Lecture 3: Maximum likelihood estimation

Ciaran Evans

Logistics

- HW 1 due Friday on Canvas
- Office hours:
  - Wednesday 2-3 pm
  - Thursday 9:30-10:30 am
- Bowling on Friday!

Motivation: fitting a *linear* regression model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_k X_{i,k} + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

Suppose we observe data $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$, where $X_i = (1, X_{i,1}, ..., X_{i,k})^T$.

How do we fit this linear regression model? That is, how do we estimate

$$\beta = (\beta_0, \beta_1, ..., \beta_k)^T$$

Minimize sum of squared errors (aka residual sum of squares)

choose $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ to minimize

$$SSE = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})^2$$

$$\left.\begin{aligned} \frac{\partial SSE}{\partial \beta_0} &\overset{set}{=} 0 \\ &\vdots \\ \frac{\partial SSE}{\partial \beta_k} &\overset{set}{=} 0 \end{aligned}\right\}$$

$k+1$ unknowns

$k+1$ equations

estimate $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$ solves this system

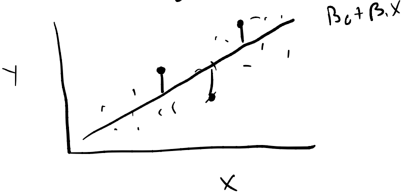
Fitting a *logistic* regression model?

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

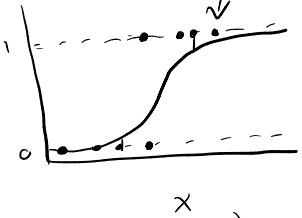
Question: Should we minimize a similar sum of squares for a *logistic* regression model? $\underline{\text{No}}$

$$\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Linear regression



Logistic regression



Linear regression:

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

error terms

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

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

not same idea of "residual"

Motivation: likelihoods and estimation

Let $Y \sim Bernoulli(p)$ 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?

sample proportion: 0.6 (closer to 0.7)

$P(\text{data} \mid p = 0.3) = (0.3)^3 (0.7)^2 = 0.013$

$P(\text{data} \mid p = 0.7) = (0.7)^3 (0.3)^2 = 0.031$

Intuition: choose value of p which 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

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

"probability" of the observed data, <u>if</u> θ is the true parameter

$L(\theta|Y)$: condition on observed data, and we want to know how the joint density/mass function of Y changes as a function of θ

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

<u>special case</u> : $Y_1, \ldots, Y_n$ iid $\quad L(\theta|Y) = \prod_{i=1}^{n} f(Y_i|\theta)$

Example: Bernoulli data

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

$P(Y=y)$

$f(y|p) = p^y (1-p)^{1-y}$

(single observation)

$y \in \{0, 1\}$

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

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

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

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

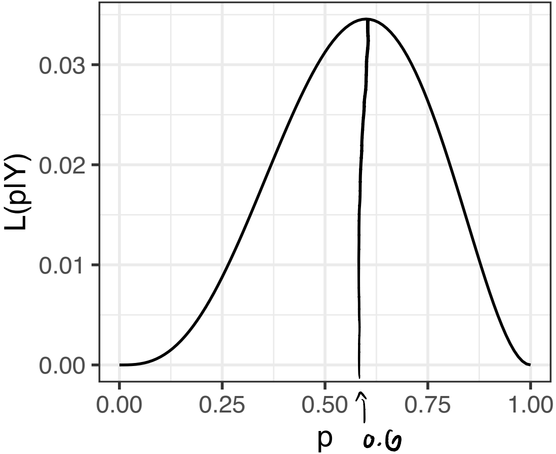
$$L(p|Y) = p^3 (1-p)^2$$

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

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

argmax_θ means "value of θ that maximizes..."

Example: *Bernoulli*(p)

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

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

maximize to estimate p:

① Take log to make life easier

- log is monotone, increasing, so if $\hat{p}$ maximize

$L(p|Y) \iff \log L(p|Y)$

$$\ell(p|Y) = \log L(p|Y) = \left(\sum_i Y_i\right) \log p + \left(n - \sum_i Y_i\right) \log(1-p)$$

② Differentiate wrt parameter of interest:

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

$$\frac{\sum_i Y_i}{p} - \frac{(n - \sum_i Y_i)}{1-p} \stackrel{set}{=} 0$$

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

(sample proportion!)

Check second derivative:

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

$$< 0$$

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

Sidebar: $\sum_i (Y_i - \beta_0)^2$ is minimized by

$$\beta_0 = \bar{Y}$$

i.e. $\sum_i (Y_i - \bar{Y})^2 \leq \sum_i (Y_i - \beta_0)^2$

$\forall \beta_0$