

СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 22. Методы регуляризации в задаче множественной линейной регрессии. Часть 2



Пример 2

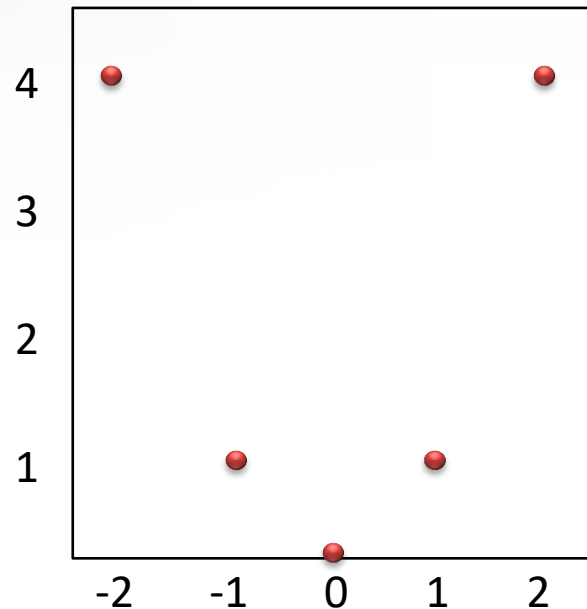
Имеется 5 наблюдений значений фактора X и отклика Y :

$(x_k, y_k): (-2, 4); (-1, 1); (0, 0); (1, 1); (2, 4); n=5$


Задача: построить полиномиальную регрессию по этим точкам. Степень полинома $p = 10$.

Задача полиномиальной регрессии сводится к задаче множественной линейной регрессии с набором факторов:

$$X_1 = X, X_2 = X^2, \dots, X_p = X^p.$$



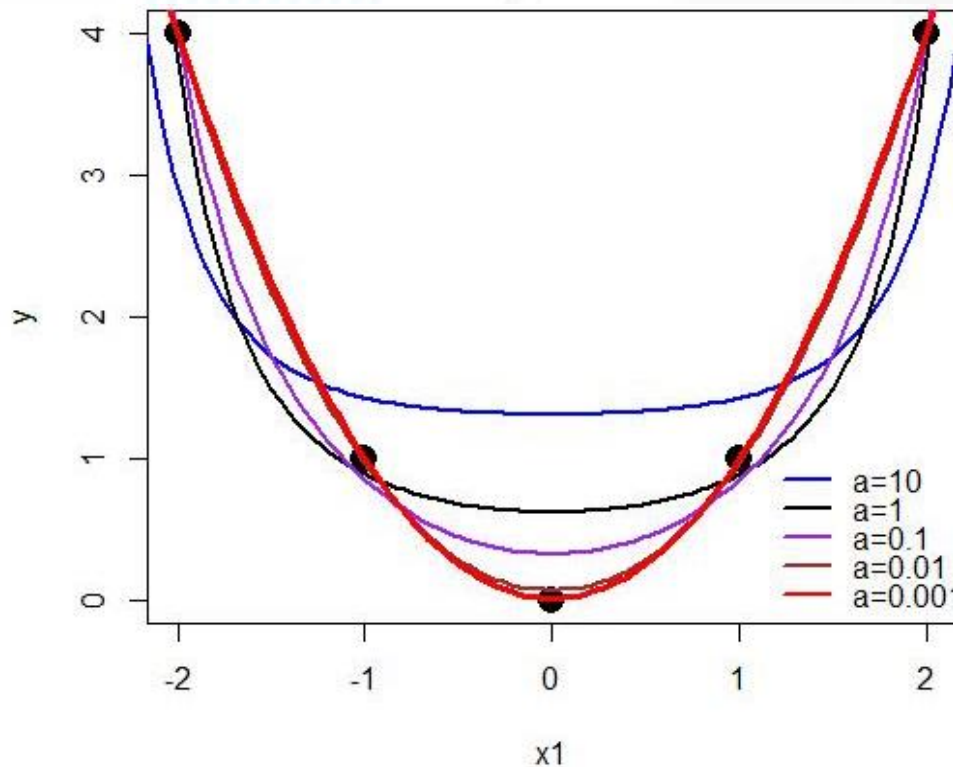
Пример 2



	1	2	3	4	5
(Intercept)	1.309515e+00	6.206787e-01	3.249698e-01	6.946130e-02	1.195367e-02
V1	.	1.822752e-18	4.105493e-17	2.240432e-16	3.478002e-16
V2	9.338630e-02	2.172650e-01	4.744654e-01	8.591770e-01	9.631473e-01
V3	-2.349304e-19	-4.428233e-18	-4.637001e-17	-2.756894e-16	-4.240427e-16
V4	2.001625e-02	3.906499e-02	4.156606e-02	3.800476e-02	2.118786e-02
V5	8.443704e-21	2.666364e-19	2.919109e-18	3.616119e-17	6.426918e-17
V6	4.792930e-03	8.931313e-03	6.570190e-03	-8.753660e-04	-1.811850e-03
V7	1.804045e-21	6.944738e-20	7.635718e-19	3.440821e-18	3.698614e-18
V8	1.185187e-03	2.181991e-03	1.359141e-03	-4.104496e-04	-4.816969e-04
V9	3.860143e-22	1.755928e-20	1.800678e-19	3.103876e-19	3.254313e-19
V10	2.954841e-04	5.429914e-04	3.241598e-04	4.562459e-05	3.488022e-05



Пример 2

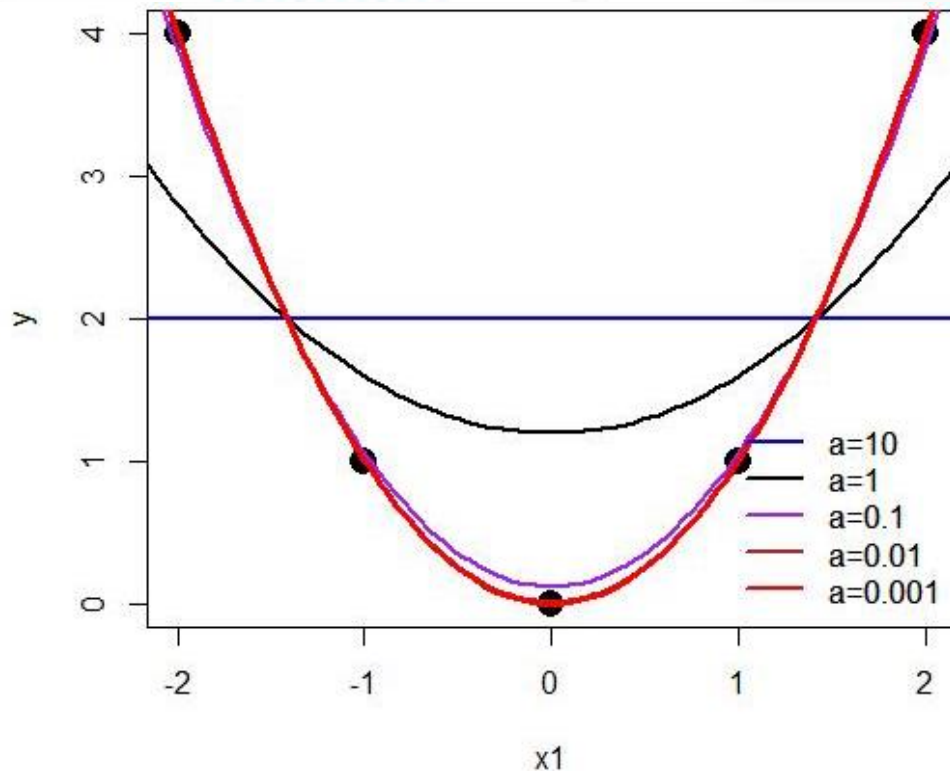


λ	RSS
10	4.43
1	0.49
0.1	0.15
0.01	0.0072
0.001	0.00021


Пример 2

	1	2	3	4	5
(Intercept)	2	1.1952286	0.1195229	0.01195229	0.001195229
v1
v2	.	0.4023857	0.9402386	0.99402386	0.999402386
v3
v4
v5
v6
v7
v8
v9
v10

Пример 2



λ	RSS
10	14
1	5
0.1	0.05
0.01	0.0005
0.001	0.000005



Выбор штрафного параметра λ

В среде R в пакете **glmnet** реализована процедура перекрестной проверки, которая позволяет найти подходящее значение λ . Эта процедура разделяет наблюдения на контрольную и обучающую выборки. Задается последовательность параметров штрафа. По обучающей выборке для каждого значения параметра штрафа строятся оценки регрессионных коэффициентов. Затем полученная регрессионная модель проверяется на контрольной выборке.

Перекрестная проверка в R реализуется с помощью процедуры **cv.glmnet()**.

Пример. Обращение к процедуре и использование оптимального значения λ :

```
ridge.cv=cv.glmnet(x,y,alpha=0)
lambda0=ridge.cv$lambda.min
ridge.mod=glmnet(x,y,alpha=0,lambda= lambda0)
```

Эквивалентные формулировки

Для каждого значения λ найдется такое значение s , что гребневая регрессия может быть сведена к решению задачи условной оптимизации

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \rightarrow \min_{\beta}$$

при условии

$$\sum_{j=1}^p \beta_j^2 \leq s.$$

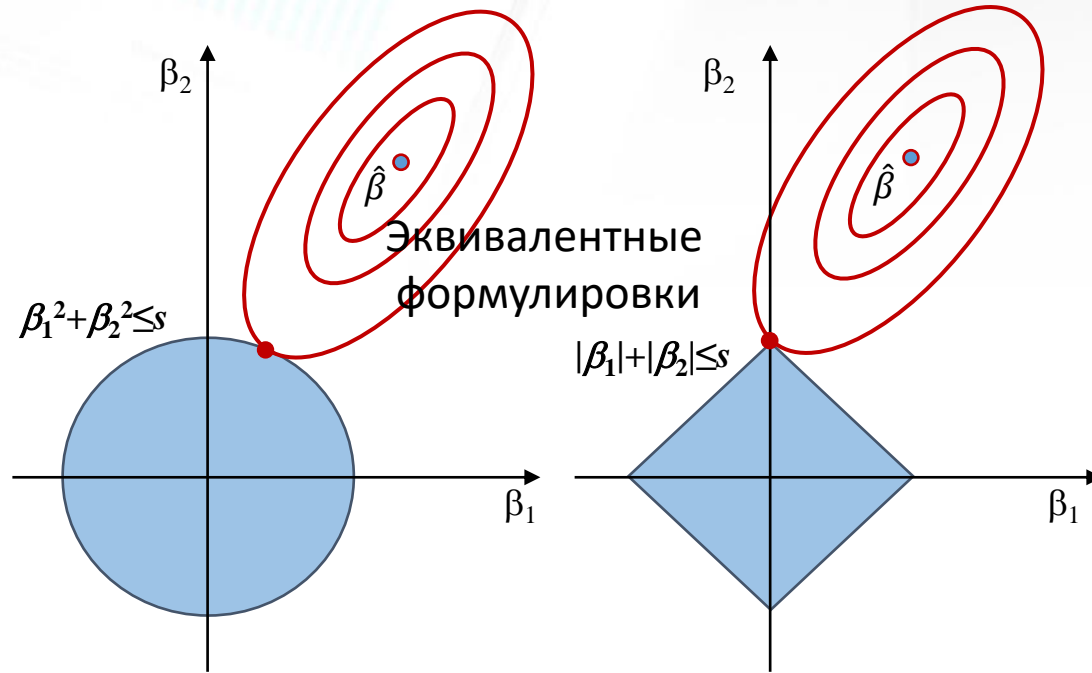
Для каждого значения λ найдется такое значение s , что метод лассо может быть сведен к решению задачи условной оптимизации

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \rightarrow \min_{\beta}$$

при условии

$$\sum_{j=1}^p |\beta_j| \leq s.$$

Эквивалентные формулировки



Пример. Частный случай $p = 2$. Красным цветом изображены линии уровня функции RSS . Синим цветом - области, задаваемые ограничениями. Оптимальное значение коэффициентов β_1, β_2 соответствуют точке первого касания линии уровня и области ограничений.



Сравнение гребневой и лассо регрессий

Лассо лучше работает в задачах, в которых относительно небольшое число факторов имеют существенные значения коэффициентов. Гребневая регрессия работает лучше, когда отклик является функцией многих факторов с коэффициентами примерно одной величины.

Реализация метода гребневой регрессии значительно проще реализации метода лассо.

Как и в гребневой регрессии, лассо регрессия уменьшает дисперсию за счет небольшого увеличения смещения и соответственно обеспечивает более точный прогноз.

В отличие от гребневой регрессии, лассо производит отбор переменных, что позволяет легче интерпретировать ее модели.

Пример 3: случай $n=p$.

Рассмотрим частный случай $n = p$ и единичной матрицы значений факторов X , $\beta_0 = 0$. МНК дает:

$$RSS = \sum_{j=1}^p (y_j - \beta_j)^2 \rightarrow \min_{\beta}.$$

Решение:

$$\hat{\beta}_j = y_j$$

Гребневая и лассо регрессии дают следующие решения:

$$\sum_{j=1}^p (y_j - \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \rightarrow \min_{\beta},$$

Решение:

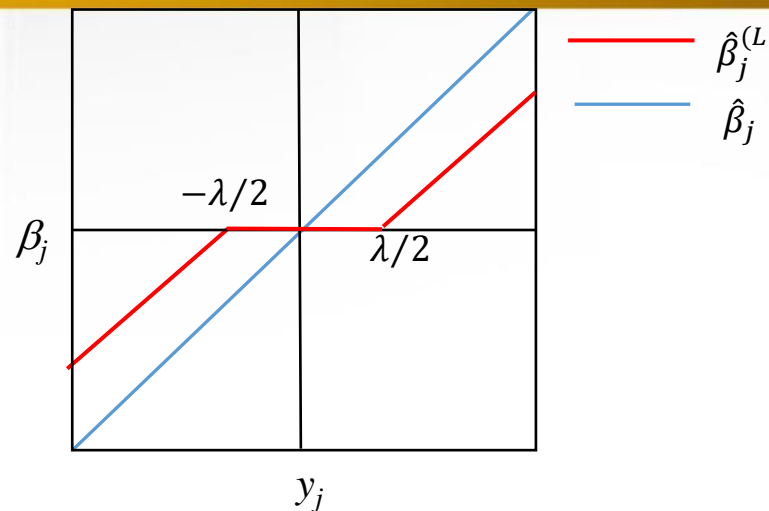
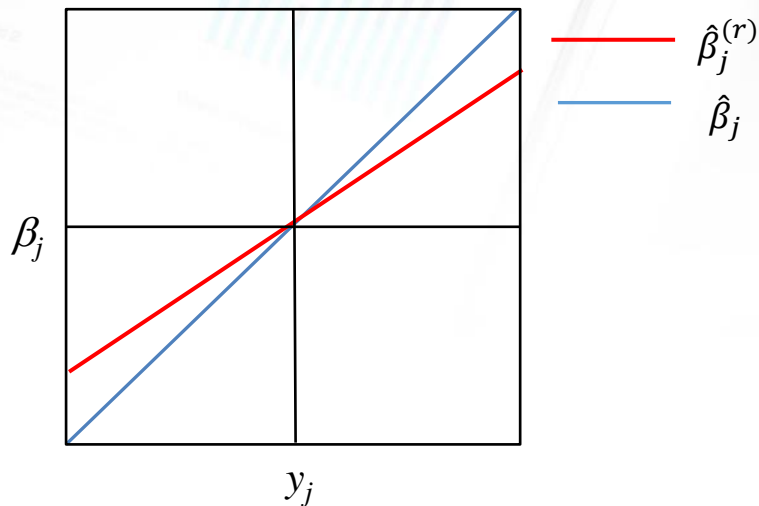
$$\hat{\beta}_j^{(r)} = y_j / (1 + \lambda)$$

$$\sum_{j=1}^p (y_j - \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \rightarrow \min_{\beta},$$

Решение:

$$\hat{\beta}_j^{(L)} = \begin{cases} y_j - \lambda/2, & \text{если } y_j > \lambda/2, \\ y_j + \lambda/2, & \text{если } y_j < -\lambda/2, \\ 0, & \text{если } |y_j| \leq \lambda/2 \end{cases}$$

Пример 3: случай $n=p$



Свойство отбора переменных: в методе лассо при $|y_i| \leq \lambda/2$ оценки параметров регрессии β обращаются в нуль.

Свойство робастности оценок по методу лассо: ошибка оценки параметров β в гребневой регрессии линейно растет с ростом отклика. Ошибка метода лассо остается постоянной.

Заключение

- Приведена мотивация и постановка задачи регуляризации решения задачи регрессии в случае больших размерностей.
- Сформулирована процедура регуляризации по методу гребневой регрессии и проведено ее сравнение с МНК-регрессии.
- Сформулирована процедура регуляризации по методу лассо и проведено ее сравнение с гребневой и МНК регрессиями.
- Приведены примеры методов лассо и гребневой регрессии.
- Методы гребневой и лассо регрессий «стягивают» к нулю оценки коэффициентов множественной линейной регрессии в том смысле, что уменьшается норма вектора этих оценок при увеличении коэффициента штрафа λ .
- Гребневая регрессия не обращает в нуль коэффициенты множественной линейной регрессии даже при больших значениях коэффициента штрафа λ .
- Лассо регрессия, в отличие от МНК и гребневой регрессии, осуществляет выбор подмножества значимых факторов, то есть некоторые коэффициенты множественной линейной регрессии обращаются в нуль, что упрощает интерпретацию результатов регрессионного анализа

Контрольные вопросы и задания

1. Строго вывести решения задач из примера 3 для гребневой и лассо регрессий.
2. Рассмотреть набор данных x, y : $(-2, -7)$, $(-1, 0)$, $(0, 1)$, $(1, 2)$, $(2, 9)$. Построить регрессионные полиномы степени $p=11$ методами гребневой и лассо регрессий. Сделать выводы о значимых факторах, структуре восстанавливаемой модели и приемлемом значении коэффициента штрафа.
3. В предыдущей задаче добавить к значениям отклика шум, распределенный по нормальному закону с нулевым средним и стандартным отклонением $(0.1, 0.2$ и $0.3)$. Построить регрессионные полиномы степени $p=11$ методами гребневой и лассо регрессий. Сделать выводы о влиянии шума на оценки коэффициентов регрессий.