Эконометрика. Лекция 10. Три сюжета напоследок

Три сюжета

- Квантильная регрессия
- Алгоритм случайного леса
- Байесовский подход

Квантильная регрессии

Моделировать можно не только среднее, но и медиану или другой определённый квантиль.

Классическая регрессия — модель для среднего

Предпосылки классической модели:

- $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$
- экзогенность, $E(\varepsilon_i | x_i) = 0$
- другие предпосылки

Следствие:

$$E(y_i | x_i) = \beta_1 + \beta_2 x_i$$

Минимизация суммы квадратов

Модель: $E(y_i|x_i) = \beta_1 + \beta_2 x_i$

- Сумма квадратов остатков, $Q(\hat{\beta}_1, \hat{\beta}_2) = \sum_i (y_i - \hat{y}_i)^2$
- При минизации $Q(\hat{\beta}_1, \hat{\beta}_2)$ получаем состоятельные оценки $\hat{\beta}_1$, $\hat{\beta}_2$

Медианная регрессия

Модель: $Med(y_i|x_i) = \beta_1 + \beta_2 x_i$

Алгоритм получения оценок

- Сумма модулей остатков, $M(\hat{\beta}_1, \hat{\beta}_2) = \sum_i |y_i - \hat{y}_i|$
- Минимизируя $M(\hat{\beta}_1, \hat{\beta}_2)$ получаем состоятельные оценки $\hat{\beta}_1$, $\hat{\beta}_2$

Пример у неоновой доски

Найдите оценку $\hat{\beta}$ медианной регрессии:

$$Med(y_i|x_i) = \beta x_i$$

Набор данных:

y	x
1	1
2	5
6	5

Медианная и классическая регрессия

- Классическая: от каких факторов зависит $E(y_i|x_i)$?
- Медианная: от каких факторов зависит $Med(y_i|x_i)$?
- Если распределение ε_i симметрично, то оба подхода дают асимптотически одинаковые оценки

Медианная регрессия: минусы

- Нет явных формул для оценок коэффициентов и стандартных ошибок
- Только асимптотические свойства оценок коэффициентов

Медианная регрессия: плюсы

- Взгляд на данные с другой стороны
- Более устойчивые оценки в случае “выбросов” в ε_i

Произвольная квантиль

- Медиана, $Med(y_i)$, — квантиль 50%

$$P(y_i \leq Med(y_i)) = 0.5$$

- Квантиль порядка τ, q_τ:

$$P(y_i \leq q_\tau) = \tau$$

(здесь картинка)

Квантильная регрессия

Модель: $q_\tau(y_i|x_i) = \beta_1^\tau + \beta_2^\tau x_i$

- Зависимость для разных квантилей может быть разная!

Асимметричная сумма модулей остатков:

$M(\hat{\beta}_1, \hat{\beta}_2) = \sum_i w_i \cdot |y_i - \hat{y}_i|$

где веса w_i равны:

$$w_i = \begin{cases} (1-\tau), \ y_i < \hat{y}_i \\ \tau, \ y_i \geq \hat{y}_i \end{cases}$$

- Минимизируя $M(\hat{\beta}_1, \hat{\beta}_2)$ получаем состоятельные оценки $\hat{\beta}_1$, $\hat{\beta}_2$

Квантильная регрессия стоимости квартир

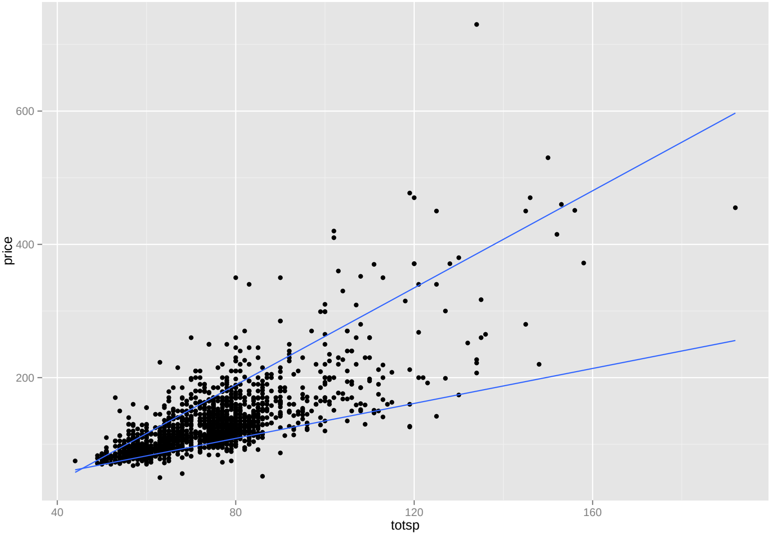
недорогое жильё (10%-ый квантиль):

$\widehat{price}_i = 3.9 + 1.3 totsp_i$

дорогое жильё (90%-ый квантиль):

$\widehat{price}_i = -102.4 + 3.6 totsp_i$

Квантильная регрессия стоимости на графике



Алгоритм случайного леса

- Очень хорошо прогнозирует
- Не объясняет, как устроены данные

Две версии алгоритма

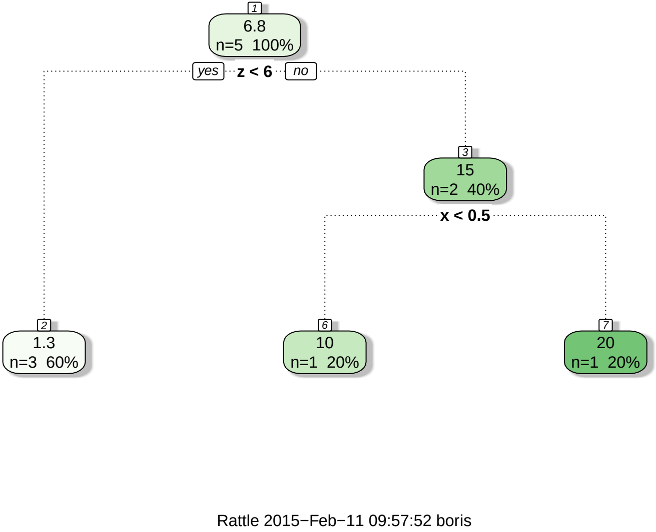
- Для непрерывной y_i
- Для качественной y_i

Каждый мужчина должен посадить дерево

Набор данных

y	x	z
1	1	-2
1	0	3
2	0	-4
10	0	9
20	1	9

Каждый мужчина должен посадить дерево



(рассказываю, что прогнозировать легко)

Как посадить дерево?

- Из имеющихся k переменных случайно отбираем $k' = \lceil k/3 \rceil$ переменных
- Из отобранных k' переменных выбираем ту, которая даёт наилучшее деление ветви дерева на две
- Повторяем до тех пор, пока в каждом терминальном узле остаётся больше $nodesize = 5$ наблюдений

Наилучшее деление

До деления: $RSS = 274.8$

$\{1, 1, 2, 10, 20\}$, $\hat{y} = \bar{y} = 6.8$,

После разбиения: $RSS = RSS_1 + RSS_2 = 50.67$

Слева: $\{1, 1, 2\}$, $\hat{y} = \bar{y} = 6.8$, $RSS_1 = 0.67$

Справа: $\{10, 20\}$, $\hat{y} = \bar{y} = 15$, $RSS_2 = 50$

Алгоритм случайный

Повторное применение алгоритма к тому же набору данных даст слегка другие оценки

Неоновая доска

Пример построения классификационного дерева

y	x
1	1
2	2
9	3
10	4
10	5

Мужчина, владеющий R, может посадить целый лес!

- Случайным образом отбираем (с повторениями) n наблюдений из исходных n наблюдений
- Сажаем дерево по случайной подвыборке
- Повторяем до получения $n_{tree} = 500$ деревьев

Прогноз случайного леса:

- Каждое из $n_{tree} = 500$ деревьев даёт свой прогноз $\hat{y}_i$
- Усредняем и получаем финальный прогноз

Байейсовский подход

Опишем наше незнание параметра θ в виде априорного закона распределения!

Пример. Неизвестная вероятность

- $p \in [0; 1]$

Априорная плотность:

$$f(p) = \begin{cases} 1, & p \in [0; 1] \\ 0, & \text{иначе} \end{cases}$$

(здесь картинка)

Пример. Неизвестный положительный коэффициент

- $\beta \in [0; +\infty)$

Априорная плотность:

$$f(\beta) = \begin{cases} exp(-\beta), \ \beta \in [0; \infty) \\ 0, \quad \text{иначе} \end{cases}$$

(здесь картинка)

Модель

Модель задаёт закон распределения наблюдений, y_i, при фиксированном значении параметров

Например,

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$$

Кристально-чистая логика байесовского подхода

Определяем:

- Априорное распределение, $f(\theta)$
- Модель для данных, $f(y|\theta)$

По формуле условной вероятности получаем:

- Апостериорное распределение, $f(\theta|y)$

Формула условной вероятности

$$f(\theta|y) = \frac{f(y|\theta) \cdot f(\theta)}{f(y)} \sim f(y|\theta) \cdot f(\theta)$$

Пример у неоновой доски

Караси, щуки

- нет информации
- Бабушка: караси встречаются чаще щук!

Как описать сложную функцию плотности?

(картинка)

- Большая выборка независимых значений случайной величины r:

$r_1, r_2, r_3, \ldots, r_{10000}$

- Можно оценить всё: $E(r)$, $E(r^2)$, $P(r > 0)$

Монте-Карло по схеме Марковской цепи

MCMC (Markov Chain Monte Carlo)

Заменяет формулу условной вероятности

Алгоритм MCMC

На входе:

- Априорное распределение, $f(\theta)$
- Модель для данных, $f(y|\theta)$

На выходе:

- Большая выборка из апостериорного распределения, $f(\theta|y)$

(картинка)

Алгоритм случайный

Повторное применение алгоритма к тому же набору данных даст слегка другие оценки

Плюсы байесовского подхода

- Можно задавать вопросы про неизвестные параметры:

$P(\beta_3 > 0|y)$, $P(\beta_3 = 0|y)$, $E(\beta_3|y)$?

- Апостериорное распределение есть всегда!

даже при жесткой мультиколлинеарности и полном отсутствии наблюдений

Минусы байесовского подхода

- Его не все знают
- Может требовать больших объемов вычислений

MCMC и логит

“Идеальное прогнозирование” — ситуация, в которой ML оценки логит-модели не существуют

Логит-модель

$y_i \in \{0, 1\}$.

$$y_i = \begin{cases} 1, y_i^* \geq 0 \\ 0, y_i^* < 0 \end{cases}$$

Скрытая переменная: $y_i^* = \beta_1 + \beta_2 x_i + \varepsilon_i$.

Априорное распределение для логит модели

$\beta \sim N(b_0, B_0^{-1})$

Гиперпараметры:

b_0 — априорное среднее

B_0 — априорная матрица точности

$B_0^{-1} = Var(\beta)$

Выбор априорных гиперпараметров

Традиционно:

$b_0 = (0, 0, \dots, 0)'$

$$B_0 = \begin{pmatrix} d & 0 & 0 & \dots \\ 0 & d & 0 & \dots \\ 0 & 0 & d & \dots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

d — мало

Пример проблемной ситуации

y	x
0	1
0	2
1	3

Логит и пробит оценки не существуют

Логит со вкусом Байеса

Априорно: $\beta_1 \sim N(0, 10^2)$, $\beta_2 \sim N(0, 10^2)$

Апостериорные средние:

$\hat{y}_i^* = -10.8 + 4.5x_i$

$$y_i = \begin{cases} 1, y_i^* \geq 0 \\ 0, y_i^* < 0 \end{cases}$$

Регрессия пик-плато

Модель: $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$

Вариант априорного распределения пик-плато

- $\beta_j | \gamma_j, \tau_j^2 \sim N(0, \gamma_j \cdot \tau_j^2)$
- $\gamma_j = \begin{cases} 1, & \text{с вероятностью } 1/2 \\ 0, & \text{с вероятностью } 1/2 \end{cases}$
- $\tau_j^2 \sim \Gamma^{-1}(a_1, a_2)$
- $\sigma^2 \sim \Gamma^{-1}(b_1, b_2)$

Гиперпараметры: a_1, a_2, b_1, b_2

Регрессия пик-плато

Позволяет напрямую отвечать на вопрос:

Чему равна вероятность $P(\beta_2 = 0|y)$?

Пример с машинами

(картинка из первой лекции)

$\widehat{dist}_i = 12.81 + 0.28 speed_i + 0.01 speed_i^2$

Апостериорные вероятности:

$P(\beta_{speed} = 0|y) = 0.15$

$P(\beta_{speed^2} = 0|y) = 0.05$

Большое спасибо

Нам не удалось решить все наши задачи. Решения, что мы находим, лишь ставят перед нами новые вопросы. В каком-то смысле, мы также мало знаем, как и раньше. Но мы верим, что наше незнание стало глубже, а не знаем мы всё более важные вещи.

Большое спасибо тем, кто прошел вместе с нами этот курс до конца!