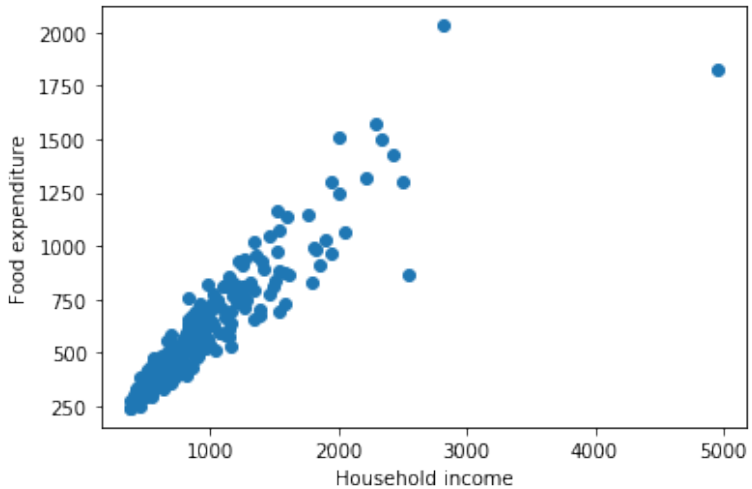
Thuật toán EM cho hồi quy phân vị

Nguyễn Thị Ngọc Anh, Nguyễn Mạnh Tuấn
Viện Toán ứng dụng và Tin học
Đại học bách khoa Hà Nội

Ngày 26 tháng 2 năm 2022

Đặt vấn đề

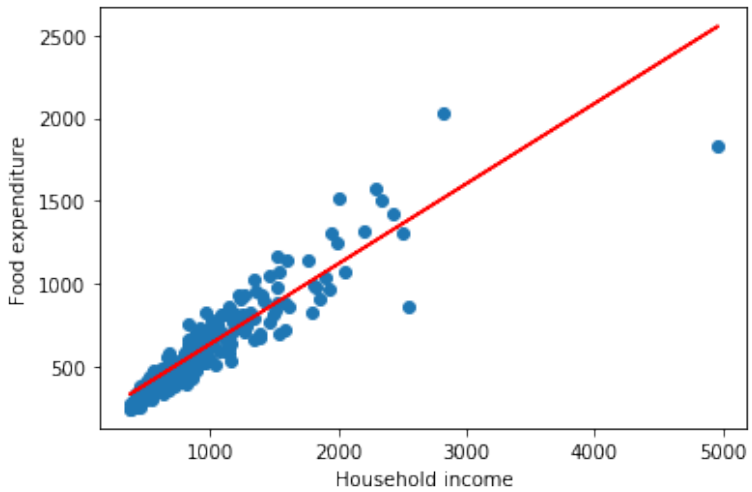
Xét bộ dữ liệu gồm quan sát về mối tương quan giữa số tiền mua thức ăn và thu nhập của 235 hộ gia đình thuộc tầng lớp lao động ở Bỉ được thu thập bởi Ernst Engel (1857) như sau:



Hình: Biểu đồ mô tả của dữ liệu Engel

Đặt vấn đề

Ta thấy dữ liệu này có phương sai thay đổi, càng lớn khi x (Household income) càng lớn. Ta thử chạy hồi quy tuyến tính:



Hình: Đường hồi quy tuyến tính

Phát biểu bài toán

Xét biến ngẫu nhiên 1 chiều $y \in \mathbb{R}$ và vector ngẫu nhiên x k chiều $x = (x_1, \ldots, x_k)^T \in \mathbb{R}^k$ và $x_1 = 1$.
Mối quan hệ của y và x được thể hiện bằng mô hình hồi quy phân vị tại mức $p \in (0, 1)$ như sau:

$$y\big|_p = F^{-1}_{y|x}(p) = x^T.\beta^*(p),$$

Với $\beta^*(p) = \left(\beta^*_1(p), \ldots, \beta^*_k(p)\right)^T \in \mathbb{R}^k$ là hệ số hồi quy phân vị thực.

Bài toán tối ưu

Xét $\{(y_i, x_i)\}_{i=1}^n$ là các quan sát được của (y, x). Ước lượng tham số cho β được cho bởi:

$$\hat{\beta}(p) \in \min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta) = \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_p(y_i - x_i^T \beta)$$

Với $\rho_p(u) = u\{p - 1(u < 0)\}$ là hàm lỗi ở mức phân vị p

Bài toán thống kê

Ta xét bài toán thống kê tương đương:
Xét mô hình hồi quy:

$$y = x^T.\beta(p) + \epsilon \tag{1}$$

Với $\beta(p)$ là tham số mô hình, ϵ là lỗi ngẫu nhiên và có dạng phân phối Laplace không đối xứng: $ALD(0, \sigma, p)$

Phân phối Laplace không đối xứng

Nếu $V \sim ALD(\mu, \sigma, p)$ thì hàm xác suất của V có dạng:

$$f(v; \mu, \sigma, p) = \frac{p(1-p)}{\sigma} exp\left(- \rho_p\left(\frac{v-\mu}{\sigma}\right)\right) \qquad (2)$$

Ví dụ

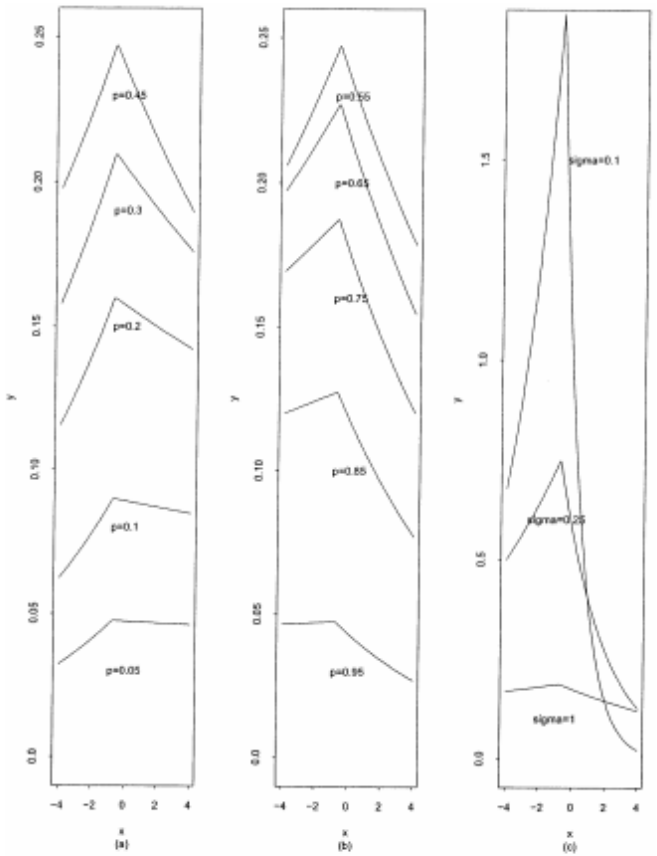


Figure 1. ALD densities (a) for $0 < p < 0.5$ with $\sigma = 1$; (b) for $0.5 < p < 1$ with $\sigma = 1$; (c) for $0.1 \leq \sigma \leq 1$ with $p = 0.75$.

Ước lượng tham số thống kê

- Do $\epsilon \sim ALD(0, \sigma, p)$ nên $y \sim ALD(\langle x, \beta(p) \rangle, \sigma, p)$.
- Khi biết các quan sát $\{y_i, x_i\}_{i=1,,n}$, ta có thể ước lượng tham số $\beta(p)$ bằng cách phương pháp cực đại hóa hàm hợp lý (MLE).

Ta định nghĩa hàm hợp lý như sau:

$$L(y; \beta(p), \sigma, p) = f(y_1, y_2, \ldots, y_n; \beta(p), \sigma, p) \quad (3)$$

$$= \prod_{i=1}^{n} f(y_i; \beta(p), \sigma, p) \quad (4)$$

Hàm log-likelihood

Từ đó, ta có hàm log-likelihood:

$$logL(y_i; \beta(p), \sigma, p) = \sum_{i=1}^{n} logf(y_i; \beta(p), \sigma, p) \quad (5)$$

$$= \sum_{i=1}^{n} \left[log(p) + log(1-p) - log(\sigma) - \frac{1}{\sigma} \rho_p \left(y_i - x_i^T . \beta(p) \right) \right] \quad (6)$$

Ước lượng hợp lý cực đại

Khi đó,

$$logL(y_i; \beta(p), \sigma, p)$$

$$= \sum_{i=1}^{n} \left[log(p) + log(1-p) - log(\sigma) - \frac{1}{\sigma} \rho_p \left(y_i - x_i^T \beta(p) \right) \right] \quad (7)$$

$$\Rightarrow L \to \max \iff \sum_{i=1}^{n} \rho_p \left(y_i - x_i^T . \beta(p) \right) \to \min \quad (8)$$

Bài toán thống kê

Khi biết các quan sát được $\{y_i, x_i\}_{i=1,\ldots,n}$ của $\{y, x\}$ và quan hệ giữa y và x:

$$y = x^T.\beta(p) + \epsilon \tag{9}$$

Với ϵ là lỗi ngẫu nhiên và có dạng phân phối Laplace không đối xứng: $ALD(0, \sigma, p)$, ta cần ước lượng tham số $\beta(p)$ theo phương pháp MLE.

Biểu diễn của lại phân phối của lỗi ϵ

Ta có thể biểu diễn $\epsilon \sim ALD(0, \mu, \tau)$ dưới dạng:

$$\epsilon = \theta z + \tau\sqrt{z}u \tag{10}$$

Với $\theta = \frac{1-2p}{p(1-p)}$, $\tau^2 = \frac{2}{p(1-p)}$, $z \sim exp(1)$, $u \sim N(0,1)$ và z độc lập với u.

Biểu diễn của lại phân phối của lỗi ϵ

Khi đó, ta có thể biểu diễn phương trình hồi quy dưới dạng:

$$y = x^T.\beta(p) + \theta z + \tau\sqrt{z}u \sim N(x^T.\beta(p) + \theta z, \tau^2 z) \quad (11)$$

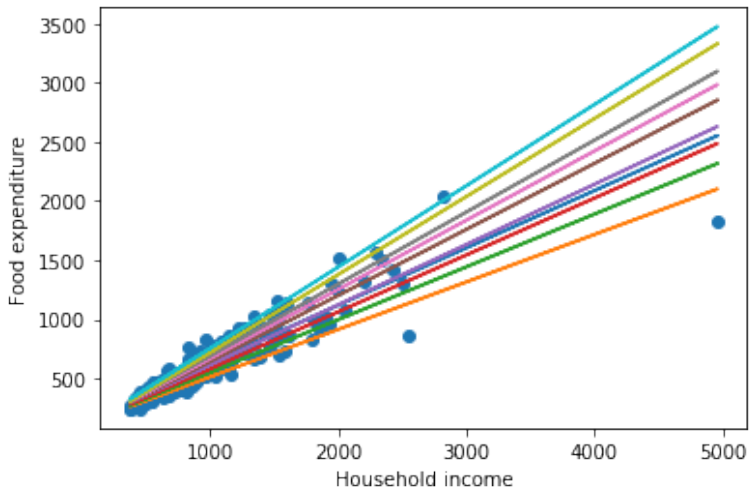
Ước lượng bằng EM

Ở phương trình trên, ta xem z là dữ liệu không quan sát được.
Tian et al (2014) đưa ra công thức cập nhật của thuật toán EM cho bài toán như sau:

1. Khởi tạo
$$\hat{\beta}^{(0)}(p) = \hat{\beta}_{OLS}$$
2. Cập nhật
$$\hat{\beta}^{(t+1)}(p) = (XW^{(t)}X^T)^{-1}(XW^{(t)}Y - \theta X.1_n)$$

Kết quả chạy

Kết quả chạy tại các mức phân vị $0.1, 0.2, \ldots, 0.9$ trên bộ dữ liệu Engel với 50 vòng lặp:



Hình: Kết quả chạy thuật toán EM trên bộ dữ liệu Engel