

Todo list

Проверить! Нет ли у β_1 особого положения?	3
на картинке три c : очень большое — дающее мнк решение, меньше — ненулевые β , маленькое — одна из β равна 0	3
Может ли появиться мультимодальность? В точности ли на моду или только примерно? .	3

1 Конвенция

y — вектор столбец зависимых переменных, наблюдаемый случайный

β — вектор столбец неизвестных параметров, ненаблюдаемый, неслучайный

\hat{y} — прогноз y полученный по некоторой модели, наблюдаемый, случайный

$\hat{\beta}$ — оценки β

X — матрица всех объясняющих переменных

ε

$\hat{\varepsilon}$

Некоторые авторы используют обозначения:

Y и y для разных вещей, $y = Y - \bar{Y}$.

2 Семинар 1

Неформальное определение. Если матрица A квадратная, то её определителем называется площадь/объём параллелограмма/параллелепипеда образованного векторами-столбцами матрицы. Знак определителя задаётся порядком следования векторов.

Свойства определителя:

1. $\det(AB) = \det(A) \det(B) = \det(BA)$, если A и B квадратные
2. $\det(A) = \prod \lambda_i$

Определение. Если матрица A квадратная, то её следом называется сумма диагональных элементов, $\text{tr}(A) = \sum a_{ii}$.

Свойства следа:

1. $\text{tr}(AB) = \text{tr}(BA)$, если AB и BA существуют. При этом A и B могут не быть квадратными матрицами.
2. $\text{tr}(A) = \sum \lambda_i$

Добавить про геометрический смысл следа, <http://mathoverflow.net/questions/13526/geometric-interpretation-of-trace>.

Определение. Вектор x называется собственным вектором матрицы A , если при умножении на матрицу A он остается на той же прямой, т.е. $Ax = \lambda x$

Определение. Число λ называется собственным числом матрицы A , если есть вектор x , который при умножении на матрицу A изменяется в λ раз, т.е. $Ax = \lambda x$.

Метод наименьших квадратов (МНК), ordinary least squares (OLS):

Есть n наблюдений, y_1, \dots, y_n . Есть модель, которая даёт прогнозы, $\hat{y}_1, \dots, \hat{y}_n$. Эта модель зависит от вектора неизвестных параметров, β . МНК предлагает в качестве оценок неизвестных параметров взять такое $\hat{\beta}$, чтобы минимизировать $\sum (y_i - \hat{y}_i)^2$.

3 Семинар 2

Контрольная-1

4 Картинка

Утверждение. $\text{sCorr}^2(y, \hat{y}) = R^2$

Доказательство. По определению, $\text{sCorr}(y, \hat{y}) = \frac{(y - \bar{y})(\hat{y} - \bar{\hat{y}})}{|y - \bar{y}||\hat{y} - \bar{\hat{y}}|}$. Поскольку в регрессии присутствует свободный член, $\bar{\hat{y}} = \bar{y}$. Значит,

$$\text{sCorr}(y, \hat{y}) = \frac{(y - \bar{y})(\hat{y} - \bar{y})}{|y - \bar{y}||\hat{y} - \bar{y}|} = \cos(y - \bar{y}, \hat{y} - \bar{y}) \quad (1)$$

По определению, $R^2 = \frac{|\hat{y} - \bar{y}|^2}{|y - \bar{y}|^2} = \cos^2(y - \bar{y}, \hat{y} - \bar{y})$

Опыт: лучший результат у меня получается с обозначением $(\bar{y}, \dots, \bar{y})$.

5 Мегаматрица

След и математическое ожидание можно переставлять, $\mathbb{E}(\text{tr}(A)) = \text{tr}(\mathbb{E}(A))$.

Математическое ожидание квадратичной формы

$$\mathbb{E}(x'Ax) = \text{tr}(A \text{Var}(x)) + \mathbb{E}(x')A\mathbb{E}(x) \quad (2)$$

Доказательство. Мы будем пользоваться простым приёмом. Если u — это скаляр, вектор размера 1 на 1, то $\text{tr}(u) = u$.

Поехали,

$$\mathbb{E}(x'Ax) = \mathbb{E}(\text{tr}(x'Ax)) = \mathbb{E}(\text{tr}(Axx')) = \text{tr}(\mathbb{E}(Axx')) = \text{tr}(A\mathbb{E}(xx')) \quad (3)$$

По определению дисперсии, $\text{Var}(x) = \mathbb{E}(xx') - \mathbb{E}(x)\mathbb{E}(x')$. Поэтому:

$$\text{tr}(A\mathbb{E}(xx')) = \text{tr}(A(\text{Var}(x) + \mathbb{E}(x)\mathbb{E}(x'))) = \text{tr}(A \text{Var}(x)) + \text{tr}(A\mathbb{E}(x)\mathbb{E}(x')) \quad (4)$$

□

6 Разное

1. Гипотеза H_0 по-английски читается как «H naught»
2. При проверке гипотезы об адекватности регрессии НЕЛЬЗЯ писать $H_0 : R^2 = 0$. Гипотеза имеет смысл проверять о ненаблюдаемых неизвестных константах. Проверить гипотезу о том, что $R^2 = 0$ легко. Для этого не нужно знать ничего из теории вероятностей, достаточно просто сравнить посчитанное значение R^2 с нулём. Более того, даже корректировка $\mathbb{E}(R^2) = 0$ неверна. Случайная величина R^2 всегда неотрицательна, поэтому при любых разумных предположениях на ε окажется, что $\mathbb{P}(R^2 > 0) > 0$. А это приведёт к тому, что $\mathbb{E}(R^2) > 0$ даже если Y никак не зависит от X . Единственный правильный вариант — $H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$ и $H_a : \exists i \geq 2 : \beta_i \neq 0$.

7 Ridge/Lasso regression

LASSO — Least Absolute Shrinkage and Selection Operator. Метод построения регрессии, предложенный Robert Tibshirani в 1995 году.

Вспомним обычный МНК:

$$\min_{\beta} (y - X\beta)'(y - X\beta) \quad (5)$$

LASSO вместо исходной задачи решает задачу условного экстремума:

$$\min_{\beta} (y - X\beta)'(y - X\beta) \quad (6)$$

при ограничении $\sum_{j=1}^k |\beta_j| \leq c$.

Проверить! Нет ли у β_1 особого положения?

Естественно, при больших значениях c результат LASSO совпадает с МНК. Что происходит при малых c ?

Для наглядности рассмотрим задачу с двумя коэффициентами β : β_1 и β_2 . Линии уровня целевой функции — эллипсы. Допустимое множество имеет форму ромба с центром в начала координат.

на картинке три c : очень большое — дающее мнк решение, меньше — ненулевые β , маленькое — одна из β равна 0

То есть при малых c LASSO обратит ровно в ноль некоторые коэффициенты β .

Применим метод множителей Лагранжа для случая, когда ограничение $\sum_{j=1}^k |\beta_j| \leq c$ активно, то есть выполнено как равенство.

$$L(\beta, \lambda) = (y - X\beta)'(y - X\beta) + \lambda \left(\sum_{j=1}^k |\beta_j| - c \right) \quad (7)$$

Необходимым условием первого порядка является $\partial L / \partial \beta = 0$. Это условие первого порядка не изменится, если мы зачеркнём c в выражении. Таким образом мы получили альтернативную формулировку метода LASSO:

$$\min_{\beta} (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^k |\beta_j| \quad (8)$$

LASSO пытается минимизировать взвешенную сумму $RSS = (y - X\beta)'(y - X\beta)$ и «размера» коэффициентов $\sum_{j=1}^k |\beta_j|$.

Мы не будем вдаваться в численные алгоритмы, которые используются при решении этой задачи.

Ridge regression отличается от LASSO ограничением $\sum \beta_j^2 \leq c$. Также как и LASSO Ridge regression допускает альтернативную формулировку:

$$\min_{\beta} (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^k \beta_j^2 \quad (9)$$

Также как и LASSO Ridge regression тоже приближает значения коэффициентов β_j к нулю. Принципиальное отличие LASSO и RR. В LASSO крайнее решение с несколькими коэффициентами равными нулю является типичной ситуацией. В RR коэффициент β_j может оказаться точно равным нулю только по чистой случайности.

LASSO допускает байесовскую интерпретацию...

Предположим, что априорное распределение параметров следующее:

...

Тогда мода апостериорного распределения будут приходится в точности (?) на оценки LASSO.

Может ли появиться мультимодальность? В точности ли на моду или только примерно?