Maximum Likelihood

May 3, 2016

Maximum Likelihood

- Suppose that we have data:

$$\begin{matrix} y_1 & x_{11} & \cdots & x_{1m} \\ y_2 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{nm} \end{matrix}$$

- What if y is categorical or otherwise limited?

Maximum Likelihood

- Suppose that we have data:

$$\begin{matrix} y_1 & x_{11} & \cdots & x_{1m} \\ y_2 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{nm} \end{matrix}$$

- What if y is categorical or otherwise limited? MAXIMUM LIKELIHOOD!

Maximum Likelihood

- Remember that we had:

$$p(\beta|y) = \frac{\overbrace{p(y|\beta)}^{\text{Likelihood}}\ p(\beta)}{p(y)}$$

Maximum Likelihood

- Remember that we had:

$$p(\beta|y) = \frac{\overbrace{p(y|\beta)}^{\text{Likelihood}} \; p(\beta)}{p(y)}$$

- Probability = function of an outcome given a fixed value for the parameter (before data available).

Maximum Likelihood

- Remember that we had:

$$p(\beta|y) = \frac{\overbrace{p(y|\beta)}^{\text{Likelihood}}\;\; p(\beta)}{p(y)}$$

- Probability = function of an outcome given a fixed value for the parameter (before data available).
- Likelihood = function of a parameter for a fixed outcome (after data available).

Maximum Likelihood

- Remember that we had:

$$p(\beta|y) = \frac{\overbrace{\mathcal{L}(\beta|y)}^{\text{Likelihood}} \; p(\beta)}{p(y)}$$

- Probability = function of an outcome given a fixed value for the parameter (before data available).
- Likelihood = function of a parameter for a fixed outcome (after data available).

Maximum Likelihood

- Let's forget about everything except the likelihood!

$$\mathcal{L}(\beta|y) = p(y|\beta)$$

- How do we use this?

Maximum Likelihood

- Let's forget about everything except the likelihood!

$$\mathcal{L}(\beta|y) = p(y|\beta)$$

- How do we use this? Define a cost function!

Maximum Likelihood

- Let's forget about everything except the likelihood!

$$\mathcal{L}(\beta|y) = p(y|\beta)$$

- How do we use this? Define a cost function! How?!

Maximum Likelihood

- We need some way to represent:

$$\mathcal{L}(\beta | y_1, y_2, \ldots, y_n) = p(y_1, y_2, \ldots, y_n | \beta)$$

- Remember that if A and B are independent then:

$$p(A, B) = p(A)p(B)$$

Maximum Likelihood

- We need some way to represent:

$$\mathcal{L}(\beta|y_1, y_2, \dots, y_n) = p(y_1, y_2, \dots, y_n|\beta)$$

- Remember that if A and B are independent then:

$$p(A, B) = p(A)p(B)$$

- So if we are willing to assume that $y_1, y_2, \dots, y_n$ are independent we have:

$$\mathcal{L}(\beta|y_1, y_2, \dots, y_n) = p(y_1|\beta)p(y_2|\beta), \dots, p(y_n|\beta) = \prod_i p(y_i|\beta)$$

Maximum Likelihood

- We have a cost function:

$$\mathcal{L}(\beta|y) = \prod_i p(y_i|\beta)$$

- So what is maximum likelihood then?

Maximum Likelihood

- We have a cost function:

$$\mathcal{L}(\beta|y) = \prod_i p(y_i|\beta)$$

- So what is maximum likelihood then?

- It is the solution to:

$$\max_\beta \left\{ \prod_i p(y_i|\beta) \right\}$$

Maximum Likelihood

- We have a cost function:

$$\mathcal{L}(\beta|y) = \prod_i p(y_i|\beta)$$

- So what is maximum likelihood then?

- It is the solution to:

$$\ln\left\{\max_\beta \left\{\prod_i p(y_i|\beta)\right\}\right\}$$

Maximum Likelihood

- We have a cost function:

$$\mathcal{L}(\beta|y) = \prod_i p(y_i|\beta)$$

- So what is maximum likelihood then?

- It is the solution to:

$$\max_\beta \left\{ \ln \left\{ \prod_i p(y_i|\beta) \right\} \right\}$$

Maximum Likelihood

- We have a cost function:

$$\mathcal{L}(\beta|y) = \prod_i p(y_i|\beta)$$

- So what is maximum likelihood then?

- It is the solution to:

$$\max_\beta \left\{ \sum_i \ln \{p(y_i|\beta)\} \right\}$$

Maximum Likelihood

- We know that OLS is the Best Linear Unbiased Estimator. Does MLE have similar properties?

Maximum Likelihood

- We know that OLS is the Best Linear Unbiased Estimator. Does MLE have similar properties?
- Yes!
  - MLE is consistent: with infinite data it will estimate the correct β
  - MLE is efficient: no consistent estimator has lower asymptotic mean squared error than MLE

Maximum Likelihood

- There are always three steps in MLE:

  1. Write down the likelihood;

  2. Take the natural log and simplify;

  3. Maximize.

Maximum Likelihood: A simple example

- Suppose that we have n coin flips and we observe k heads. What is the most likely value for the probability of heads, π?

- Suppose $y_1 = H, y_2 = H, y_3 = T, \ldots, y_{n-1} = H, y_n = T$.

- Step (1) write down likelihood of π – the probability of n flips leading to k heads is:

$$\left(\begin{array}{c} n \\ k \end{array}\right) \pi\pi(1-\pi)\ldots\pi(1-\pi) = \left(\begin{array}{c} n \\ k \end{array}\right) \pi^k (1-\pi)^{n-k}$$

Maximum Likelihood: A simple example

- So what next?

Maximum Likelihood: A simple example

- So what next? Step (3) maxmimize:
  - Take derivative

$$
\begin{aligned}
0 &= \frac{d}{d\pi}\left\{\binom{n}{k}\pi^k(1-\pi)^{n-k}\right\} \\
0 &= \frac{d}{d\pi}\left\{\pi^k(1-\pi)^{n-k}\right\} \\
0 &= \frac{d}{d\pi}\left\{\pi^k\right\}(1-\pi)^{n-k} + \pi^k\frac{d}{d\pi}\left\{(1-\pi)^{n-k}\right\} \\
0 &= k\pi^{k-1}(1-\pi)^{n-k} - (n-k)\pi^k(1-\pi)^{n-k-1}
\end{aligned}
$$

  - Solve for π

$$\pi = \frac{k}{n}.$$

Maximum Likelihood: Linear Regression

- Suppose we have our usual linear model:

$$y = x\beta + \varepsilon$$

- Remember that a central assumption was that $\varepsilon_i \sim N(0, \sigma^2)$.

Maximum Likelihood: Linear Regression

- Suppose we have our usual linear model:

$$y = x\beta + \varepsilon$$

- Remember that a central assumption was that $\varepsilon_i \sim N(0, \sigma^2)$. This means that $y_i \sim N(x_i\beta, \sigma^2)$.

Maximum Likelihood: Linear Regression

- Suppose we have our usual linear model:

$$y = x\beta + \varepsilon$$

- Remember that a central assumption was that $\varepsilon_i \sim N(0, \sigma^2)$. This means that $y_i \sim N(x_i\beta, \sigma^2)$.

- Step (1) write down the likelihood of β and σ^2:

$$\mathcal{L}(\beta, \sigma^2 | y, x) = \prod_i N(x_i\beta, \sigma^2 | y_i)$$

Maximum Likelihood: Linear Regression

- Step (2) log and simplify:

$$\mathcal{L}(\beta, \sigma^2 | y, x) = \prod_i N(x_i\beta, \sigma^2 | y_i)$$

$$\mathcal{L}(\beta, \sigma^2 | y, x) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i\beta - y_i)^2}{2\sigma^2} \right\}$$

$$\begin{aligned} \ln \mathcal{L}(\beta, \sigma^2 | y, x) &= \sum_i \ln \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i\beta - y_i)^2}{2\sigma^2} \right\} \right\} \\ &= \sum_i -\frac{\ln\{2\pi\sigma^2\}}{2} - \frac{(x_i\beta - y_i)^2}{2\sigma^2} \end{aligned}$$

Maximum Likelihood: Linear Regression

- Step (3) maximize:

$$\max_{\beta,\sigma}\left\{\sum_i -\frac{\ln\{2\pi\sigma^2\}}{2} - \frac{(x_i\beta - y_i)^2}{2\sigma^2}\right\}$$

- Do this on a computer!

Maximum Likelihood: Logistic Regression

- In linear regression y is continuous and we assume:

$$y_i = x_i\beta + \varepsilon_i$$

- In logistic regression y is binary (coded 1 or 0) and we assume:

$$p(y_i = 1) = \frac{\exp\{x_i\beta\}}{1 + \exp\{x_i\beta\}}$$

Maximum Likelihood: Logistic Regression

- How would we apply MLE here? Step (1) write down the likelihood:

$$
\begin{aligned}
\mathcal{L}(\beta|x,y) &= \prod_i \left[p(y_i=1)\right]^{y_i}\left[p(y_i=0)\right]^{1-y_i} \\
&= \prod_i \left[p(y_i=1)\right]^{y_i}\left[1-p(y_i=1)\right]^{1-y_i} \\
&= \prod_i \left[\frac{\exp\{x_i\beta\}}{1+\exp\{x_i\beta\}}\right]^{y_i}\left[1-\frac{\exp\{x_i\beta\}}{1+\exp\{x_i\beta\}}\right]^{1-y_i}
\end{aligned}
$$

- Step (2) log and simplify:

$$
\begin{aligned}
\ln \mathcal{L}(\beta|x,y) &= \sum_i y_i \ln\left[\frac{\exp\{x_i\beta\}}{1+\exp\{x_i\beta\}}\right] \\
&\qquad + (1-y_i)\ln\left[1-\frac{\exp\{x_i\beta\}}{1+\exp\{x_i\beta\}}\right] \\
&= \sum_i y_i x_i\beta - \ln\left[1+\exp\{x_i\beta\}\right]
\end{aligned}
$$

Maximum Likelihood: Logistic Regression

- Step (3) maximize:

$$\max_{\beta} \left\{ \sum_i y_i x_i \beta - \ln\left[1 + \exp\{x_i \beta\}\right] \right\}$$

- Again, do this on a computer!