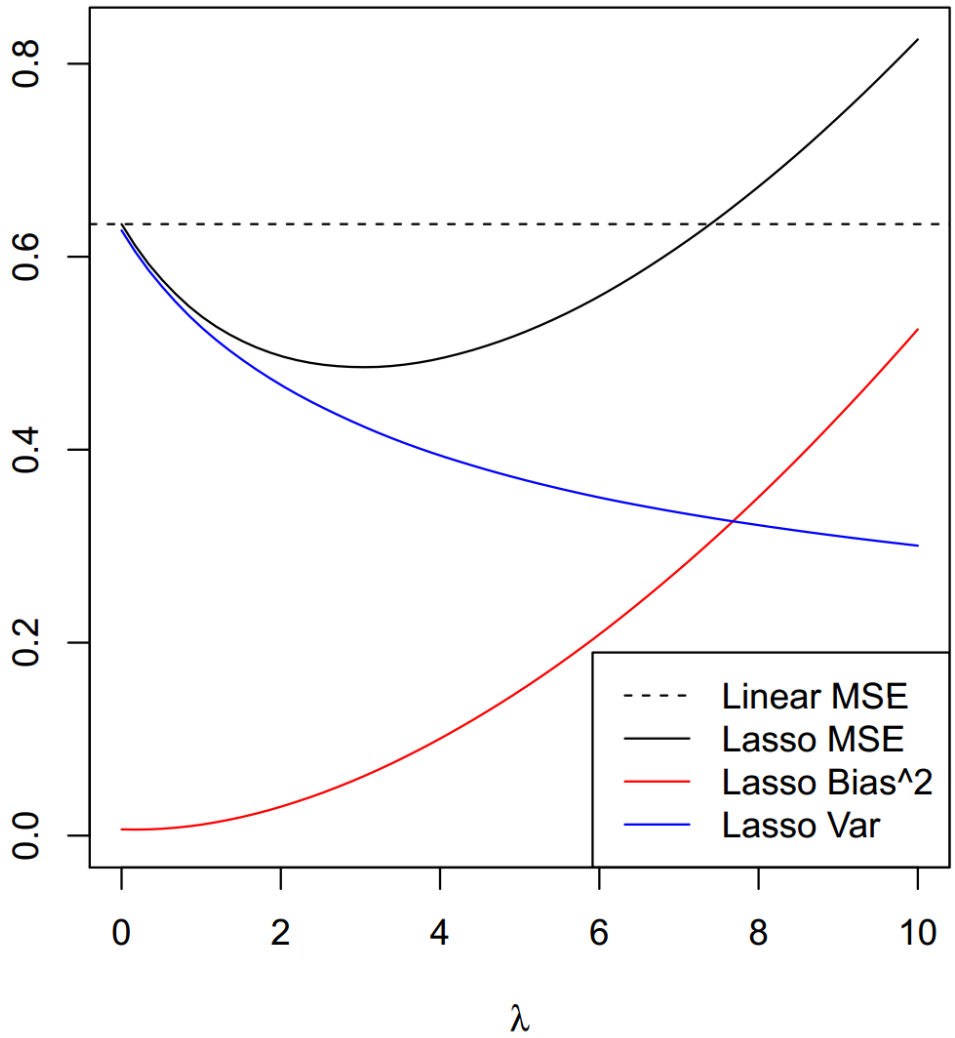
Выражения в замкнутом виде не существует.

l_1-регуляризация позволяет обнулять коэффициенты.

Всё остальные детали те же, что у гребневой регрессии: λ определяет баланс между смещением и дисперсией, θ_0 не входит в регуляризатор, признаки нужно стандартизировать, итоговую модель необходимо приводить к записи в исходных величинах, λ выбирается по кросс-валидации.

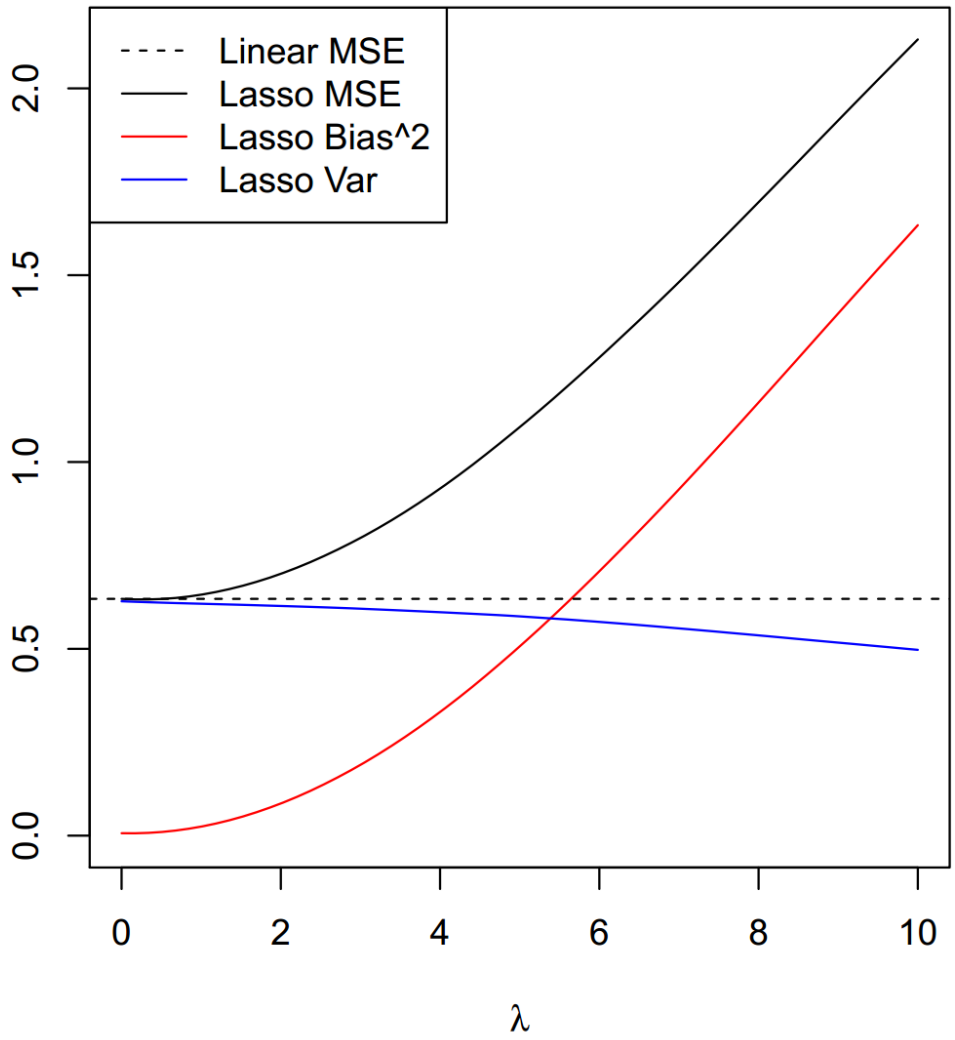
Пример 1

10 больших коэффициентов (между 0.5 и 1) и 20 маленьких (между 0 и 0.3).



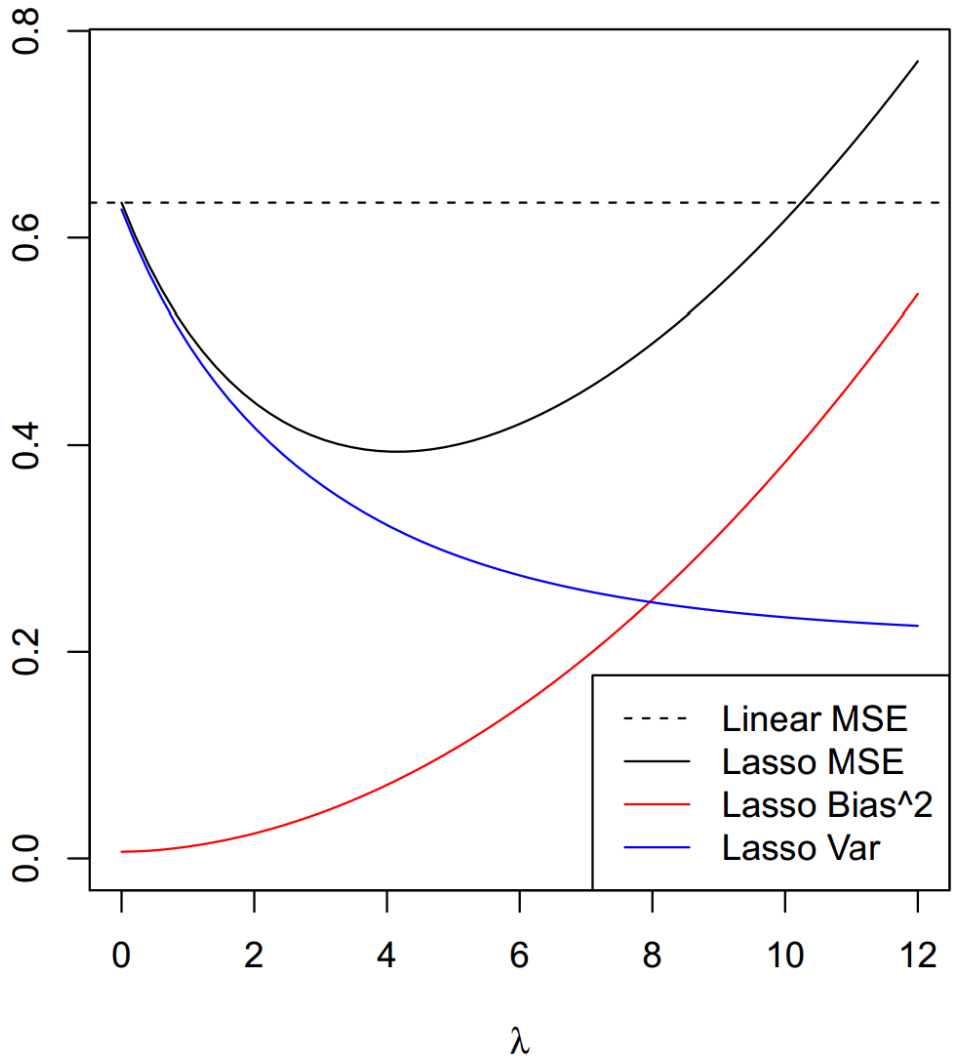
Пример 2

30 больших коэффициентов (между 0.5 и 1).



Пример 3

10 больших коэффициентов (между 0.5 и 1) и 20 нулевых.



Особенности

- Алгоритм LARS позволяет получить значения коэффициентов при всех λ.
- Лассо можно применять даже при $k > n$, но получится не больше n ненулевых коэффициентов.
- При $n > k$ и наличии высоко коррелированных предикторов лассо проигрывает по ошибке предсказания гребневой регрессии.
- Если среди признаков есть группа высоко коррелированных, лассо, как правило, отбирает только один из них.

Эластичная сеть

Elastic net:

$$\hat{\theta}^{en} = \left(1 + \frac{\lambda_2}{n}\right) \underset{\theta}{\operatorname{argmin}} \left(\|y - X\theta\|_2^2 + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_2^2\right).$$

Другая форма записи:

$$\hat{\theta}^{en} = \underset{\theta}{\operatorname{argmin}} \left(\theta^T \left(\frac{X^T X + \lambda_2 I_n}{1 + \lambda_2}\right) \theta - 2y^T X\theta + \lambda_1 \|\theta\|_1\right).$$

Лассо в аналогичном виде:

$$\hat{\theta}^{lasso} = \underset{\theta}{\operatorname{argmin}} \left(\theta^T X^T X\theta - 2y^T X\theta + \lambda_1 \|\theta\|_1\right).$$

Роль параметров регуляризации

- При $\lambda_1 = 0$ получаем гребневую регрессию с параметром λ_2.
- При $\lambda_2 = 0$ получаем лассо с параметром λ_1.
- При $\lambda_1 = \infty$ получаем $\hat{\theta}_j^{en} = 0$.
- При $\lambda_2 = \infty$ получаем univariate soft thresholding:

$$\hat{\theta}_j^{UST} = \left(\left|y^T x_j\right| - \frac{\lambda_1}{2}\right)_+ \operatorname{sign}\left(y^T x_j\right),\ j = 1, \ldots, k.$$

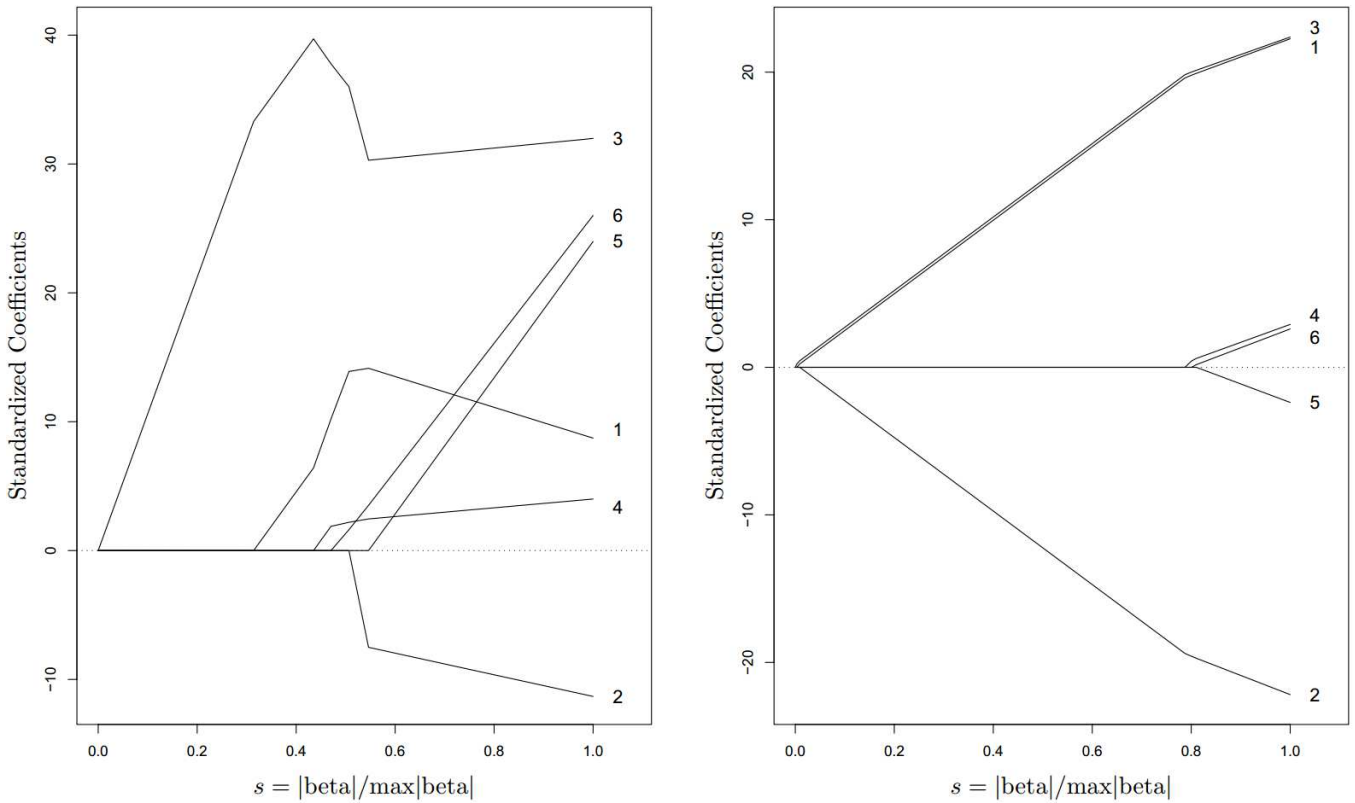
Значения параметров выбираются кросс-валидацией, λ_2 по небольшой сетке (например, $(0.0, 0.01, 0.1, 1, 10, 100)$), а λ_1 — по непрерывной кривой, получаемой алгоритмом LARS-EN.

Пример

Пусть $Z_1, Z_2 \sim N(0,1), \quad y \sim N\left(Z_1 + 0.1Z_2, 1\right),$

$$X_1 = Z_1 + \varepsilon_1, \qquad X_2 = -Z_1 + \varepsilon_2, \qquad X_3 = Z_1 + \varepsilon_3,$$
$$X_4 = Z_2 + \varepsilon_4, \qquad X_5 = -Z_2 + \varepsilon_5, \qquad X_6 = Z_2 + \varepsilon_6,$$

$\varepsilon_i \sim N(0, 1/16)$.



Слева лассо, справа эластичная сеть с $\lambda_2 = 0.5$.

Прикладная статистика
9. Регрессионный анализ.

Рябенко Евгений
riabenko.e@gmail.com