Спецкурс "Регрессия"

Шкляев А.В.

17 января 2019 г.

1 Асимптотические свойства оценок вне нормальной модели

1.1 Регрессия с ограничениями

Итак, мы рассмотрели оценку

$$\widehat{a}_{MD} = \widehat{a} - U^{-1}C^t(CU^{-1}C^t)^{-1}(C\widehat{a} - \vec{r}).$$

Мы доказали, что такая оценка имеет асимптотическую дисперсию

$$S_{MD} = (U^{-1}C^t(CU^{-1}C^t)^{-1}C)S(U^{-1}C^t(CU^{-1}C^t)^{-1}C)^t.$$

Мы остановились на следующей теореме:

Теорема 1. *Пусть U^*, $\vec{a}$ таковы, что $\vec{a}^t S_{MD}(U)\vec{a} \geq \vec{a}^t S_{MD}(U^*)\vec{a}$ при всех положительно определенных матрицах U. Тогда $U^* = S^{-1}$. Соответствующая оценка при этом имеет вид*

$$\widehat{a}_{EMD} = \widehat{a} - \widehat{S}C^t(C\widehat{S}C^t)^{-1}(C\widehat{a} - \vec{r}),$$

а ее асимптотическая ковариационная матрица

$$S_{EMD} = S - SC^t(CSC^t)^{-1}CS.$$

Доказательство. Представим матрицу S в виде $\widetilde{S}\widetilde{S}^t$, пользуясь симметричностью и положительной определенностью. Тогда

$$\vec{a}^t S_{MD}\vec{a} = ||E - (\widetilde{S}^t U\widetilde{S})^{-1}\widetilde{S}^t C^t(C\widetilde{S}(\widetilde{S}^t U\widetilde{S})^{-1}\widetilde{S}^t C^t)^{-1}C\widetilde{S})^t\widetilde{S}\vec{a}||^2.$$

Положим $\widetilde{S}^t U\widetilde{S} = \widetilde{U}$, $\widetilde{C} = C\widetilde{S}$, $\widetilde{a} = \widetilde{S}\vec{a}$. Вектор

$$\vec{b} = \widetilde{U}^{-1}\widetilde{C}^t(\widetilde{C}\widetilde{U}^{-1}\widetilde{C}^t)^{-1}\widetilde{C}\widetilde{\vec{a}}$$

является проекцией вектора $\widetilde{a}$ на пространство $L = \{\widetilde{C}\vec{b}\}$, где скалярное произведение задается

$$\langle x, y\rangle_{\widetilde{U}^{-1}} = x^t\widetilde{U}^{-1}y.$$

При этом

$$\vec{a}^t S_{MD}\vec{a} = ||\widetilde{a} - \vec{b}||^2,$$

где $\widetilde{a} - \vec{b}$ — проекция $\vec{a}$ на $L^\perp$ (в смысле скалярного произведения $\langle x, y\rangle_{\widetilde{U}^{-1}}$). Следовательно, задача поиска оптимального U равносильна поиску такой матрицы скалярного произведения, что проекция (относительно скалярного произведения, заданного матрицей U^{-1}) заданного вектора на заданное пространство имеет наименьшую длину (относительно стандартного скалярного произведения). Но мы знаем какой

вектор в L наилучший с точки зрения минимального расстояния от него до данного вектора $\widetilde{a}$ — это его проекция. Значит, наилучшее U это то, при котором $\vec{b}$ есть обычная ортогональная проекция $\vec{a}$ на L. Очевидно, что это достигается при $\widetilde{U}^{-1} = E$, то есть $U^{-1} = S$ (при этом мы не утверждаем, что только при таком). $\square$

1.2 Погрешность условия

Добавляя условие, мы можем задаться вопросом — а что случится, если условие ошибочно? Пусть $C\vec{a} = \vec{r}^* \neq \vec{r}$. Тогда

$$\widehat{a}_{MD} = \widehat{a} - U^{-1}C^t(CU^{-1}C^t)^{-1}(C\widehat{a} - \vec{r}) =$$
$$\widehat{a} - U^{-1}C^t(CU^{-1}C^t)^{-1}(C\widehat{a} - \vec{r}^*) + U^{-1}C^t(CU^{-1}C^t)^{-1}(\vec{r} - \vec{r}^*)$$

Таким образом,

$$\widehat{a}_{MD} \xrightarrow{P} \vec{a} + U^{-1}C^t(CU^{-1}C^t)^{-1}(\vec{r} - \vec{r}^*).$$

Таким образом, оценка будет асимптотически смещенной. При этом

$$\sqrt{n}(\widehat{a}_{MD} - \vec{a} - U^{-1}C^t(CU^{-1}C^t)^{-1}(\vec{r} - \vec{r}^*)) \xrightarrow{d} Z \sim \mathcal{N}(0, S_{MD})$$

по тем же причинам, что и прежде. Поэтому оценить дисперсию ошибок мы все равно сможем верно, а вот устранить смещение нет.

В случае, если $C\vec{a}_n = \vec{r} + \vec{\delta}n^{-1/2}$ мы получаем

$$\sqrt{n}(\widehat{a}_{MD} - \vec{a}_n) = (E - U_n^{-1}C^t(CU_n^{-1}C^t)^{-1}C)\sqrt{n}(\widehat{a} - \vec{a}_n) - U_n^{-1}C^t(CU_n^{-1}C^t)^{-1}\vec{\delta}.$$

Первое слагаемое как и прежде сходится к $\mathcal{N}(0, S_{MD})$, а второе к константе $\vec{\delta}^* = U^{-1}C^t(CU^{-1}C^t)^{-1}\vec{\delta}$.

2 Регрессия с нелинейными ограничениями

В случае, если ограничения нелинейны $r(\vec{a}) = 0$ и $r(\vec{a}) \geq 0$, то оценки строятся из соображений

$$\min_{r(\vec{a})=0} ||\widehat{a} - \vec{a}||_{U_n}, \ \min_{r(\vec{a})=0} ||\widehat{a} - \vec{a}||_{X^tX}, \ \min_{r(\vec{a})\geq 0} ||\widehat{a} - \vec{a}||_{U_n}, \ \min_{r(\vec{a})\geq 0} ||\widehat{a} - \vec{a}||_{X^tX},$$

соответственно.

2.1 Ридж-регрессия и лассо-регрессия

Рассмотрим ограничение $||a||_C^2 \leq b$, где C — заданная положительно определенная матрица, b — заданная константа. Для общности будем считать, что мы накладываем такого рода ограничения на оценку методом наименьших расстояний, хотя чаще всего в качестве оценки рассматривают обычную оценку МНК.

Тогда метод множителей Лагранжа приводит нас к задаче оптимизации

$$||\widehat{a}_{MD} - \vec{a}||_U + \lambda(||\vec{a}||_C^2 - b),$$

где $\lambda \geq 0$. Предположим, что λ мы уже нашли и $\lambda > 0$ (иначе $\widehat{a}_{MD}$ уже удовлетворяла дополнительным условиям). Тогда наша задача сводится к минимизации

$$||\widehat{a}_{MD} - \vec{a}||_U^2 + \lambda||\vec{a}||_C^2 = \widehat{a}_{MD}^t U\widehat{a}_{MD} + \vec{a}^t(U + \lambda C)\vec{a} - \widehat{a}_{MD}^t U\vec{a} - \vec{a}^t U\widehat{a}_{MD}.$$

Дифференцируя по a_i, получаем

$$2\langle (U+\lambda C)_i, \vec{a}\rangle - 2\langle (U)_i, \widehat{a}_{MD}\rangle = 0,$$

что в матричном виде записывается

$$(U_n + \lambda C)\vec{a} = U\widehat{a}_{MD}.$$

Таким образом,

$$\widehat{a}_{Ridge} = (U+\lambda C)^{-1}U\widehat{a}_{MD}.$$

В частном случае $U = X^tX$ получаем

$$(X^tX+\lambda C)^{-1}X^tX(X^tX)^{-1}X^t\vec{Y} = (X^tX+\lambda C)^{-1}X^t\vec{Y}.$$

Зачастую рассматривают $C = E$. Перейдем в базис, в котором U диагонализируется в матрицу D. Тогда

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

Поэтому мы видим что при $\lambda > 0$ коэффициенты $\widehat{a}_{MD}$ уменьшаются, причем маленькие коэффициенты увеличиваются больше. Это позволяет практически обнулить маленькие коэффициенты.

Другим популярным вариантом ограничений является лассо-регрессия, в которой рассматривается $||\vec{a}||_{L_1} \le b$. В этом случае оценка будет минимизировать

$$||\widehat{a}_{MD} - \vec{a}||_U + \lambda(||\vec{a}||_{L_1} - b).$$

Опять же при фиксированном λ мы занимаем минимизацией

$$(\widehat{a}_{MD}-\vec{a})^tU(\widehat{a}_{MD}-\vec{a}) + \lambda\sum_{i=1}^{n}|a_i|.$$

Всевозможные a_i можно разбить на случаи, соответствующий конкретным знакам a_i, T — матрица, на диагонали которой стоят числа ± 1, соответствующие этим знакам. Тогда наша задача переписывается в виде

$$(\widehat{a}_{MD}-\vec{a})^tU(\widehat{a}_{MD}-\vec{a}) + \lambda\vec{e}^tT\vec{a},$$

где $T = (t_{i,i})$, $t_{i,i} = \operatorname{sgn} a_i$. При ненулевых коэффициентах мы получаем дифференцированием формулу

$$-2U\widehat{a}_{MD} + 2U\vec{a} + \lambda T\vec{e} = 0,\ \vec{a}_{Lasso} = \widehat{a}_{MD} - \frac{1}{2}\lambda U^{-1}T\vec{e}.$$

Матрицы U и T могут приведены ортогональным преобразованием к каноническому виду. При этом T останется собой, а U станет диагональной матрицей D. Опять же при этом мы увидим, что

$$(a_{Lasso})_i = (\widehat{a}_{MD})_i - \frac{\lambda t_i}{2d_{i,i}}.$$

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

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

Таким образом, мы уменьшаем коэффициент $\widehat{a}_{MD}$ на константу, если же он при этом становится отри-

цательным, то обнуляем его. За счет этого ридж-регрессия позволяет фильтровать часть предикторов с маленькими коэффициентами.

Полезно взглянуть на это геометрически. Мы решаем задачу минимизации расстояния между фиксированной $\widehat{a}$ и всеми $\vec{a}$, лежащими внутри некоторого единичного круга по норме L^2 или L_1 с точки зрения расстояния $||\widehat{a} - \vec{a}||_U$. В координатах, в которых U диагонализируется, линия уровня

$$||\widehat{a} - \vec{a}||_U^2 = const$$

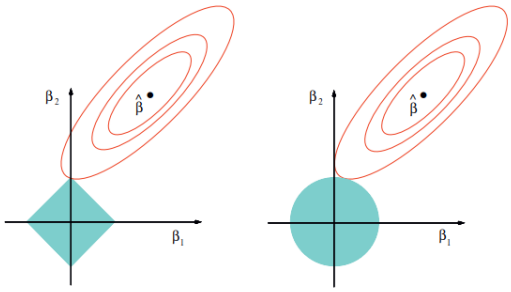
задает эллипс

$$\sum_{i=1}^{k} d_i(\widehat{a} - \vec{a})^2 = const.$$

Наша минимизация есть поиск эллипса наименьшего размера, касающегося круга

$$||\vec{a}||_{L_p} \leq b.$$

При $p = 1$ зачастую таким эллипсом будет эллипс, проходящий через вершину квадрата. При $p = 2$

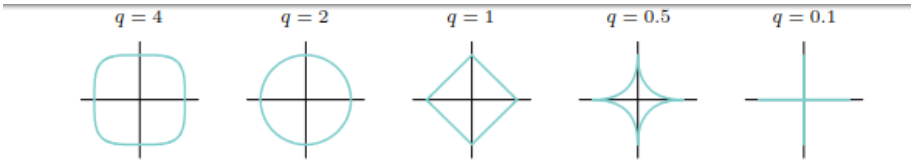


эллипс будет касаться окружности.

Более общая форма предлагает минимизировать

$$||\widehat{a}_{MD} - \vec{a}||_U + \lambda||\vec{a}||_{L_q}$$

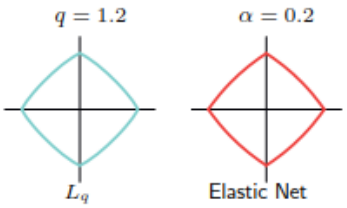
при $q > 0$. Однако, при $q > 1$ мы теряем свойство лассо-регрессии обнулять часть коэффициентов.



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

$$||\widehat{a} - \vec{a}||_{U_n} + \lambda_1(\lambda_2||\vec{a}||_{L_1} + (1 - \lambda_2)||\vec{a}||_{L_2}).$$

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


$q = 1.2$
$\alpha = 0.2$
L_q
Elastic Net