Maximum likelihood estimation

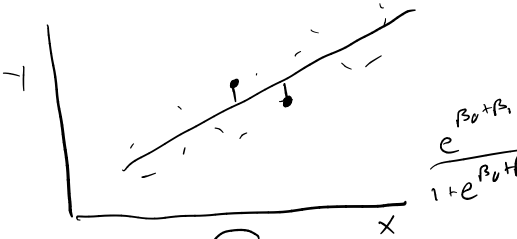
Ciaran Evans

Fitting a *logistic* regression model?

Linear regression: minimize $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2$

Question: Should we minimize a similar sum of squares for a *logistic* regression model? NO

Linear regression



$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

additive error term

Logistic regression

$\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x}}$

$Y_i | X_i \sim \text{Bernoulli}(p_i)$

$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$

Motivation: likelihoods and estimation

Let $Y \sim \textit{Bernoulli}(p)$ be a Bernoulli random variable, with $p \in [0, 1]$. We observe 5 independent samples from this distribution:

$$Y_1 = 1,\ Y_2 = 1,\ Y_3 = 0,\ Y_4 = 0,\ Y_5 = 1$$

The true value of p is unknown, so two friends propose different guesses for the value of p: 0.3 and 0.7. Which do you think is a "better" guess? $\hat{p} = 0.6$ (closer to 0.7)

more 1s than 0s $\Rightarrow$ 0.7 better guess than 0.3

$P(\text{data} \mid p = 0.3) = (0.3)^3 (1-0.3)^2 = 0.013$

$P(\text{data} \mid p = 0.7) = (0.7)^3 (1-0.7)^2 = 0.031$

Intuition: choose value of p that makes data "more likely"

Likelihood

Definition: Let $\mathbf{y} = (Y_1, ..., Y_n)$ be a sample of n observations, and let $f(\mathbf{y}|\theta)$ denote the joint pdf or pmf of $\mathbf{y}$, with parameter(s) θ. The *likelihood function* is

"probability" of the observed data, if θ is the true parameter

$$\underbrace{L(\theta|\mathbf{y})}_{\text{function of } \theta, \text{ given observed data } Y} = f(\mathbf{y}|\theta)$$

$L(\theta|Y) \geq 0$ since $f(y|\theta) \geq 0$

Special case: $Y_1, \ldots, Y_n$ are iid

$$\Rightarrow L(\theta|Y) = \prod_{i=1}^{n} f(Y_i|\theta)$$

Example: Bernoulli data

Let $Y_1, \dots, Y_n \overset{iid}{\sim}$ Bernoulli(p)

$$f(y_i \mid p) = P(Y = y_i) = p^{y_i} (1-p)^{1-y_i} \qquad y_i \in \{0,1\}$$

$$L(p \mid Y_1, \dots, Y_n) = \prod_{i=1}^{n} f(Y_i \mid p)$$

$$= \prod_{i=1}^{n} p^{Y_i} (1-p)^{1-Y_i}$$

$$= p^{\Sigma_i Y_i} (1-p)^{n - \Sigma_i Y_i}$$

<u>ex</u>: $Y = (1, 1, 0, 0, 1)$

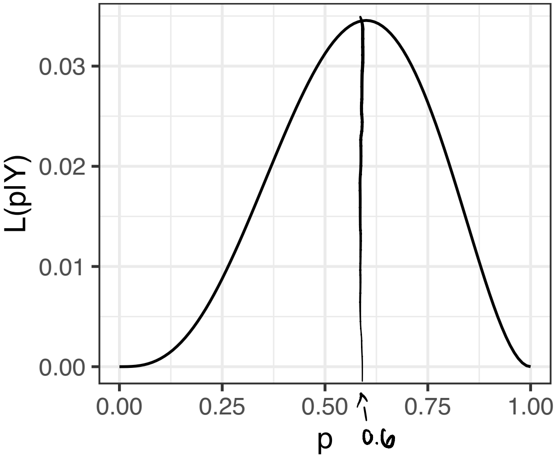
$$L(p \mid Y) = p^3 (1-p)^2$$

Example: Bernoulli data

$Y_1, ..., Y_5 \overset{iid}{\sim} Bernoulli(p)$, with observed data

$$Y_1 = 1, \;\; Y_2 = 1, \;\; Y_3 = 0, \;\; Y_4 = 0, \;\; Y_5 = 1$$

$L(p|\mathbf{y}) = p^3(1-p)^2$



Maximum likelihood estimator

Definition: Let $\mathbf{y} = (Y_1, ..., Y_n)$ be a sample of n observations. The *maximum likelihood estimator* (MLE) is

$$\widehat{\theta} = \text{argmax}_\theta \ L(\theta|\mathbf{y})$$

"value of θ that maximizes..."

Example: *Bernoulli(p)*

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bernoulli}(p)$$

$$L(p \mid Y) = p^{\sum_i Y_i} (1-p)^{n - \sum_i Y_i}$$

maximize to estimate p:

1. Take log to make life easier:

$$\ell(p \mid Y) = \log L(p \mid Y)$$

$$= \left(\sum_i Y_i\right) \log p + \left(n - \sum_i Y_i\right) \log(1-p)$$

2. Differentiate wrt parameter of interest:

$$\frac{d}{dp} \ell(p \mid Y) = \frac{\sum_i Y_i}{p} - \frac{\left(n - \sum_i Y_i\right)}{1-p} \overset{set}{=} 0$$

$$\Rightarrow \frac{\sum_i Y_i}{p} = \frac{n - \sum_i Y_i}{1-p}$$

$$\Rightarrow \hat{p} = \frac{1}{n} \sum_i Y_i = \bar{Y}$$

(sample proportion!)

Example: *Bernoulli*(p)

$$\frac{d}{dp} \ell(p|Y) = \frac{\sum_i Y_i}{p} - \frac{(n - \sum_i Y_i)}{1-p}$$

Check second derivative:

$$\frac{d^2}{dp^2} \ell(p|Y)\Big|_{p=\bar{Y}} = -\frac{\sum_i Y_i}{p^2} - \frac{(n - \sum_i Y_i)}{(1-p)^2}\Big|_{p=\bar{Y}}$$

$$< 0$$

$\Rightarrow \quad \hat{p} = \frac{1}{n} \sum_i Y_i = \bar{Y}$ maximizes $\ell(p|Y)$

Example: $N(\theta, 1)$

$Y_1, \dots, Y_n \overset{iid}{\sim} N(\theta, 1)$ $\qquad \theta \in (-\infty, \infty)$

$$f(y_i | \theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y_i - \theta)^2\right\}$$

$$L(\theta | Y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(Y_i - \theta)^2\right\}$$

$$= (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_i (Y_i - \theta)^2\right\}$$

$$\ell(\theta | Y) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (Y_i - \theta)^2$$

$$\frac{d}{d\theta} \ell(\theta | Y) = -\frac{1}{2} \sum_{i=1}^{n} 2(Y_i - \theta)(-1) = \sum_{i=1}^{n} (Y_i - \theta) \overset{set}{=} 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i = n\theta$$

$$\Rightarrow \boxed{\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}}$$

$$\frac{d^2}{d\theta^2} \ell(\theta | Y) = -n < 0 \quad \checkmark$$

$\Rightarrow$ unique maximum