Lecture 3: Maximum likelihood estimation

Recap: ways of fitting a *linear* regression model

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

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

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

Summary: three ways of fitting linear regression models

- Minimize SSE, via derivatives of
  $$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i,1} - \cdots - \beta_k X_{i,k})^2$$
- Minimize $||Y - \widehat{Y}||$ (equivalent to minimizing SSE)
- Maximize likelihood (for *normal* data, equivalent to minimizing SSE)

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 $Y \sim \mathrm{Bernoulli}(p)$ 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?

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

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

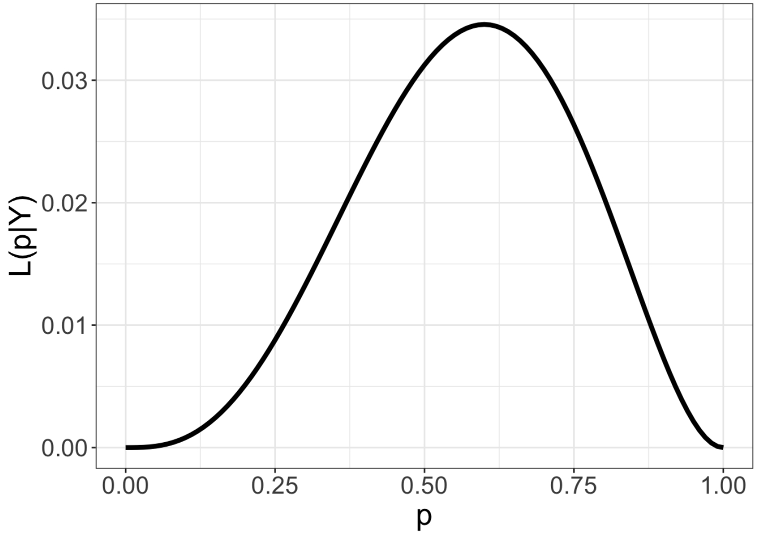
Example: Bernoulli data

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

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

Example: Bernoulli(p)