Todo list

Проверить! Нет ли у β_1 особого положения? . . . 2
на картинке три c: очень большое — дающиее мнк решение, меньше — ненулевые β, маленькое — одна из β равна 0 . . . 2
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, $\mathrm{tr}(A) = \sum a_{ii}$.
Свойства следа:

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

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, ..., y_n. Есть модель, которая даёт прогнозы, $\hat{y}_1$, ..., $\hat{y}_n$. 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. $\widehat{\mathrm{Corr}}^2(y,\hat{y}) = R^2$

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

$$\widehat{\mathrm{Corr}}(y,\hat{y}) = \frac{(y-\bar{y})(\hat{y}-\bar{y})}{|y-\bar{y}||\hat{y}-\bar{y}|} = \cos(y-\bar{y}, \hat{y}-\bar{y}) \tag{1}$$

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 $(\bar{y},\ldots,\bar{y})$.

5 Разное

1. Гипотеза H_0 по-английски читается как «H naught»
2. При проверке гипотезы об адекватности регрессии НЕЛЬЗЯ писать $H_0 : R^2 = 0$.
   Гипотезы 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 предпосылках 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 — $H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$ и $H_a : \exists i \geqslant 2 : \beta_i \neq 0$.

6 Ridge/Lasso regression

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

Вспомним обычный МНК:

$$\min_{\beta}(y - X\beta)'(y - X\beta) \tag{2}$$

LASSO вместо исходной задачи решает задачу условного экстремума:

$$\min_{\beta}(y - X\beta)'(y - X\beta) \tag{3}$$

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

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 — дающиее 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 $\sum_{j=1}^{k}|\beta_j| \leqslant c$ активно, то есть выполнено как равенство.

$$L(\beta, \lambda) = (y - X\beta)'(y - X\beta) + \lambda \left(\sum_{j=1}^{k} |\beta_j| - c \right) \tag{4}$$

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:

$$\min_{\beta} (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{k} |\beta_j| \tag{5}$$

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 $\sum \beta_j^2 \leqslant c$. Также как и LASSO Ridge regression допускает альтернативную формулировку:

$$\min_{\beta} (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{k} \beta_j^2 \tag{6}$$

Также как и LASSO Ridge regression тоже приближает значения коэффициентов β_j к нулю. Принципиальное отличие LASSO и RR. В LASSO краевое 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?