Maximum likelihood estimation for regression models

Ciaran Evans

Maximum likelihood estimation and linear regression

Let $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$ be iid samples from the model

$$Y_i | \mathbf{x}_i \sim N(\mu_i, \underrightarrow{\sigma^2})$$

$$\mu_i = \mathbf{x}_i^T \beta \quad = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

where the distribution of $\mathbf{x}_i$ does not depend on β or σ^2.

$$L(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} f(x_i, Y_i \mid \beta, \sigma^2) = \prod_{i=1}^{n} f(x_i) f(Y_i | x_i, \beta, \sigma^2)$$

does not involve β or σ^2

$$= \left(\prod_{i=1}^{n} f(x_i) \right) \left(\prod_{i=1}^{n} f(Y_i | x_i, \beta, \sigma^2) \right)$$

does involve β and σ^2

$$\propto \prod_{i=1}^{n} f(Y_i | x_i, \beta, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (Y_i - x_i^T\beta)^2 \right\}$$

SSE

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - x_i^T\beta)^2 \right\}$$

$\Rightarrow$ choosing β to maximize L is equivalent to minimizing SSE!

Maximum likelihood estimation and logistic regression

Let $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$ be iid samples from the model

$$Y_i|\mathbf{x}_i \sim Bernoulli(p_i) \qquad f(Y_i|x_i,\beta) = p_i^{Y_i}(1-p_i)^{1-Y_i}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T\beta \qquad p_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

where the distribution of $\mathbf{x}_i$ does not depend on β.

$$L(\beta|\mathbf{y},\mathbf{X}) \propto \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i,\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

(up to a constant) $= \prod_{i=1}^{n} \left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)^{Y_i} \left(\frac{1}{1+e^{x_i^T\beta}}\right)^{1-Y_i}$

$$\ell(\beta|y,X) = \sum_{i=1}^{n} \left\{ Y_i \log\left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right) + (1-Y_i)\log\left(\frac{1}{1+e^{x_i^T\beta}}\right)\right\}$$

$$= \sum_{i=1}^{n} \left\{ Y_i x_i^T\beta - \log\left(1+e^{x_i^T\beta}\right)\right\}$$

Maximizing

$$\begin{bmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_k} \end{bmatrix}$$

$$\ell(\beta|\mathbf{y},\mathbf{X}) \overset{\text{(up to a constant)}}{=} \sum_{i=1}^{n} \left\{ Y_i \mathbf{x}_i^T \beta - \log(1 + e^{\mathbf{x}_i^T \beta}) \right\}$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \left\{ \underbrace{\frac{\partial}{\partial \beta} Y_i x_i^T \beta}_{Y_i x_i} - \underbrace{\frac{\partial}{\partial \beta} \log\left(1 + e^{x_i^T \beta}\right)}_{-\frac{1}{1+e^{x_i^T\beta}} \cdot e^{x_i^T \beta} \cdot x_i} \right\}$$

$$\left(\frac{\partial}{\partial u} a^T u = a\right)$$

(chain rule)

$$= \sum_{i=1}^{n} \left\{ Y_i x_i - \boxed{\frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}} x_i \right\}$$

p_i

$$= \sum_{i=1}^{n} (Y_i - p_i) x_i = \underbrace{X^T}_{X \,=\, \text{design matrix}} \left(\underbrace{y}_{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}} - \underbrace{p}_{\begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}}\right)$$

Score

Definition (score): Let $\mathbf{y} = (Y_1, ..., Y_n)$ be a sample of n observations from some distribution with parameter vector $\boldsymbol{\theta}$. Let $L(\boldsymbol{\theta}|\mathbf{y})$ be the likelihood function, and $\ell(\boldsymbol{\theta}|\mathbf{y}) = \log L(\boldsymbol{\theta}|\mathbf{y})$ the log-likelihood.

The **score**, which we will denote $U(\boldsymbol{\theta})$, is the gradient of the log-likelihood with respect to $\boldsymbol{\theta}$:

$$U(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\mathbf{y}).$$

Example: For logistic regression: $U(\beta) = \mathbf{X}^T(\mathbf{y} - \mathbf{p})$

Question: How would I solve $\mathbf{X}^T(\mathbf{y} - \mathbf{p}) = \mathbf{0}$?

Challenge: p is a nonlinear function of β

In fact, there is no closed form solution for β!

Newton's method

We want to find β^* such that $U(\beta^*) = \mathbf{0}$. Issue: no closed form solution!

Idea: Approximate $U(\beta^*)$ with a first-order Taylor expansion:

Taylor expansion of $g(x)$ around a;

$$g(x) \approx g(a) + g'(a)(x-a)$$

$$U(\beta^*) \approx u(\beta) + \left(\frac{\partial u(\beta)}{\partial \beta}\right)(\beta^* - \beta)$$

(if β is close enough to β^*)

Let $\beta^{(0)}$ be an initial guess for β^*

Want to iteratively update and improve this initial guess

$$\underset{\underset{0}{\shortparallel}}{u(\beta^*)} \approx u(\beta^{(0)}) + \left(\frac{\partial u}{\partial \beta}\Big|_{\beta=\beta^{(0)}}\right)(\beta^* - \beta^{(0)})$$

$$\Rightarrow \quad \beta^* \approx \underbrace{\beta^{(0)} - \left(\frac{\partial u}{\partial \beta}\Big|_{\beta=\beta^{(0)}}\right)^{-1} u(\beta^{(0)})}_{\text{we can evaluate this!}}$$

$$u(\beta) = \frac{\partial}{\partial \beta} \ell(\beta | y, x) \qquad H(\beta) = \frac{\partial u}{\partial \beta} = \frac{\partial^2}{\partial \beta^2} \ell(\beta | y, x)$$

second derivative! (Hessian matrix)

Newton's method

- Want β^* such that $U(\beta^*) = \mathbf{0}$
- Begin with initial estimate ~~$\beta^{(k)}$~~ $\beta^{(0)}$
- Iterative updates:

$$\beta^{(r+1)} = \beta^{(r)} - \left(\mathbf{H}(\beta^{(r)})\right)^{-1} U(\beta^{(r)})$$

$$U(\beta) = \frac{\partial \ell}{\partial \beta} \in \mathbb{R}^{k+1} \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \in \mathbb{R}^{k+1}$$

(gradient)

$$H(\beta) = \frac{\partial^2 \ell}{\partial \beta^2} \in \mathbb{R}^{(k+1)\times(k+1)}$$

(Hessian)

$$= \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_k} \\ \vdots & & & \ddots \\ \frac{\partial^2 \ell}{\partial \beta_k \partial \beta_0} & & \cdots & \frac{\partial^2 \ell}{\partial \beta_k^2} \end{bmatrix}$$

The Hessian

$$U(\beta) = \frac{\partial}{\partial \beta} \ell(\beta | \mathbf{y}, \mathbf{X}) = \mathbf{X}^T(\mathbf{y} - \mathbf{p})$$

$$\mathbf{H}(\beta) = \frac{\partial}{\partial \beta} U(\beta) = \frac{\partial}{\partial \beta} \mathbf{X}^T(\mathbf{y} - \mathbf{p})$$

$$= -\frac{\partial}{\partial \beta} X^T p$$

(chain rule)

$$= \left(-\frac{\partial p}{\partial \beta} \right) X$$

$$p = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} \in \mathbb{R}^n$$

$$\beta \in \mathbb{R}^{k+1}$$

$$\frac{\partial p}{\partial \beta} = \left[\frac{\partial p_1}{\partial \beta} \quad \frac{\partial p_2}{\partial \beta} \quad \cdots \quad \frac{\partial p_n}{\partial \beta} \right] \in \mathbb{R}^{(k+1) \times n}$$

So : need to find $\frac{\partial p_i}{\partial \beta}$

$$p_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

Putting everything together

Want to maximize the log likelihood $\ell(\beta|\mathbf{y}, \mathbf{X})$.