Глава 4

Линейная регрессия

Мы с вами научились обнаруживать наличие зависимости факторов, однако, можем ли мы предсказывать одни параметры на основе других и описывать эту зависимость?

Рассмотрим набор точек (X_i, Y_i), где Y — зависимая переменная, зависящая от координат вектора X, которые называют предикторами. Задача регрессионного анализа — оценить функцию $f(x) = \mathbf{E}(Y|X = x)$. Таким образом, $Y_i = f(X_i)+\varepsilon_i$, где $\mathbf{E}(\varepsilon|X) = 0$. В рамках первых тем мы будем рассматривать случай, когда X является вектором, причем $f(x)$ – линейная функция x.

Введем несколько важных параметров, которые понадобятся нам в дальнейшем. Пусть $\widehat{f}(x)$ – оценка функции регрессии. Тогда величины $r_i = Y_i - \widehat{f}(X_i)$. Если мы в точности безупречно оценили бы $f(x)$, то это были бы ε_i. А так это, вообще говоря, зависимые величины с различным распределением.

Величина $RSS = \sum_{i=1}^{n} r_i^2$ называется остаточной суммой квадратов.

4.1 Линейная регрессия. Нормальная модель

Начнем со случая, когда ε_i предполагаются н.о.р. $\mathcal{N}(0, \sigma^2)$ и независящими от X. Для простоты будем предполагать X независящими от ε. Тем самым, можно фиксировать X и рассматривать его как вектор параметров: распределение $\mathbf{P}(\varepsilon \in \cdot|\vec{X} = \vec{x})$ не зависит от x, а величины ε_i при этом остаются н.о.р.

Эту задачу мы рассматривали в курсе статистического анализа.

4.1.1 Простая линейная регрессия

Самой простой формой является случай, когда предиктор один и в этом случае

$$Y_i = b_0 + b_1 X_i + \varepsilon_i,$$

где X_i — случайные или детерминированные величины, $\mathbf{E}(\varepsilon_i|X_i) = 0$, $\mathbf{D}(\varepsilon_i|X_i) = \sigma^2$ — константа, а ε_i независимы при условии X. В данном случае мы предполагаем нормальность ε_i.

В этом случае нетрудно построить ОМП для параметров b_0, b_1, ε и убедиться, что

- ОМП для b_0, b_1 получается минимизацией $RSS = \sum_{i=1}^{n} r_i^2$, то есть мы выбираем такие оценки b_0 и b_1, что сумма квадратов остатков минимальна. Такие оценки называются МНК – оценками методом наименьших квадратов.

- Непосредственно оценки выглядят как

$$\widehat{b_1} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2},\ \widehat{b_0} = \overline{Y} - \overline{X}\widehat{b_1}.$$

- ОМП для σ^2 представляет собой RSS/n.

- Оценки $\widehat{b_0}$, $\widehat{b_1}$ являются несмещеными, состоятельными и асимптотически нормальными с ковариационной матрицей (при условии $X_1, ..., X_n$)

$$\sigma^2\Sigma = \frac{\sigma^2}{nS_X^2}\begin{pmatrix} \overline{X^2} & -\overline{X} \\ -\overline{X} & 1 \end{pmatrix},$$

  и не зависят от RSS, имеющей распределение χ^2_{n-2}.

Отсюда нетрудно построить асимптотические доверительные интервалы для b_0, b_1:

$$b_0 \in \left(\widehat{b_0} - \frac{\sqrt{RSS\overline{X^2}}t_{1-\alpha/2}}{S_X\sqrt{(n-2)n}}, \widehat{b_0} + \frac{\sqrt{RSS\overline{X^2}}t_{1-\alpha/2}}{S_X\sqrt{(n-2)n}}\right),$$
$$b_1 \in \left(\widehat{b_1} - \frac{\sqrt{RSS}t_{1-\alpha/2}}{S_X\sqrt{(n-2)n}}, \widehat{b_1} + \frac{\sqrt{RSS}t_{1-\alpha/2}}{S_X\sqrt{(n-2)n}}\right),$$

где t_β – квантиль распределения Стьюдента с $n-2$ степенями свободы. Можно построить асимптотическое доверительное множество (эллипс) для пары параметров

$$\frac{n-2}{RSS}(\widehat{b_0} - b_0, \widehat{b_1} - b_1)\Sigma^{-1}(\widehat{b_0} - b_0, \widehat{b_1} - b_1)^t \le f_{1-\alpha},$$

где f – квантиль распределения Фишера-Снедекора с $2, n-2$ степенями уровня $1-\alpha$

4.1.2 Общая линейная модель для нормальных данных

В более общем случае предполагается, что регрессоров несколько. Таким образом, у нас есть матрица X с n строками и k столбцами. Тогда

$$\vec{Y} = X\vec{a} + \vec{\varepsilon},$$

где ε_i – н.о.р. $\mathcal{N}(0, \sigma^2)$ величины. Также допустим, что X^tX невырождена (иначе можно удалить часть линейно зависимых предикторов).

Обратите внимания, что столбцы матрицы X (то есть предикторы, на основе которых мы оцениваем Y) не являются, вообще говоря, независимыми переменными. Например,

мы можем взять $X_{\cdot,1} = (x_1, \ldots, x_n)$, $X_{\cdot,2} = 1$ и получим простую линейную модель из прошлого раздела. А можем взять $X_{\cdot,1} = (x_1^2, \ldots, x_n^2)$, $X_{\cdot,2} = (x_1, \ldots, x_n)$ и $X_{\cdot,3} = (1, \ldots, 1)$ и получим квадратичную модель, в которой Y с точностью до погрешности представляет собой

В этом случае мы также можем составить правдоподобие и получить

$$L(X, \vec{Y}, \vec{a}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}||\vec{Y} - X\vec{a}||^2\right).$$

Максимизация правдоподобия путем дифференцирования для $\vec{a}$ опять же приводит к оценкам методом наименьших квадратов, то есть

$$||\vec{Y} - X\vec{a}||^2 \to \min.$$

Это приводит к оценкам

$$\widehat{a} = (X^tX)^{-1}X^t\vec{Y}, \quad \widehat{\sigma^2} = \frac{1}{n}RSS, \quad RSS = \sum_{i=1}^{n} ||\vec{Y} - X\widehat{a}||^2, \tag{4.1}$$

Несмещенная оценка для дисперсии будет использовать в (4.1) коэффициент $1/(n-k)$ перед RSS вместо $1/n$.

При этом

$$\widehat{a} \sim \mathcal{N}(\vec{a}, (X^tX)^{-1}\sigma^2).$$

Геометрически задача максимизации правдоподобия по $\vec{a}$ представляет собой поиск вектора вида $X\vec{a}$, наиболее близкого к $\vec{Y}$.

Решением данной задачи является вектор $X\vec{a}$, являющейся проекцией $\vec{Y}$ на пространство $L = \{X\vec{a}, \vec{a} \in \mathbb{R}^k\}$. Первая формула в (4.1) представляет собой простое следствие из того, что оператор проекции на такое пространство имеет вид $X(X^tX)^{-1}X^t$. Соответственно, RSS представляет собой квадрат длины проекции вектора $\vec{Y}$ на пространство, ортогональное к L.

Отсюда, вспоминая конец курса "Статистический анализ данных", мы опять же можем получить, что $RSS/\sigma^2 \sim \chi^2_{n-k}$, причем RSS не зависит от коэффициентов $\vec{a}$ проекции X на L.

Таким образом,

$$\frac{1}{\sigma^2}RSS \sim \chi^2_{n-k}, \quad \frac{\widehat{a} - \vec{a}}{\sigma} \sim \mathcal{N}(0, (X^tX)^{-1}),$$

откуда

$$\frac{1}{\sigma^2}(\widehat{a} - \vec{a})^tX^tX(\widehat{a} - \vec{a}) \sim \chi^2_k, \quad \frac{(\widehat{a} - \vec{a})^tX^tX(\widehat{a} - \vec{a})/k}{RSS/(n-k)} \sim F_{k,n-k},$$

где F – распределение Фишера-Снедекора.Этот вывод опирается на известный нам из первого семестра факт о том, что величина $\vec{X}\Sigma^{-1}\vec{X}^t$, где $X \sim \mathcal{N}(0, \Sigma)$, распределена по закону χ^2_d, где d – размерность вектора. Отсюда может быть построено доверительное множество для параметра $\vec{a}$.

В Python есть несколько вариантов запуска линейной регрессии. Можно использовать pandas и LinearRegression(). Ее нужно сперва инициализировать в некоторой переменной, а затем применять к этой переменной метод fit, указав регрессоры и зависимую переменную. Можно использовать statmodels.formula.api (будем для краткости называть его sm), у которого запустить sm.OLS(y,x).fit(). Метод summary() от результата даст анализ модели.

Метод predict позволяет предсказать значение функции $f(x)$ в точке x.

4.2 Линейная регрессия без предположения нормальности

4.2.1 Свойства МНК оценок

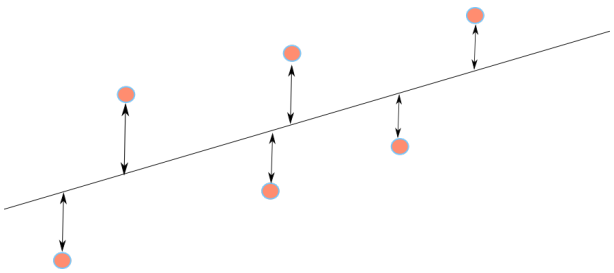
В прошлом разделе мы предполагали, что ε_i не зависят от X, н.о.р. и имеют $\mathcal{N}(0, \sigma^2)$ распределение. Начнем с того, что откажемся от условия нормальности, сохранив условие $\mathbf{E}\varepsilon = 0$. Итак, пусть

$$\vec{Y} = X\vec{a} + \vec{\varepsilon}.$$

Рассмотрим те же оценки МНК, полученные минимизацией

$$||\vec{Y} - X\vec{a}|| \to \min.$$

Это уже не ОМП, но все же некоторые оценки. В случае одного предиктора X мы будем минимизировать сумму квадратов расстояний по вертикали от Y до прямой aX. Отметим, что задача принципиально асимметрична, прямая, которая строилась бы для



зависимой переменной X с регрессором Y, была бы другой.

Теперь оценка МНК уже не является ОМП, однако, из тех же соображений, что и раньше остается несмещенной. Более того, она обладает следующим замечательным свойством:

Теорема 1 (Теорема Гаусса-Маркова)**.** *Оценка МНК имеет минимальную дисперсию среди несмещенных оценок вида* $a_1(X)Y_1 + \cdots + a_n(X)Y_n$.

Оптимальности в классе всех оценок достигнуть не удается. Для нормального распределения ε_i из соображений регулярности МНК-оценка как ОМП будет асимптотически эффективной (в действительности, в нормальной модели она оптимальна в классе всех оценок). Для, например, распределения Лапласа ε_i с плотностью $\exp(-|x|)$ ОМП, опять

же будет асимптотически эффективной. Однако, это уже будет не оценка МНК, а оценка, минимизирующая абсолютную сумму остатков

$$\sum_{i=1}^{n} |y_i - (X\vec{a})_i| \to \min .$$

4.2.2 Асимптотические свойства оценок МНК

Можем ли мы утверждать о том, что оценка МНК $\widehat{a}$ будет состоятельной? Задача осложняется тем, что с ростом n меняется матрица X, а значит нам нужно описать изменение матрицы X. Зато предполагая случайность, независимость и одинаковую распределенность строк X_i, мы автоматически получаем сильные асимптотические свойства.

Зачастую случайность предикторов – естественное предположение, вытекающее из устройства эксперимента. Например, мы можем применять линейную модель для изучения характера зависимости некоторых параметров (например, роста и веса) различных объектов, которые естественно предполагать независимыми для разных объектов.

При этом нам даже не требуется чтобы ε_i были независимыми от X_i. Наложим следующие условия:

1. Векторы (X_1, ε_1) н.о.р.;
2. Ранг $Q = \mathbf{E}X^tX$ равен k;
3. $\mathbf{E}X_{1,i}\varepsilon_1 = 0$;
4. $\mathbf{E}\varepsilon_1^2 < \infty$;
5. $\mathbf{E}\varepsilon_1^4 < \infty$, $\mathbf{E}X_{1,i}^4 < \infty$.

Теорема 2.
- *В условиях 1)-4) оценка МНК $\widehat{a}$ состоятельна для $\vec{a}$.*
- *В условиях 1)-5) оценка МНК $\widehat{a}$ асимптотически нормальна для $\vec{a}$ с асимптотической матрицей ковариации $S = Q^{-1}RQ^{-1}$, где $R_{i,j} = \mathbf{E}X_{1,i}X_{1,j}\varepsilon_1^2$.*

В случае, если $\mathbf{E}(\varepsilon^2|X) = \sigma^2$, $R = Q\sigma^2$ и $S = Q^{-1}\sigma^2$, то есть $\widehat{a}$ независимы с ковариационной матрицей $Q^{-1}\sigma^2$.

Матрица S нам неизвестна, однако, ее можно состоятельно оценить матрицей

$$\widehat{Q}^{-1}\widehat{R}\widehat{Q}^{-1}, \quad \widehat{Q} = \frac{1}{n}X^tX, \quad \widehat{R} = \frac{1}{n}X^t\widehat{D}X,$$

а $\widehat{D}$ – диагональная матрица $n \times n$ с $Y_l - \langle X_{l,\cdot}, \widehat{a}\rangle$ на диагонали.

Впрочем, в случае $\mathbf{E}(\varepsilon|X) = 0$ достаточно использовать $\widehat{Q}^{-1}$.

Величина σ^2 при этом состоятельно оценивается RSS/n, откуда

$$\frac{n}{RSS}(\vec{Y} - X\widehat{a})^t\widehat{S}^{-1}(\vec{Y} - X\widehat{a}) \to Y \sim \chi_k^2, \ n \to \infty$$

Отсюда мы получаем все тот же асимптотический доверительный эллипс для наших параметров в случае $\mathbf{E}(\varepsilon^2|X) = const$.

Итак, в случае ненормальных ошибок оценка МНК будет состоятельной и асимптотически нормальной при достаточно общих условиях на МНК.

4.3 Лассо и ридж-регрессия

4.4 Регрессия с регуляризацией

Проблема с отбором признаков возникает из-за того, что $||y - X\vec{b}||$ увеличивается с ростом размерности пространства признаков. Мы можем показывать, что это увеличение незначительно (сравнимо с шумом), как мы это делали в прошлой темой. Альтернативой является замена $||y - X\vec{b}||$ на

$$||y - X\vec{b}||^2 + h(\vec{b}),$$

где h – некоторая функция штрафа за вектор коэффициентов. Это может быть, например, норма или полунорма, которая увеличивается с ростом коэффициентов.

Отметим две ключевые проблемы, которые при этом можно решить:

- если часть предикторов почти коллинеарна, то проблема возникает с тем, что разложить вектор по $\vec{x}_1, \vec{x}_2$ можно несколькими способами практически с одинаковой эффективностью (скажем, если $\vec{x}_1 \approx \vec{x}_2$, то $\vec{y} \approx \vec{x}_1 \approx 100\vec{x}_1 - 99\vec{x}_2$), то разложение выбирается из соображений большей коррелированности погрешности с остатком.
- если часть предикторов незначительно связана с зависимой переменной, то коэффициенты будут ненулевыми, хотя и небольшими.

За засчет штрафа h мне а) будет невыгодно выставлять большие коэффициенты для коллинеарных предикторов, если это не приводит к существенному уменьшению ошибки б) будет выгодно обнулить или поставить еще более маленьким коэффициент при малозначительном предикторе.

Базовых процедуры в этом направлении две – ридж-регрессия и лассо-регрессия.

4.4.1 Ридж-регрессия и Лассо-регрессия

Пусть размерность множества параметров есть k, а количество наблюдений n.

Рассмотрим дополнительное ограничение $g(\vec{b}) \leq c$, где g — некоторая неперерывная функция, c — заданная константа. Такая задача называется регрессией со сжатием (shrinkage). В этом случае, если $\widehat{\beta}$ удовлетворяет этому условию, то мы оставим оценку МНК, а если нет, то сместим ее в сторону искомого множества. При этом

$$r(\vec{b}) = ||\vec{y} - X\vec{b}||^2 = ||y - X\widehat{\beta}||^2 + (\vec{\beta} - \widehat{\beta})^t X^t X(\vec{\beta} - \widehat{\beta}),$$

что нетрудно получить из формулы квадрата суммы (здесь $||||$ — стандартная норма $\mathbb{R}^n$).

Тем самым линии уровня нашей разности квадратов (то есть кривые $\vec{b}$ на которых $r(\vec{b}) = const$) представляют собой эллипсоиды с центром в $\widehat{\beta}$. Мы должны найти как можно меньший из семейства гомотетичных эллипсоидов, имеющий общие точки с множеством $g(\vec{b}) = c$. Рассмотрим два наиболее популярных примера

Пример 1. Ридж-регрессия (или гребневая регрессия) соответствует классическому расстоянию в $\mathbb{R}^n$

$$g(\vec{b}) = ||b||_2^2 = \sum b_i^2.$$

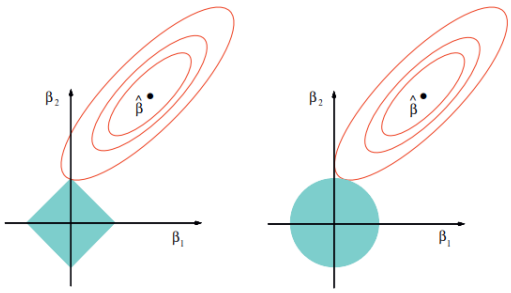
В таком случае наша геометрическая задача превращается в задачу о поиске точке касания эллипсоида и шара.

Пример 2. Лассо-регрессия соответствует манхэттанской метрике, то есть расстоянию в $\mathbb{R}^n$, заданным

$$g(\vec{b}) = ||b||_1 = \sum |b_i|.$$

В таком случае наша геометрическая задача превращается в задачу о поиске точке пересечения эллипсоида и куба. Зачастую такой точкой будет одна из вершин куба.

В обоих случаях может быть осмысленно стандартизировать данные перед применением, иначе мы будем складывать $|b_i|$ разного масштаба.



В случае лассо-регрессии мы обнулим часть коэффициентов, а в случае ридж-регрессии сможем лишь сделать их достаточно малыми.

4.4.2 Решения задач оптимизации в ридж и лассо регрессии

Ниже приведен вывод коэффициентов, который не является необходимым. Вы можете сразу заглянуть в конец раздела, где получены результаты. Задачи можно переписать в форме множителей Лагранжа — то есть оптимизации

$$(\widehat{\beta} - \vec{b})^t X^t X (\widehat{\beta} - \vec{b}) + \lambda(g(\vec{b}) - c)$$

по λ и $\vec{b}$ (правда здесь может возникнуть проблема с лассо-регрессией, где $g(x)$ не является гладкой). Пусть λ дает минимум этого выражения (отметим, что при этом λ положительна). Тогда при данном λ мы будем минимизировать выражение

$$||\widehat{\beta} - \vec{b}||^2 + \lambda g(\vec{b})$$

где $\lambda \geq 0$.

Для ридж-регрессии перейдем в диагонализирующий X^tX базис, это не изменит нормы в L^2 вектора $\vec{b}$ в силу ортогональности преобразования. Дифференцированием по b_i мы получим в качестве $\vec{b}$ оценку

$$\widehat{\beta}_{Ridge} = \frac{d_{i,i}}{d_{i,i} + \lambda}\widehat{\beta}.$$

Поэтому мы видим что при $\lambda > 0$ коэффициенты $\widehat{\beta}_{Ridge}$ уменьшаются, причем маленькие коэффициенты уменьшаются сильнее. Это позволяет практически обнулить маленькие коэффициенты.

В лассо-регрессии выведем их для случая, когда мы находимся в диагонализирующем X^tX базисе (увы, переход в такой базис меняет штрафное слагаемое). Чтобы решить проблемы с недифференцируемости, мы можем зафиксировать знаки всех b_i. Введем матрицу T на диагонали которой стоят числа ± 1, соответствующие этим знакам. Тогда наша задача переписывается в виде

$$(\widehat{\beta} - \vec{b})^t X^t X(\widehat{\beta} - \vec{b}) + \lambda \vec{e}^t T\vec{b},$$

где $T = (t_{i,i})$, $t_{i,i} = \operatorname{sgn} b_i$, $\vec{e} = (1, 1, ..., 1)$. В базисе в котором X^tX будем диагональной с d_i на диагонали, мы получим

$$\sum_{i=1}^{n}(\widehat{\beta}_i - b_i)^2 d_i + \lambda t_i b_i,$$

откуда дифференцированием получим нули производной в виде

$$b_i = \widehat{\beta} - \lambda t_i/(2d_i).$$

При этом необходимо, что знак каждого из b_i был равен $t_{i,i}$, откуда в силу $\lambda > 0$, $d_i > 0$ получим что $(\lambda t_{i,i})/(2d_{i,i})$ имеют тот же знак, что и $t_{i,i}$. Вычитая такой вектор из $\widehat{\beta}$ я могу оказаться в квадранте, соответствующем $t_{i,i}$, только если $t_{i,i}$ — знак $(\widehat{\beta})_i$. При этом λ должно быть достаточно малым, чтобы $|\lambda| < 2|\widehat{\beta}_i|d_{i,i}$. Если λ таково, что неравенство перестает выполняться, то соответствующий коэффициент b_i обнулится и мы просто исключим его из рассмотрения. Таким образом,

$$(\widehat{\beta}_{Lasso})_i = \operatorname{sgn}(\widehat{\beta}_i)\left(|\widehat{\beta}_i| - \frac{\lambda}{2d_{i,i}}\right)^+.$$

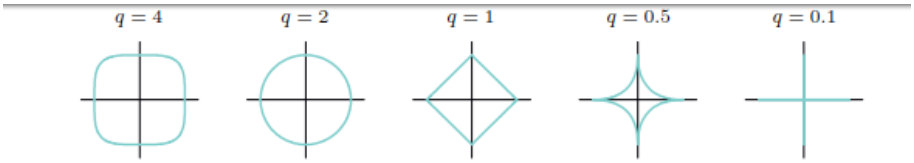
Таким образом, мы уменьшаем коэффициент $\widehat{\beta}_i$ на константу, если же он при этом становится отрицательным, то обнуляем его. За счет этого ридж-регрессия позволяет фильтровать часть предикторов с маленькими коэффициентами, в точности обнуляя их.

4.4.3 Elastic Net и другие обобщения

Более общая форма такого рода задачи (назовем ее L^q-сжатием) предлагает минимизировать

$$||\widehat{\beta} - \vec{b}||^2 + \lambda||\vec{b}||_q^q, \ ||\vec{x}||_q^q = \sum |x_i|^q$$

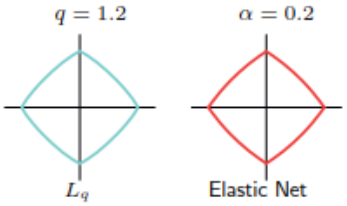
при $q > 0$. Увы, при $q > 1$ мы теряем свойство лассо-регрессии обнулять часть ко-



эффициентов. В связи с этим также используют Elastic Net, в котором предлагается рассматривать ограничения $||\vec{b}||_{L_1} \leq b$, $||\vec{b}||_{L_2} \leq c$, то есть минимизировать функционал

$$(\widehat{\beta} - \vec{b})^t X^t X(\widehat{\beta} - \vec{b}) + \lambda_1(\alpha||\vec{b}||_1/2 + (1-\alpha)||\vec{b}||_2^2).$$

В этом случае Elastic Net наследует качества лассо-регрессии и "застревает в вершинах", то есть обнуляет часть коэффициентов. Важное преимущество Elastic Net заключается



в том, что для сильно коррелирующих предикторов алгоритм будет давать им близкие коэффициенты.

4.5 Выбор параметра

При использовании методов мы должны выбрать параметр λ. Обычно этот параметр используют для адаптации модели к данным. Отметим несколько приемов

- Мы можем разделить данные x_i на обучающую ($i \in L$) и тестовую ($i \in T$) выборки. Теперь мы фиксируем λ, находим по обучающей выборке (x_i, y_i), $i \in L$, коэффициенты $\widehat{\beta}$ и считаем погрешность

$$\sum_{i \in T}(y_i^* - y_i)^2,$$

где $y_i^* = \langle \widehat{\beta}, x_i \rangle$ — прогноз тестовых значений на основе нашей модели. Теперь мы можем варьировать λ и выбирать наиболее оптимальную (то есть дающую минимальную погрешность) модель. В случае лассо-регрессии или Elastic Net часть коэффициентов при этом занулится и мы сможем определить какие переменные мало

значимы (в том числе из-за того, что они почти коллинеарны другим имеющимся переменным).

- Можно использовать вместо деления выборки кросс-валидацию, то есть делить выборку на d частей, строить для каждой из частей прогноз на основе остальных и суммировать сумму квадратов погрешностей прогноза. К сожалению, при этом может оказаться, что вы получите фактически исходную МНК модель, поскольку оценка МНК будет давать маленькую квадратичную ошибку.

- BIC (Bayesian Information Criteria) предлагает минимизировать

$$\frac{1}{n}||y - X\vec{b}||^2 + \frac{k \ln n}{n},$$

AIC (Akaike Information Criteria) предлагает минимизировать

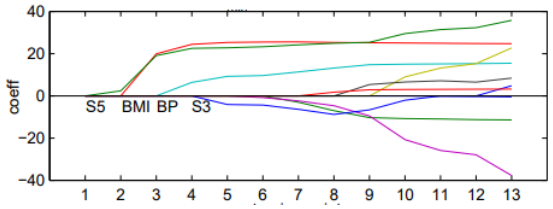
$$\frac{1}{n}||y - X\vec{b}||^2 + \frac{2k}{n},$$

где k – величина, называемая числом степеней свободы. Для Elastic Net и Lasso Regression в ее качестве можно брать количество ненулевых коэффициентов, а для Ridge берут

$$\frac{n}{RSS}\sum_{i=1}^{n} Cov(y_i, \widehat{y}_i),$$

где $\widehat{y}_i$ – прогноз для y_i, Cov – выборочная ковариация.

- Наконец, можно выбрать параметр λ графически, построив набор коэффициентов в зависимости от λ и отследив момент, когда коэффициенты начнут чрезмерно расти.



В sklearn.linear_model в python есть реализации lasso, ridge и ElasticNet алгоритмов. Более того, все методы обладают CV-модификациями, использующими кросс-валидацию для подбора параметра.

Задача 1. Моделировать данные $Y = 2X^3 + 4X + 5 + \varepsilon$, где $X \sim R[0, 1]$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ для различных σ. Применить линейную регрессию, лассо-регрессию, ридж-регрессию и Elastic Net для описания данных а) квадратичной б) кубической в) зависимостью четвертой степени.