Lecture 3: Maximum likelihood estimation

Recap: ways of fitting a *linear* regression model

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

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

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

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

- Minimize sum of squared residuals
- Maximum likelihood estimation (assume distribution for Y)
- Projection

Minimize SSE

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

Find $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ minimize SSE

$$\frac{\partial SSE}{\partial \beta_0} \overset{\text{set}}{=} 0$$

$$\frac{\partial SSE}{\partial \beta_1} \overset{\text{set}}{=} 0$$

$$\vdots$$

$$\frac{\partial SSE}{\partial \beta_k} \overset{\text{set}}{=} 0$$

$k+1$ equations

$k+1$ unknowns

$\hat{\beta}$ vector that solves this system

Projection

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

$$X = \begin{pmatrix} 1 & x_{11} & \cdots \\ 1 & x_{21} & \cdots \\ \vdots & \vdots & \\ 1 & x_{n1} & \cdots \end{pmatrix} \in \mathbb{R}^{n \times (M+1)}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_M \end{pmatrix}$$

$$\hat{Y} = X\hat{\beta} \qquad (\hat{Y} \in \text{colspace}(X))$$

want $\hat{Y}$ "close" to Y

minimize $\|Y - \hat{Y}\| = \sqrt{SSE}$

(Euclidean norm)

$\Leftrightarrow$ minimize SSE

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik},\ \sigma_\varepsilon^2)$$

$$e^a e^b = e^{a+b}$$

$$L(\beta_0, \cdots, \beta_k, \sigma_\varepsilon^2 \mid (X_1, Y_1), \cdots, (X_n, Y_n))$$

(independence)

$$= \prod_{i=1}^{n} f(Y_i \mid \beta_0, \cdots, \beta_k, \sigma_\varepsilon^2, X_i)$$

$$\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} (Y_i - \beta_0 - \beta_1 X_i - \cdots - \beta_k X_{ik})^2 \right\}$$

$$= (2\pi)^{-\frac{n}{2}} \sigma_\varepsilon^{-n} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \underbrace{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2}_{SSE} \right\}$$

maximizing likelihood $\Longleftrightarrow$ minimizing SSE
(for normal data)

Summary: three ways of fitting linear regression models

- Minimize SSE, via derivatives of

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

SSE: might not work for binary data (what should the residual be??)

- Minimize $||Y - \hat{Y}||$ (equivalent to minimizing SSE)
- Maximize likelihood (for *normal* data, equivalent to minimizing SSE)

← might work for logistic, but w/ different distribution (Bernulli?)

Which of these three methods, if any, is appropriate for fitting a logistic regression model? Do any changes need to be made for the logistic regression setting?

Step back: likelihoods and estimation

Let $Y \sim \text{Bernoulli}(p)$ 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)

$P(\text{data} \mid p=0.3) = (0.3)(0.3)(1-0.3)(1-0.3)(0.3)$

$= 0.3^3 (0.7)^2 = 0.013$

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

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

Likelihood

Definition: Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ 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

$$L(\theta|\mathbf{Y}) = f(\mathbf{Y}|\theta)$$

function of θ, given Y

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

$L(\theta|Y)$: condition on the observed data. Want to know how "probability" (joint density/mass function) of Y changes as a function of θ

- Since $f(y|\theta) \geq 0$, $L(\theta|Y) \geq 0 \quad \forall \theta$
- Special case: $Y_1, \ldots, Y_n$ iid w/ pdf or pmf f

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

Example: Bernoulli data

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

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

$$P(Y=1) = p^1 (1-p)^{1-1} = p$$

$$P(Y=0) = p^0 (1-p)^{1-0} = 1-p$$

$$L(p \mid Y_1, \ldots, Y_n) \overset{(\text{indep.})}{=} \prod_{i=1}^{n} f(Y_i \mid p)$$

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

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

Ex: $Y = (1, 1, 0, 0, 1)$

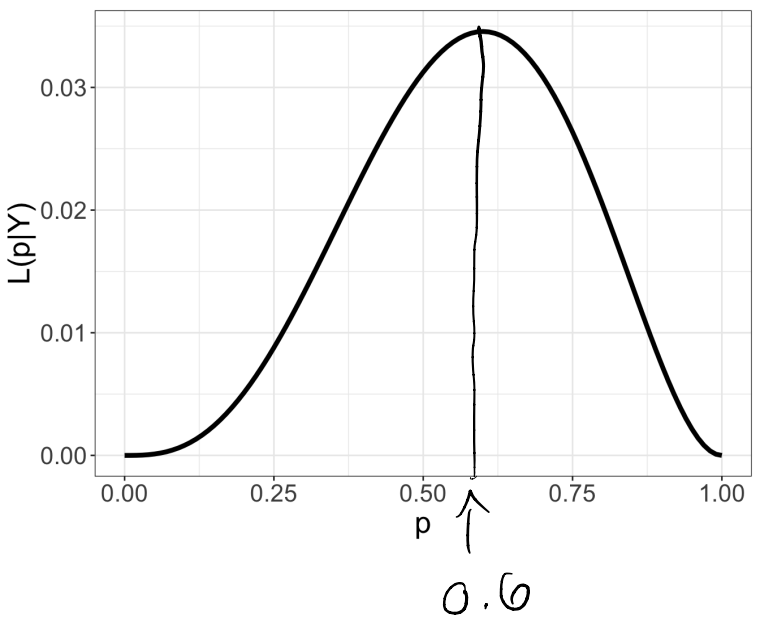
$$L(p \mid Y) = p^3 (1-p)^2$$

Example: Bernoulli data

$Y_1, \ldots, Y_5 \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, 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 $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be a sample of n observations. The *maximum likelihood estimator* (MLE) is

$$\hat{\theta} = \operatorname{argmax}_\theta L(\theta|\mathbf{Y})$$

argmax$_\theta$ means "value of θ that maximizes ..."

Example: Bernoulli(p)

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

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

Now maximize to estimate p

(1) Take log to make life easier (log is monotone increasing function, so if $\hat{p}$ maximizes $\log L(p|Y)$, then $\hat{p}$ maximizes $L(p|Y)$)

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

(2) Differentiate wrt parameter of interest:

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

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

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

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

(sample proportion!)

So $\ell(p|Y)$ has a max or a min at $p = \frac{1}{n}\sum_i Y_i$

Check second derivative:

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

$\Rightarrow \hat{p} = \frac{1}{n}\sum_i Y_i$ maximizes $\ell(p|Y)$ $\overset{<}{0}$

check $p = 0$ and $p = 1$

if $p = 0$: $L(p|Y) = \begin{cases} 0 & \text{if any } Y_i = 1 \\ 1 & \text{if all } Y_i = 0 \end{cases}$

$p = 1$: $L(p|Y) = \begin{cases} 0 & \text{if any } Y_i = 0 \\ 1 & \text{if all } Y_i = 1 \end{cases}$

if all $Y_i = 1 \Rightarrow \hat{p} = 1 = \frac{1}{n}\sum_i Y_i$

if all $Y_i = 0 \Rightarrow \hat{p} = 0 = \frac{1}{n}\sum_i Y_i$