МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

(Machine Learning and Data Mining)

Н. Ю. Золотых

http://www.uic.unn.ru/~zny/ml

Лекция 6

Борьба с переобучением

6.1. Переобучение

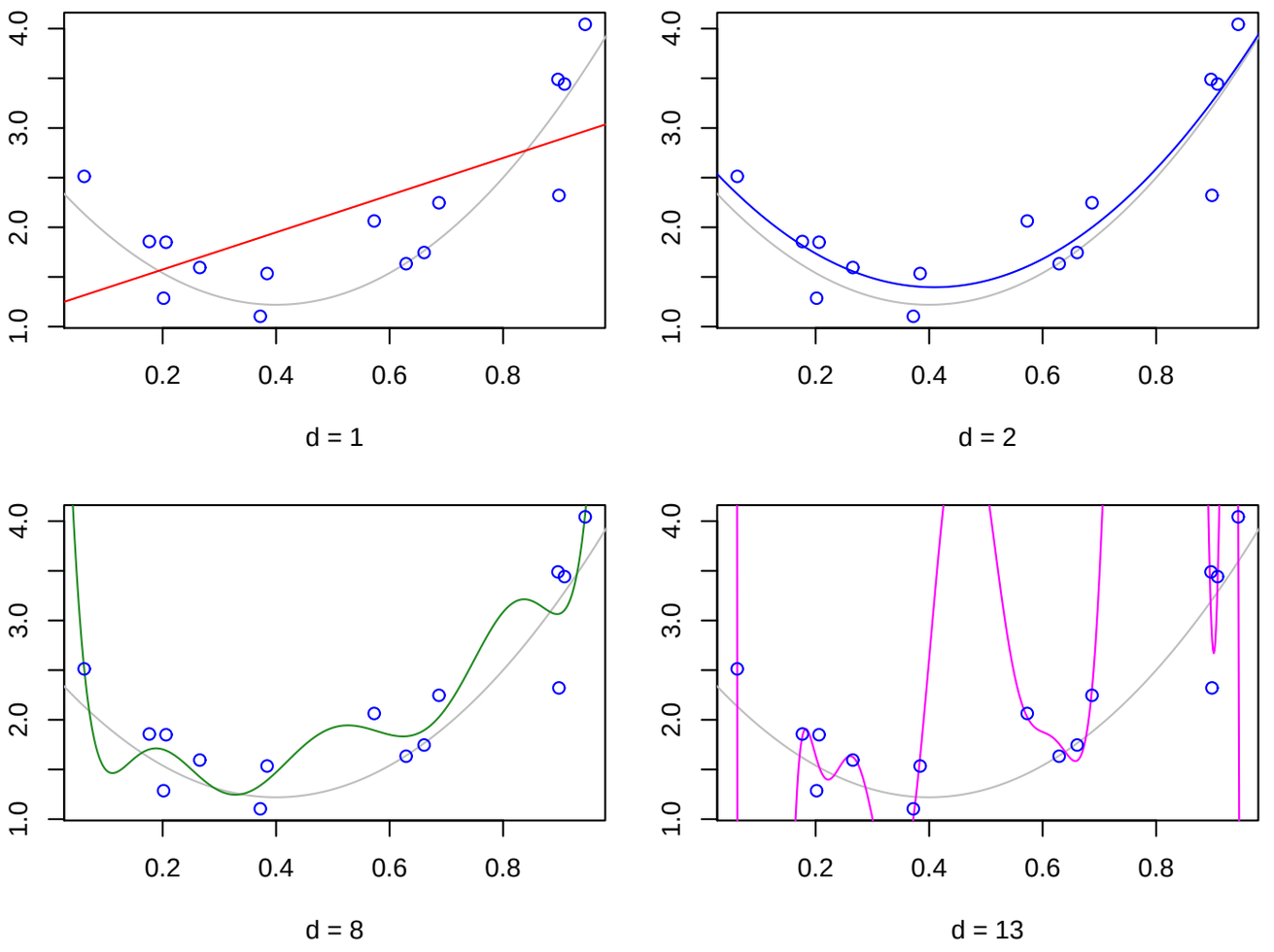
Переобучение — это явление, при котором обучающий алгоритм выдает хорошие результаты на обучающей выборке, но имеет очень плохие обобщающие свойства.

Типичное поведение ошибки на обучающей и тестовой выборках с ростом «сложности» модели:

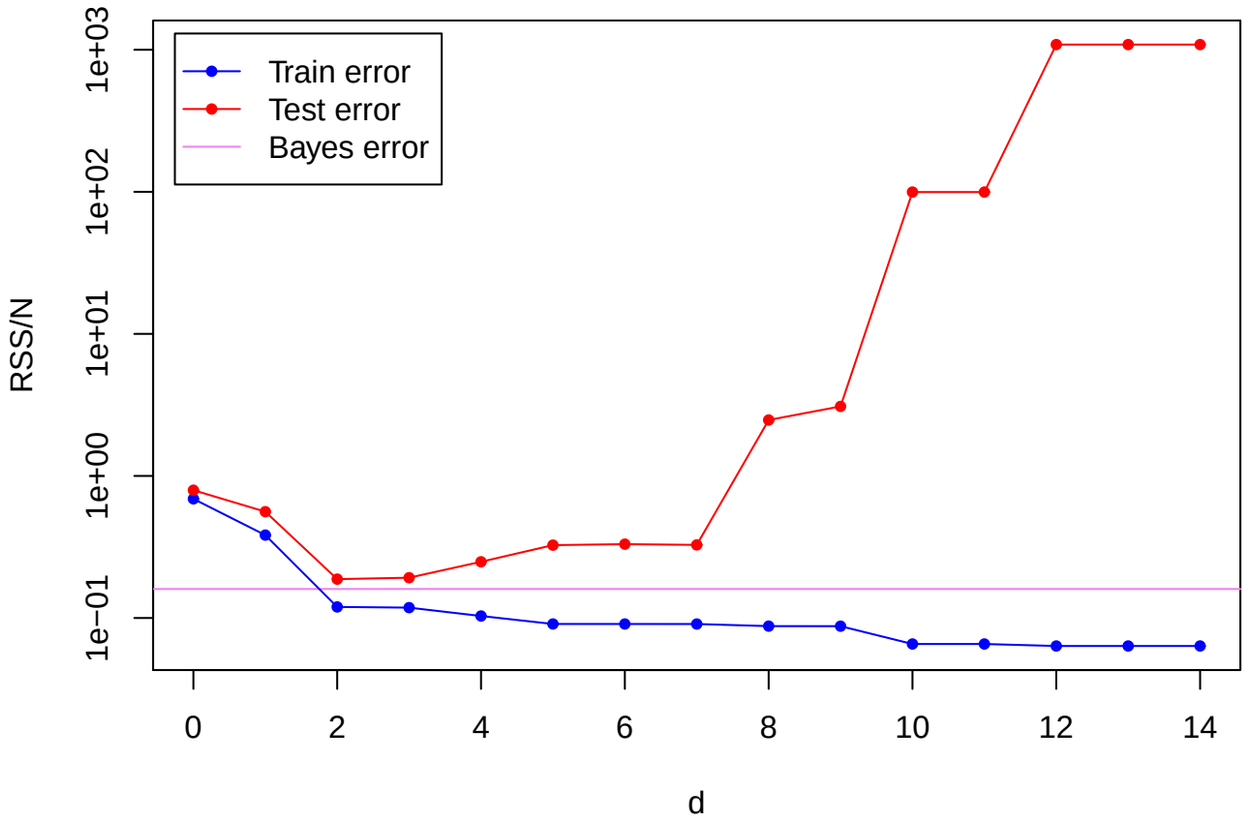
- Обычно на обучающей выборке с ростом сложности модели ошибка уменьшается.
- На тестовой выборке сперва с ростом сложности модели ошибка уменьшается, но с некоторого момента ошибка начинает расти: сказывается синдром переобучения.

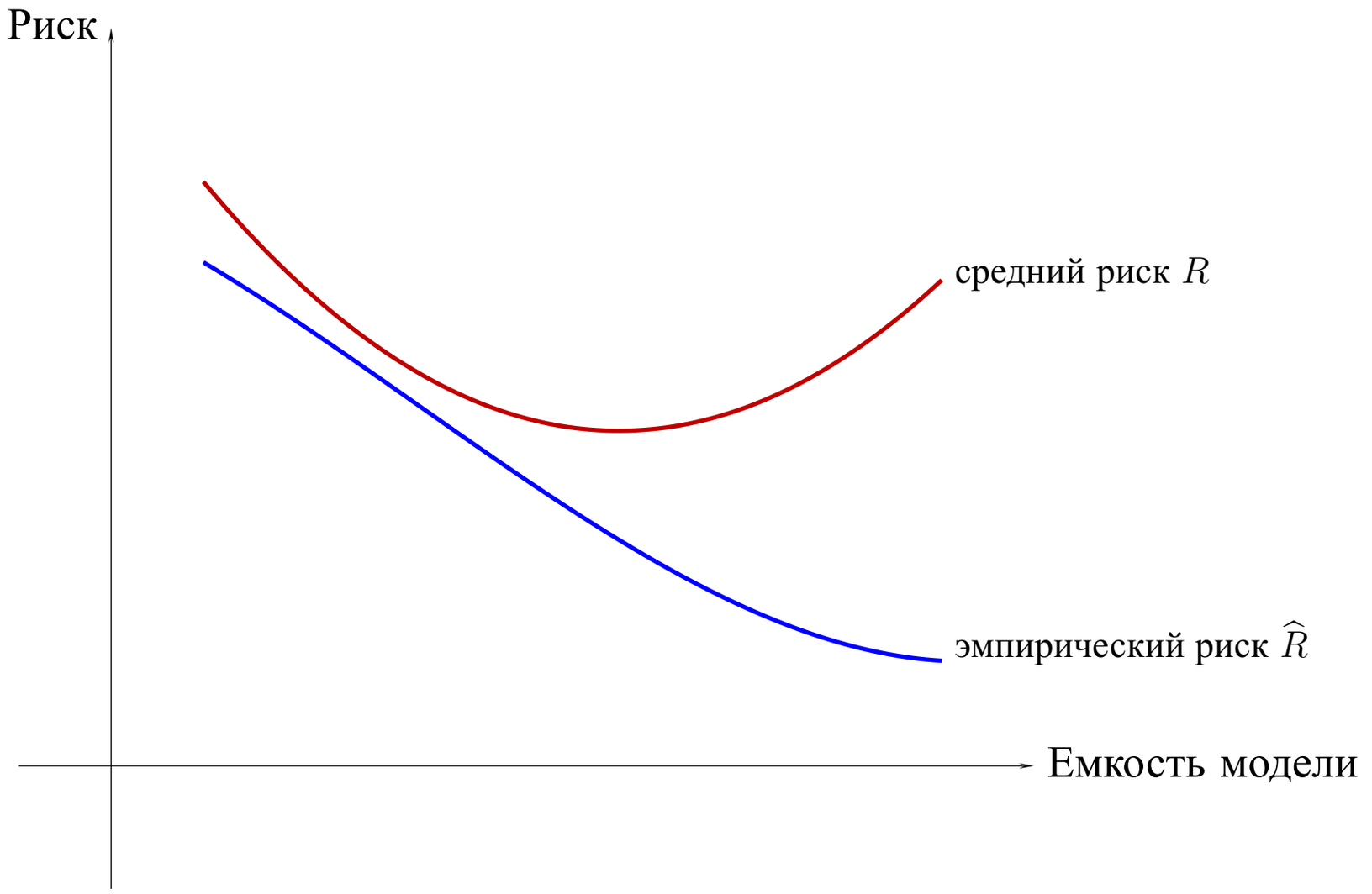
Когда модель слишком сложна, она, как правило, хорошо приспосабливается к конкретным обучающим данным, улавливая какие-то специфичные для них особенности, но не присущие всей генеральной совокупности, поэтому на тестовых данных ошибка может быть большой.

$$y = 8x^2 - 6.4x + 2.5 + N(0, 0.4)$$



RSStrain/Ntrain = 0.1196395 RSStest(3)/Ntest = 0.1876475 $\sigma^2 = 0.16$




Риск
средний риск R
эмпирический риск $\widehat{R}$
Емкость модели

6.1.1. Причины переобучения (в задаче восстановления регрессии)

- *Мультиколлинеарность*: $\mathbf{X}$ плохо обусловлена (т. е. столбцы матрицы $\mathbf{X}$ образуют систему, близкую к линейно зависимой).

$$\operatorname{cond}_2\mathbf{X} = \max_{\beta\neq 0}\frac{\|\mathbf{X}\beta\|_2}{\|\beta\|_2}\Bigg/\min_{\beta\neq 0}\frac{\|\mathbf{X}\beta\|_2}{\|\beta\|_2} = \frac{\max\sqrt{\lambda_j(\mathbf{X}^\top\mathbf{X})}}{\min\sqrt{\lambda_j(\mathbf{X}^\top\mathbf{X})}} = \frac{\max\sigma_j}{\min\sigma_j} = \frac{\sigma_1}{\sigma_d}$$

  (отношение максимального сингулярного числа матрицы $\mathbf{X}$ к минимальному)

- *Наличие неинформативных признаков*

- *Слишком мало данных*.

  Даже если все признаки информативны, данных может оказаться слишком мало.

6.1.2. Замечание о числе обусловленности

Необходимо смотреть на число обусловленности не для с.л.у., а для задачи L.S.

$$\mathbf{X}\beta \approx \mathbf{y} \qquad \Rightarrow \qquad (\mathbf{X} + \Delta\mathbf{X})\beta \approx \mathbf{y} + \Delta\mathbf{y}$$

Можно доказать, что при малых возмущениях $\Delta\mathbf{X}$ и $\Delta\mathbf{y}$

$$\frac{\|\Delta\beta\|_2}{\|\beta\|_2} \leq \varepsilon \cdot \underbrace{\left(\frac{2\,\mathrm{cond}_2\,\mathbf{X}}{\cos\theta} + \mathrm{tg}\,\theta \cdot \mathrm{cond}_2^2\,\mathbf{X} \right)}_{\mathrm{cond}_{\mathrm{LS}}(\mathbf{X},\mathbf{y})} + O(\varepsilon^2),$$

где

$$\varepsilon = \max\left\{ \frac{\|\Delta\mathbf{X}\|_2}{\|\mathbf{X}\|_2}, \; \frac{\|\Delta\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \right\}, \qquad \mathrm{cond}_2\,\mathbf{X} = \frac{\sigma_1}{\sigma_d},$$

θ — угол между $\mathbf{y}$ и $L(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d)$.

Если угол θ мал, то $\mathrm{cond}_{\mathrm{LS}}(\mathbf{X}, \mathbf{y}) \approx \mathrm{cond}_2\,\mathbf{X}$

Если $\theta \to \pi/2$ мал, то $\mathrm{cond}_{\mathrm{LS}}(\mathbf{X}, \mathbf{y}) \to \infty$

6.2. Сокращение числа параметров и «усадка» коэффициентов

Рассмотрим некоторые методы по уменьшению числа признаков (уменьшения размерности пространства признаков) и «усадке» коэффициентов.

Зачем?

- Борьба с *переобучением*: чем меньше параметров (и входных переменных) или чем меньше по величине параметры, тем проще модель и снижается возможность переобучения
- *Интерпретация*: чем меньше параметров (и входных переменных), тем проще интерпретировать модель

6.2.1. Выбор подмножества признаков

Для каждого $k \in \{0, 1, \ldots, d\}$ найдем подмножество входных параметров мощности k, для которого $\mathrm{RSS}(\beta)$ минимально.

Большой перебор!

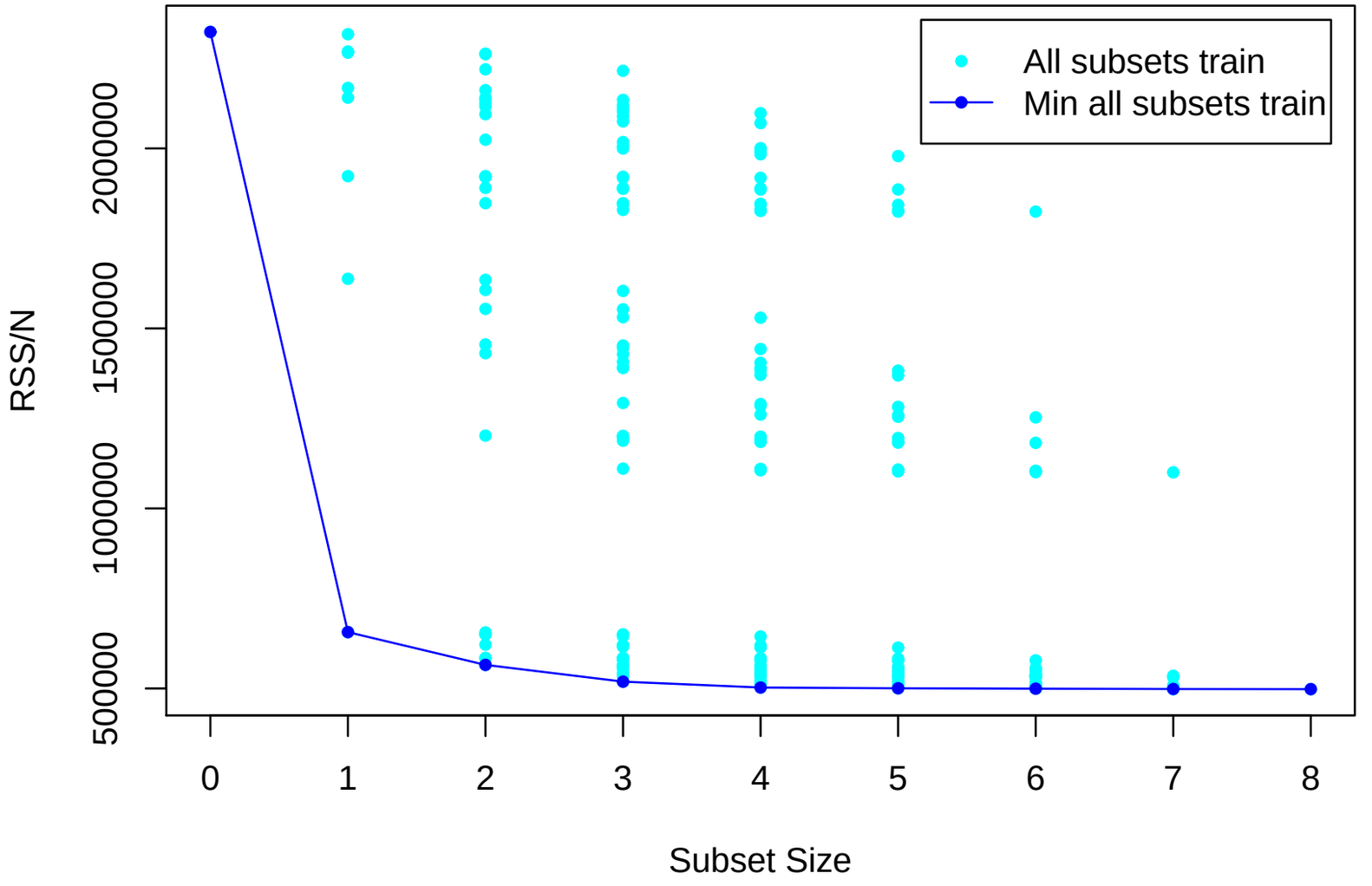
Оптимальное подмножество признаков мощности k не обязательно содержится в оптимальном подмножестве мощности $k + 1$

Задача предсказания цены квартиры

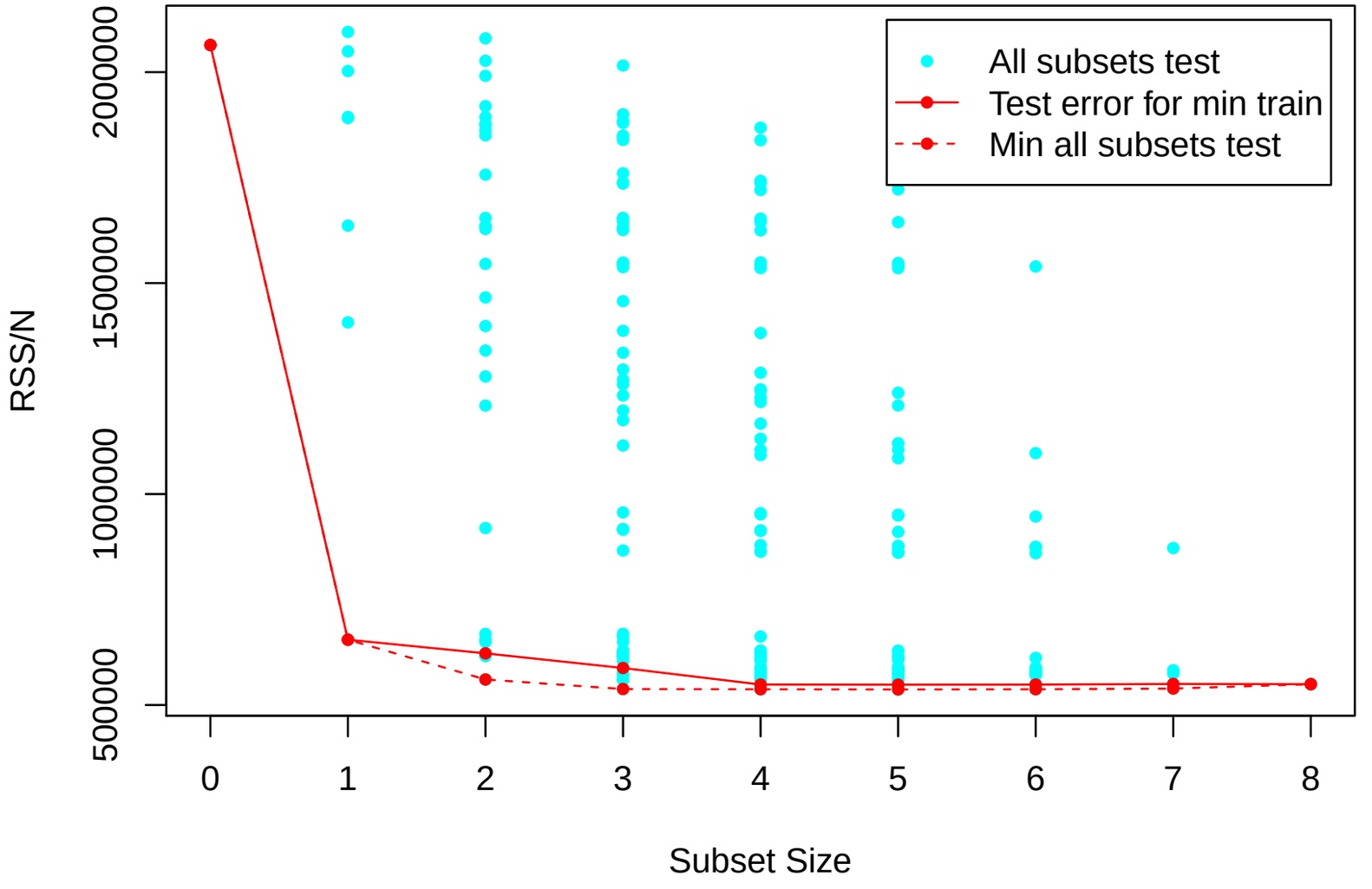
Price	цена квартиры (тыс. руб.)
Date	№ дня, в который квартира выставлена на продажу
Lat	географическая широта объекта недвижимости
Lng	географическая долгота объекта недвижимости
Housing	тип недвижимости (0 — вторичное жилье, 1 — новостройка)
Floors	количество этажей в доме
House	тип дома (Block — блочный, Brick — кирпичный, Monolithic — монолитный, Panel — панельный, Wooden — деревянный)
Rooms	количество комнат (0 — квартира-студия)
Floor	№ этажа
Area	площадь квартиры (м^2)

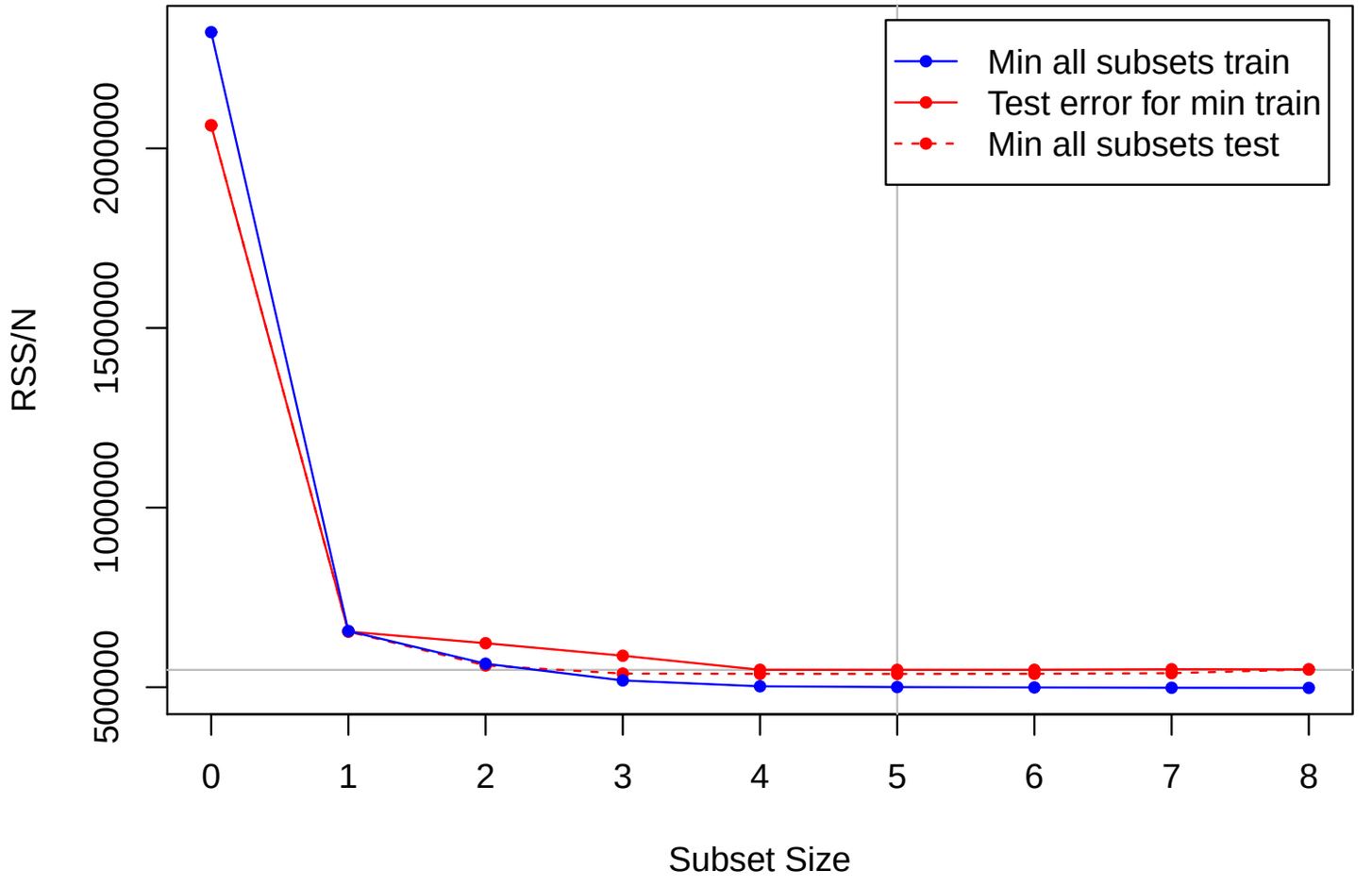
8 входных признаков, $2^8 = 256$ подмножеств

На обучающей выборке:



На тестовой выборке:





$$d = 5: \quad \widehat{R}_{\text{train}} = \frac{1}{N_{\text{train}}} \text{RSS}_{\text{train}} = 499710.7, \quad \widehat{R}_{\text{test}} = \frac{1}{N_{\text{test}}} \text{RSS}_{\text{test}} = 545722.3$$

$$y = -101009 + 2294.1\,\textsf{Lng} - 152.54\,\textsf{Housing} + 32.016\,\textsf{Floors} - 370.43\,\textsf{Rooms} + 74.944\,\textsf{Area}$$

Forward stepwise

находим константную регрессию и ошибку на тестовой выборке R_0

$J \leftarrow \emptyset$

for $k = 1, \ldots, d$

$R^* \leftarrow \infty$

for $j \in \{1, \ldots, d\} \setminus J$

строим линейную регрессию, используя признаки из $J \cup \{j\}$

вычисляем $\widehat{R}_{\text{test}}$ — ошибку этой регрессии на тестовой выборке

if $\widehat{R}_{\text{test}} < R^*$

$j^* \leftarrow j$

$R^* \leftarrow \widehat{R}_{\text{test}}$

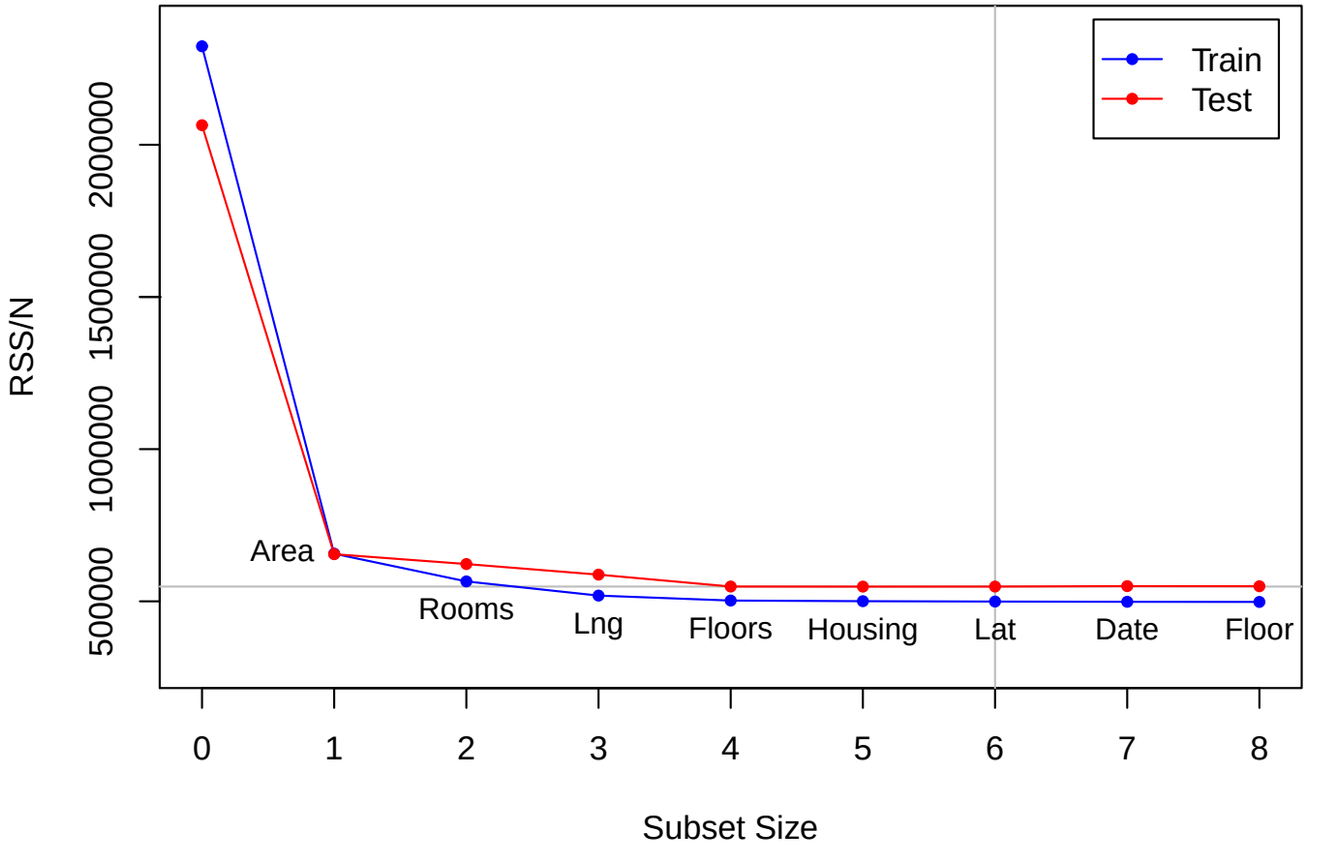
$j_k \leftarrow j^*$

$R_k \leftarrow R^*$

$J \leftarrow J \cup \{j^*\}$

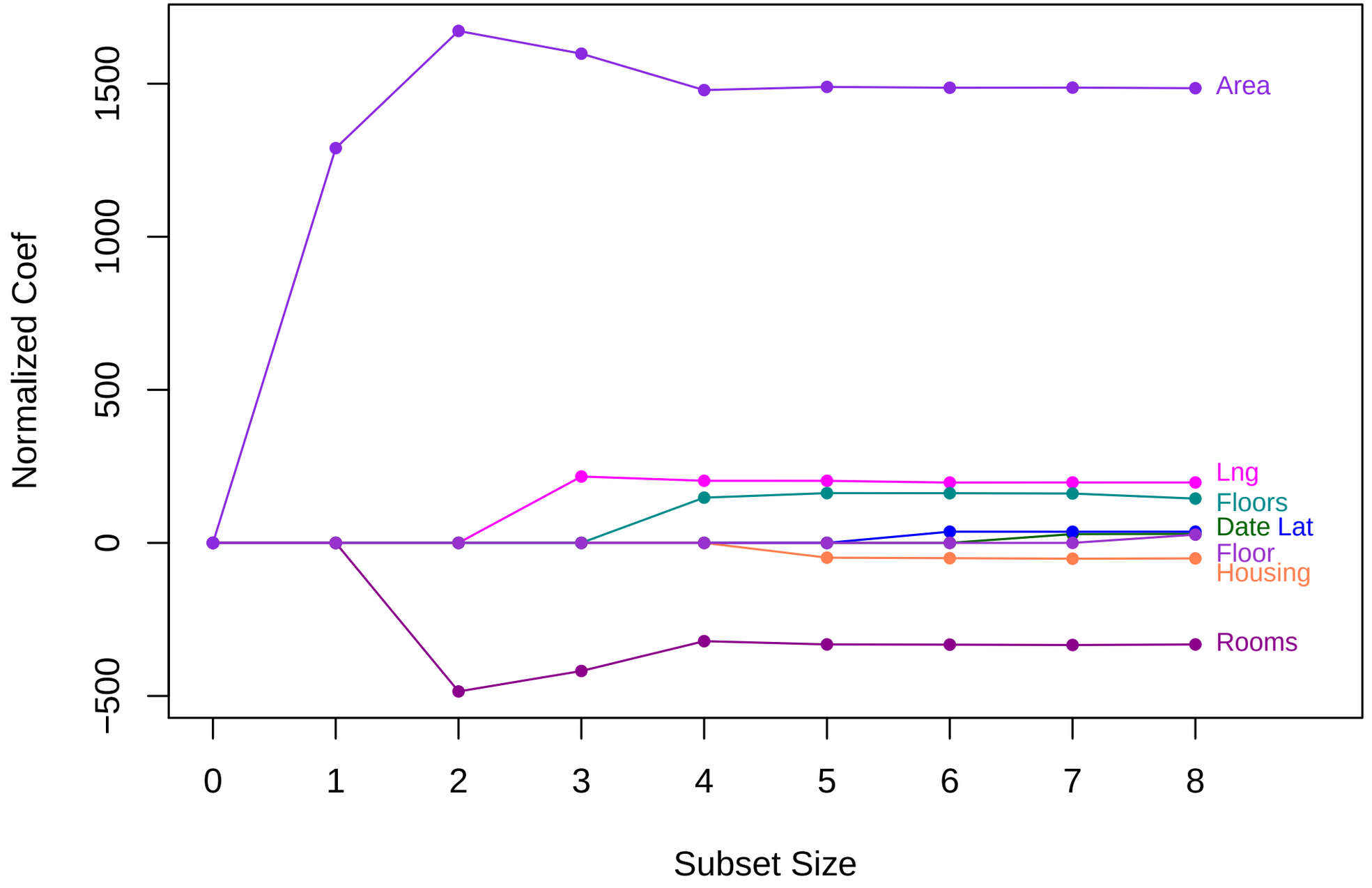
На выходе получаем последовательность номеров признаков $j_1, j_2, \ldots, j_d$ и значения ошибок $R_0, R_1, \ldots, R_d$, где R_k — ошибка на тестовой выборке регрессии, построенной по признакам $j_1, j_2, \ldots, j_k$

Другая стратегия — начав с рассмотрения всего набора входных переменных, последовательно исключать их из этого набора (backward stepwise).



$$y = -141660 + 772.88\,\mathsf{Lat} + 2229.3\,\mathsf{Lng} - 158.11\,\mathsf{Housing}$$
$$+31.941\,\mathsf{Floors} - 371.24\,\mathsf{Rooms} + 74.812\,\mathsf{Area}$$

$$k = 5: \quad \widehat{R}_{\text{train}} = \frac{1}{N_{\text{train}}}\,\text{RSS}_{\text{train}} = 545813.4, \quad \widehat{R}_{\text{test}} = \frac{1}{N_{\text{test}}}\,\text{RSS}_{\text{test}} = 497700.3$$


Normalized Coef
1500
1000
500
0
−500
0
1
2
3
4
5
6
7
8
Subset Size
Area
Lng
Floors
Date
Lat
Floor
Housing
Rooms

6.2.2. «Ридж» (гребневая) регрессия (регуляризация)

(*ridge regression*)

Добавим к RSS штрафную функцию, чувствительную к абсолютной величине коэффициентов β_j:

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y^{(i)} - \beta_0 - \sum_{j=1}^{d} x_j^{(i)} \beta_j \right)^2 + \lambda \sum_{j=1}^{d} \beta_j^2 \right\},$$

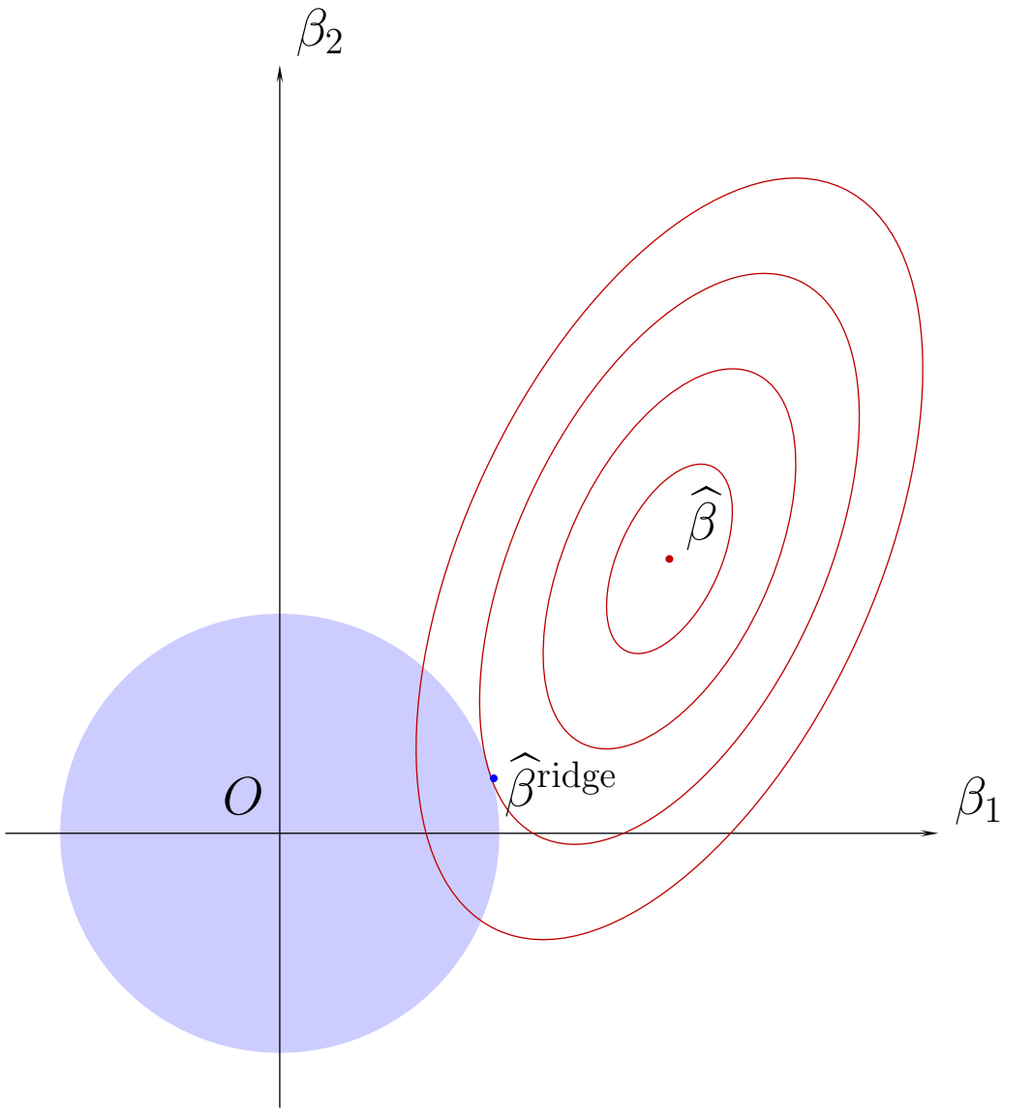
где λ — неотрицательный параметр:

чем больше λ, тем больше чувствительность штрафа к величине коэффициентов

Понятно, что задача эквивалентна следующей задаче условной минимизации:

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y^{(i)} - \beta_0 - \sum_{j=1}^{p} x_j^{(i)} \beta_j \right)^2 \right\}, \text{ при условии } \sum_{j=1}^{d} \beta_j^2 \leq s.$$

Параметр s определяется по λ, и наоборот.


β_2
$\widehat{\beta}$
O
$\widehat{\beta}^{\text{ridge}}$
β_1

Если в линейной регрессионной модели много коррелированных переменных, то параметры модели определяются с трудом и имеют большую дисперсию.

Например, X_j и $X_{j'}$ сильно зависимы.

Тогда возможна ситуация:

β_j велик и положителен, а $\beta_{j'}$ велик по абсолютному значению и отрицателен:

так как X_j и $X_{j'}$ сильно коррелируют, то вклад $\beta_j x_j$ компенсируется $\beta_{j'} x_{j'}$.

Таким образом, добавляемое ограничение на величину коэффициентов β_j должно предотвратить подобую ситуацию.

1. Коэффициент β_0 не участвует в сумме $\sum_{j=1}^{d} \beta_j^2$:

$$\sum_{i=1}^{N} \left(\beta_0 + \sum_{j=1}^{d} x_j^{(i)} \beta_j - y^{(i)} \right)^2 \to \min_{\beta}, \qquad \text{при ограничениях} \qquad \sum_{j=1}^{d} \beta_j^2 \leq s$$

2. Как правило, y центрируют, что позволяет исключить из модели коэффициент β_0:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d$$

Ридж-регрессия:

$$\sum_{i=1}^{N} \left(\sum_{j=1}^{d} x_j^{(i)} \beta_j - y^{(i)} \right)^2 \to \min_{\beta}, \qquad \text{при ограничениях} \qquad \sum_{j=1}^{d} \beta_j^2 \leq s$$

3. Как правило, **X** центрируют и нормируют на среднеквадратическое отклонение, чтобы вклад в сумму $\sum_{j=1}^{d} \beta_j^2$ каждого признака по возможности был бы одинаковым

Коэффициент β_0 может быть устранен из штрафной функции, так как он не соответствует ни одной переменной x_j.

Слагаемое β_0 можно выкинуть и из целевой функции.

Действительно, решение задачи наименьших квадратов можно разбить на два этапа:

1. *Центрирование данных*:
   - каждое $x_j^{(i)}$ заменяется на $x_j^{(i)} - \overline{x}_j$ $(i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, d)$
   - в качестве β_0 выбирается $\overline{y}$, где

$$\overline{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}, \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)}.$$

2. Гребневая регрессия (без коэффициента β_0)

Пусть центрирование произведено, следовательно, $\mathbf{X}$ имеет d (а не $d + 1$) столбцов

$$\mathrm{RSS}^{\mathrm{ridge}}(\beta, \lambda) = (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^\top\beta,$$

$$\mathrm{RSS}^{\mathrm{ridge}}(\beta, \lambda) \to \min_\beta$$

Дифференцируя, находим:

$$\frac{\partial\,\mathrm{RSS}^{\mathrm{ridge}}}{\partial\beta} = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta) + 2\lambda\beta, \qquad \frac{\partial^2\mathrm{RSS}^{\mathrm{ridge}}}{\partial\beta\,\partial\beta^\top} = 2\mathbf{X}^\top\mathbf{X} + 2\lambda\mathbf{I}.$$

откуда

$$\widehat{\beta}^{\mathrm{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \qquad (\star)$$

Итак, решение $\widehat{\beta}^{\mathrm{ridge}}$ задачи гребневой линейной регрессии также является функцией, линейно зависящей от $\mathbf{y}$.

Если $\lambda > 0$, то матрица $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ невырождена (положительно определена), даже если столбцы в $\mathbf{X}$ линейно зависимы

$(\star)$ представляет собой стандартную *регуляризацию*

Итак, от системы $\mathbf{X}\beta = \mathbf{y}$ мы перешли к $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}^\top\mathbf{y}$

6.2.3. Отступление. Регуляризация (А.Н. Тихонов)

Система $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}^\top\mathbf{y}$ называется *регуляризованной* к $\mathbf{X}\beta = \mathbf{y}$.

λ — *параметр регуляризации*, $\lambda > 0$.

Обозначим $\widehat{\beta}(\lambda)$ — (единственное) решение системы $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}^\top\mathbf{y}$.

Теорема 6.1 (А.Н. Тихонов)

Пусть λ_n — некоторая последовательность, $\lambda_n > 0$ и $\lambda_n \to 0$.

Тогда $\widehat{\beta}(\lambda_n)$ стремится к нормальному псевдорешению системы $\mathbf{X}\beta = \mathbf{y}$

(из всех псевдорешений нормальное псевдорешение имеет минимальную норму).

Если λ близко к нулю, то регуляризованная система близка к вырожденной, поэтому на практике решение $\widehat{\beta}(\lambda)$ будет найдено с большой ошибкой.

Метод регуляризации заключается в решении регуляризованных систем для конечного числа значений λ_n и дальнейшем выборе того решения, для которого норма невязки минимальна.

6.2.4. Гребневая регрессия и сингулярное разложение

Выразим $\widehat{\beta}^{\text{ls}}$ и $\widehat{\mathbf{y}}^{\text{ls}}$ (без регуляризации), используя *SVD*:

$$\widehat{\beta}^{\text{ls}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \left((\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top)^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top\right)^{-1}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top)^\top\mathbf{y} =$$

$$= \left(\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top\right)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} = (\mathbf{V}^\top)^{-1}\boldsymbol{\Sigma}^{-2}\mathbf{V}^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top\mathbf{y}$$

$$\widehat{\mathbf{y}}^{\text{ls}} = \mathbf{X}\widehat{\beta}^{\text{ls}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top\mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top\mathbf{y} = \mathbf{U}\mathbf{U}^\top\mathbf{y} = \sum_{j=1}^{d}\mathbf{u}_j\mathbf{u}_j^\top\mathbf{y}$$

Заметим, что $\mathbf{u}_j(\mathbf{u}_j^\top\mathbf{y})$ есть проекция вектора $\mathbf{y}$ на вектор $\mathbf{u}_j$, а вся сумма (как мы уже видели) есть проекция вектора $\mathbf{y}$ на подпространство, натянутое на $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d$.

Аналогично, если проводится регуляризация:

$$\widehat{\mathbf{y}}^{\text{ridge}} = \mathbf{X}\widehat{\beta}^{\text{ridge}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{U}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} = \sum_{j=1}^{d}\mathbf{u}_j\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\mathbf{u}_j^\top\mathbf{y}$$

$$\widehat{\mathbf{y}}^{\text{ridge}} = \mathbf{U}\mathbf{\Sigma}(\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{y} = \sum_{j=1}^{d} \mathbf{u}_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda}\mathbf{u}_j^\top\mathbf{y}$$

Итак, как и для линейной регрессии, линейная гребневая регрессия вычисляет проекцию $\mathbf{u}(\mathbf{u}_j^\top\mathbf{y})$ вектора $\mathbf{y}$ на вектор $\mathbf{u}_j$, но затем домножает эту проекцию на

$$\frac{\sigma_j^2}{\sigma_j^2 + \lambda} \leq 1.$$

Чем больше σ_j, тем ближе этот множитель к 1.

Чем меньше σ_j, тем ближе этот множитель к 0.

Таким образом, наибольшему уменьшению подвергаются компоненты, соответствующие меньшим сингулярным числам σ_j.

6.2.5. Лассо (R. Tibshirani)

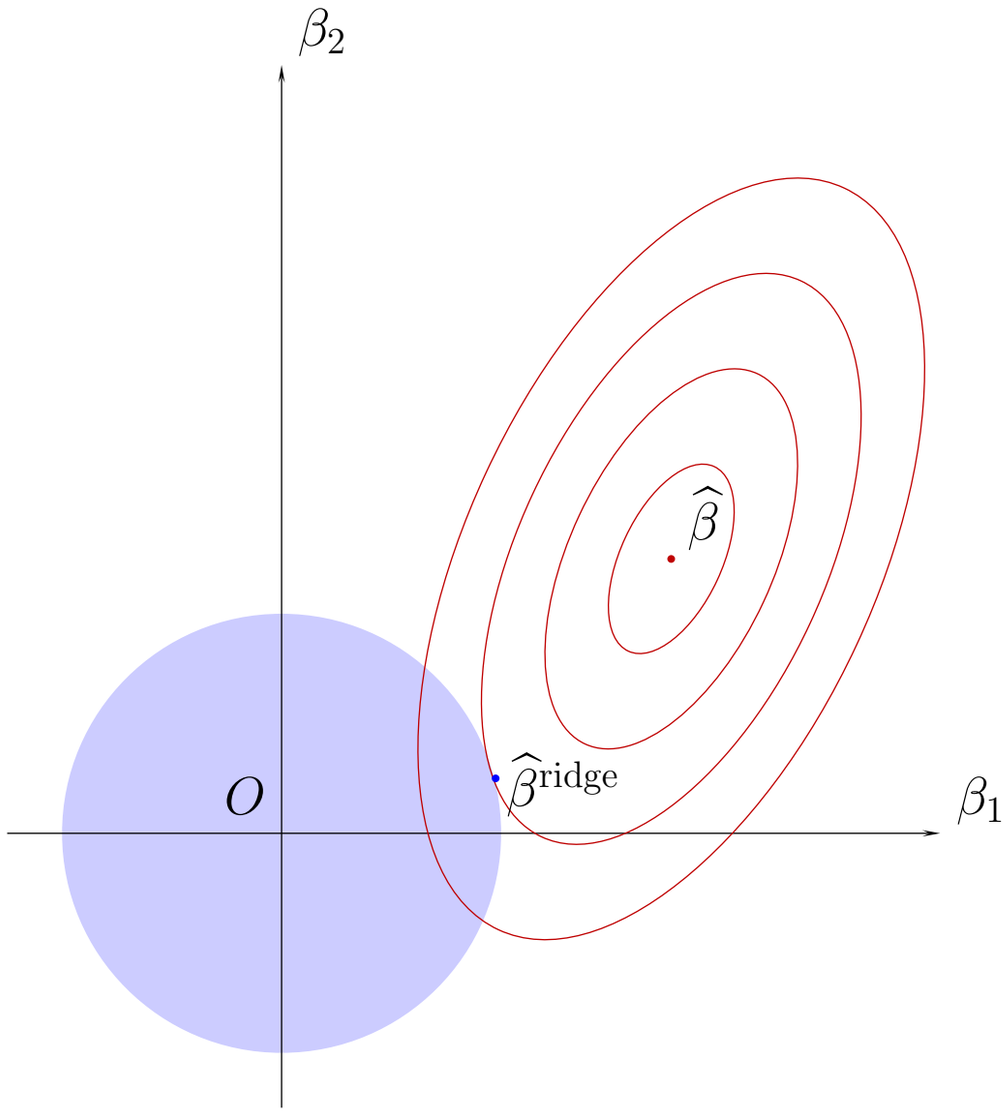
LASSO — «least absolute shrinkage and selection operator»

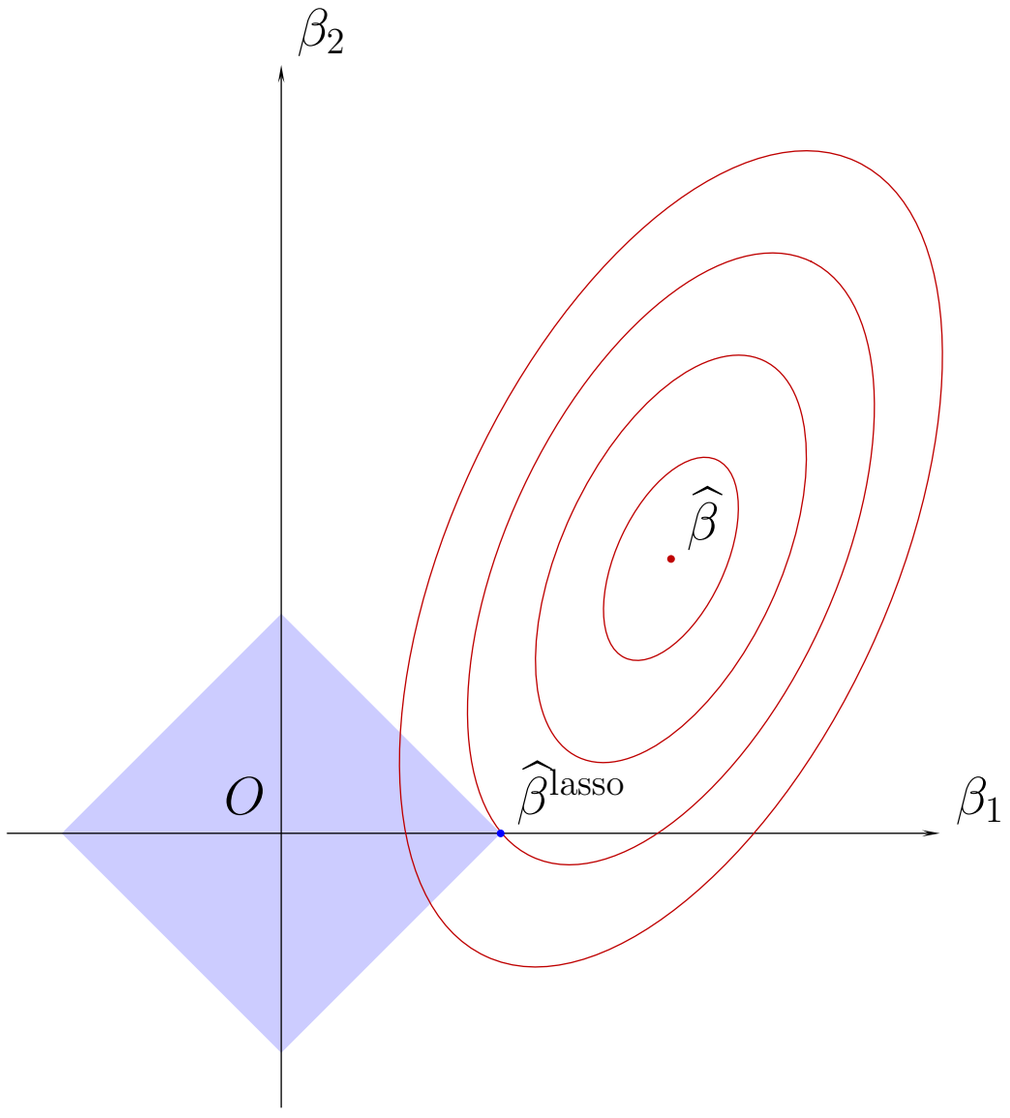
$$\widehat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y^{(i)} - \beta_0 - \sum_{j=1}^{d} x_j^{(i)} \beta_j \right)^2 \right\}, \text{ при условии } \sum_{j=1}^{d} |\beta_j| \leq s.$$

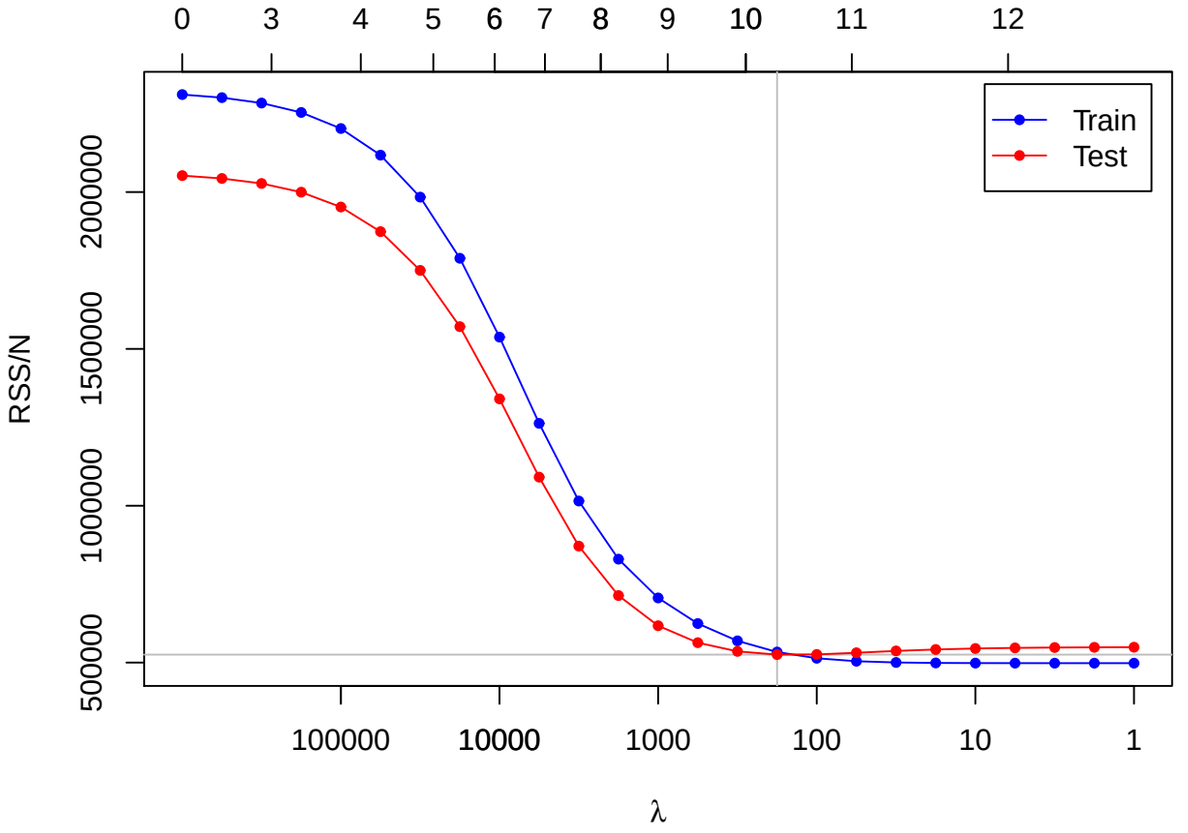
Если s достаточно мало, то в типичной ситуации часть коэффициентов $\widehat{\beta}_j^{\text{lasso}}$ равна в точности 0.

Почему?

(При гребневой регрессии такого, как правило, не происходит)


β_2
$\widehat{\beta}$
O
$\widehat{\beta}^{\text{ridge}}$
β_1

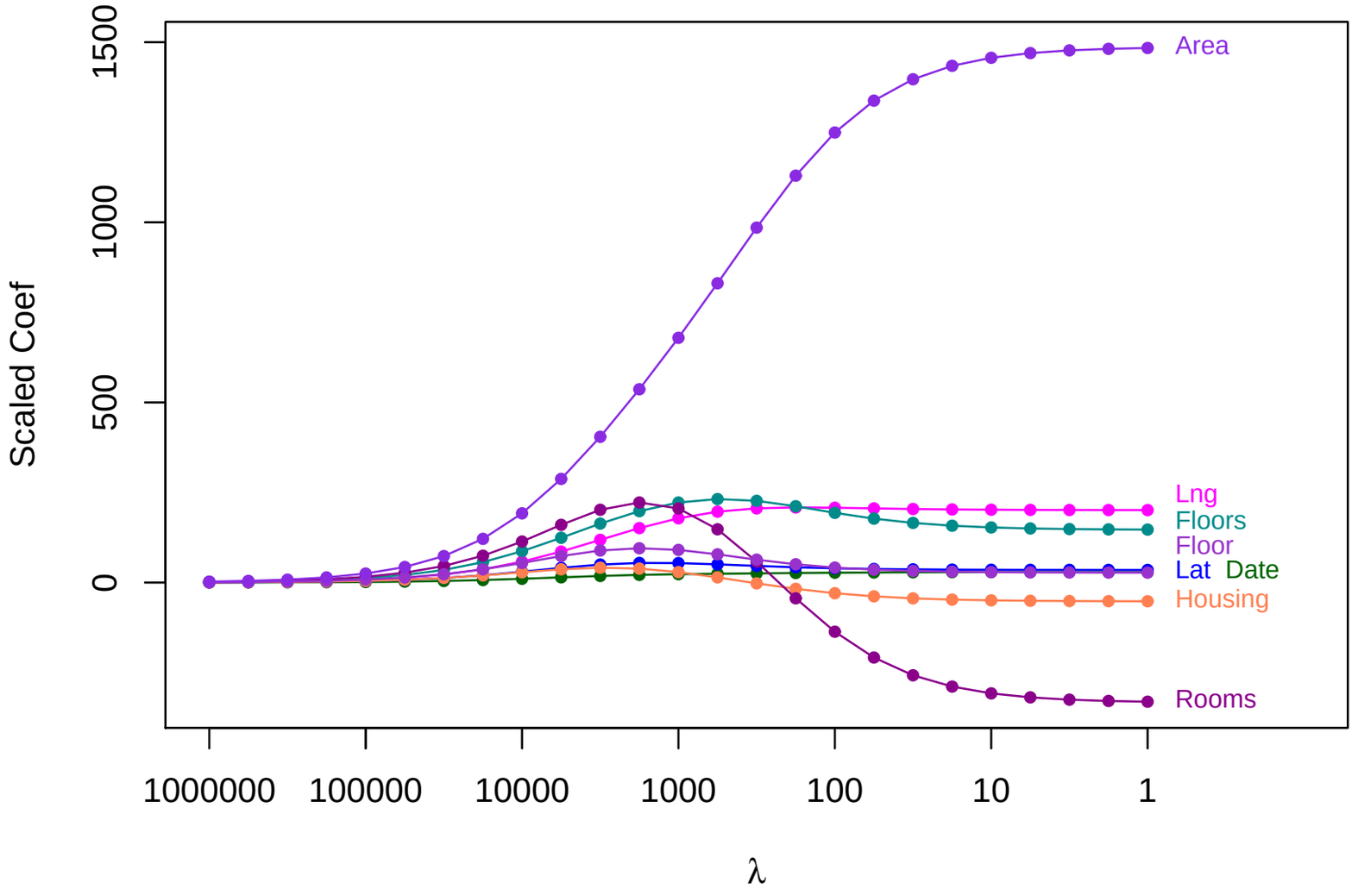

β_2
$\widehat{\beta}$
O
$\widehat{\beta}^{\text{lasso}}$
β_1



$\lambda = 177.8279 = 10^{2.25}$

$$y = 3226.1 + 0.30465\,\mathsf{Date} + 940.46\,\mathsf{Lat} + 2314.9\,\mathsf{Lng} - 53.574\,\mathsf{Housing} + 41.265\,\mathsf{Floors} - 48.654\,\mathsf{Rooms} + 13.105\,\mathsf{Floor} + 56.764\,\mathsf{Area}$$

$$\widehat{R}_{\text{train}} = \frac{1}{N_{\text{train}}}\,\text{RSS}_{\text{train}} = 533652.8, \qquad \widehat{R}_{\text{test}} = \frac{1}{N_{\text{test}}}\,\text{RSS}_{\text{test}} = 525341.3$$

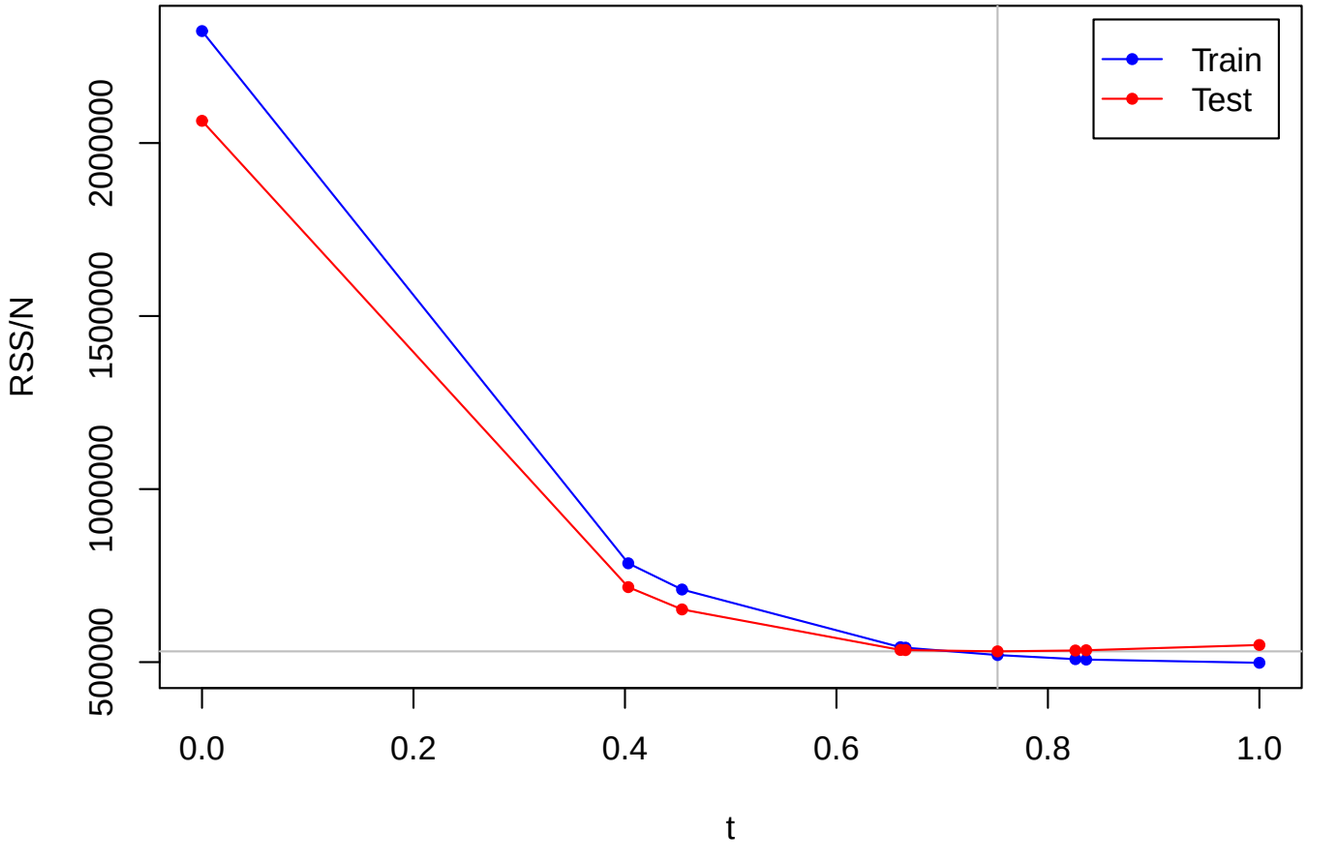

Scaled Coef
1500
1000
500
0
Area
Lng
Floors
Floor
Lat
Date
Housing
Rooms
1000000
100000
10000
1000
100
10
1
λ

Лассо. Введем параметр

$$t = \frac{s}{\sum_{j=1}^{d} |\widehat{\beta}_j|}$$

При $t = 0$ все $\widehat{\beta}_j^{\text{lasso}}$ равны нулю.

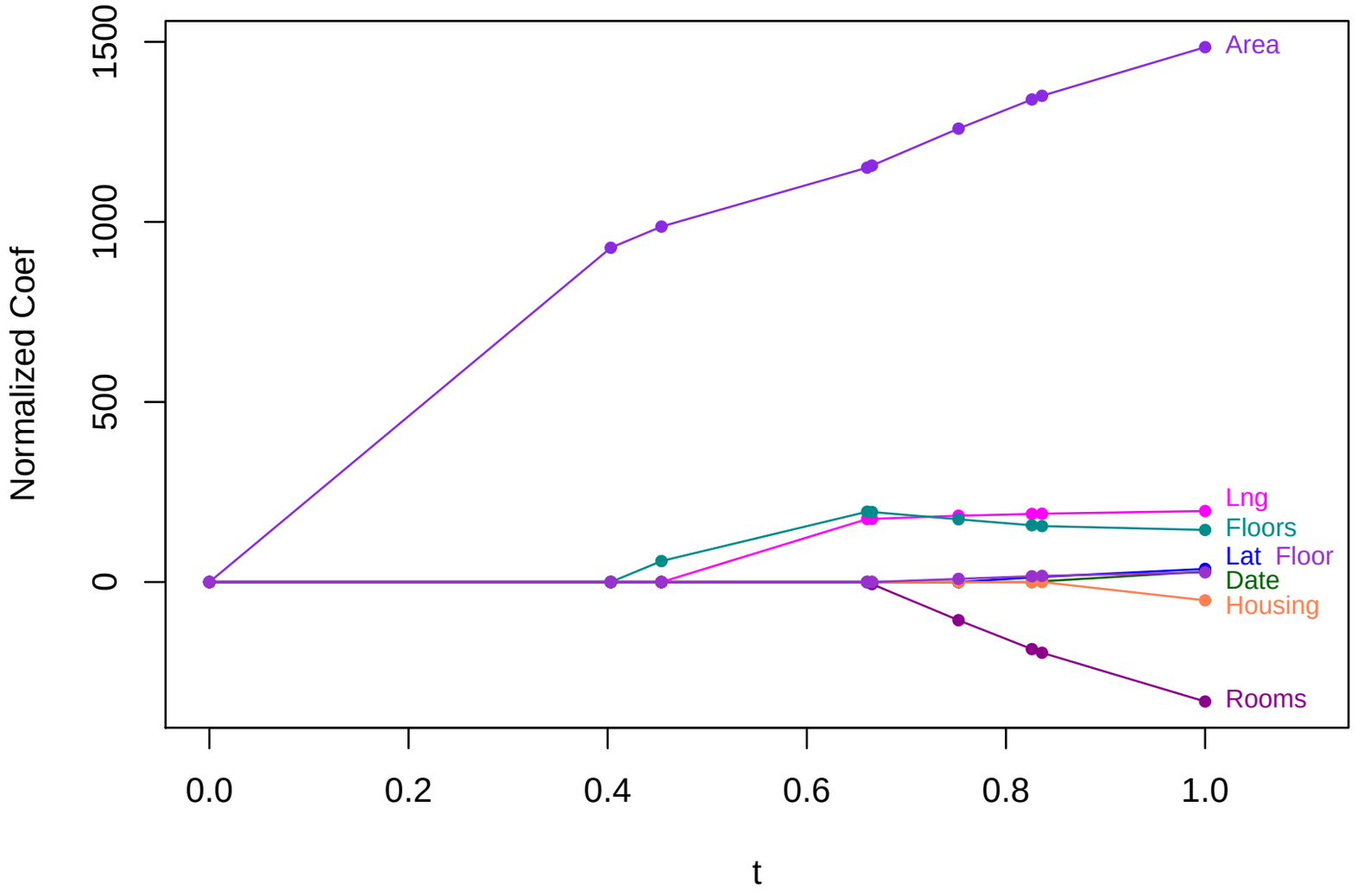
При $t = 1$ имеем $\widehat{\beta}_j^{\text{lasso}} = \widehat{\beta}_j$ $(j = 1, 2, \ldots, d)$.



$t = 0.7523712$

$$y = 3226.1 + 2083.4\,\mathsf{Lng} + 34.318\,\mathsf{Floors} - 118.41\,\mathsf{Rooms} + 2.3186\,\mathsf{Floor} + 63.356\,\mathsf{Area}$$

$$\widehat{R}_{\text{train}} = \frac{1}{N_{\text{train}}}\,\text{RSS}_{\text{train}} = 520436.5, \qquad \widehat{R}_{\text{test}} = \frac{1}{N_{\text{test}}}\,\text{RSS}_{\text{test}} = 531102.9$$


Normalized Coef
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
t
Area
Lng
Floors
Lat
Floor
Date
Housing
Rooms