Maximum Likelihood Estimation

February 20, 2013

Inference

- We want to estimate parametric models of the form:

$$y \sim f(\theta)$$

Inference

- We want to estimate parametric models of the form:

$$y \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y (the data).

Inference

- We want to estimate parametric models of the form:

$$y \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y (the data).
- What social scientists wish we could do:
  - Estimate $P(\text{unknown}|\text{known})$ or $P(f(\theta)|(y))$

Inference

- We want to estimate parametric models of the form:

$$y \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y (the data).
- What social scientists wish we could do:
  - Estimate $P(\text{unknown}|\text{known})$ or $P(f(\theta)|(y))$
  - This is the problem of inverse probability: what is the probability of the model given the data?

Posterior Inference

- What about the reverse?
    - Estimate $P(\theta|f(\cdot), y)$ or, more simply, $P(\theta|y)$

Posterior Inference

- What about the reverse?
  - Estimate $P(\theta|f(\cdot), y)$ or, more simply, $P(\theta|y)$
- Apply Bayes' theorem:

$$P(\theta|y) = \frac{P(\theta, y)}{P(y)}$$

Posterior Inference

- What about the reverse?
  - Estimate $P(\theta|f(\cdot), y)$ or, more simply, $P(\theta|y)$
- Apply Bayes' theorem:

$$
\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)}
\end{aligned}
$$

Posterior Inference

- What about the reverse?
  - Estimate $P(\theta|f(\cdot), y)$ or, more simply, $P(\theta|y)$

- Apply Bayes' theorem:

$$
\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)}
\end{aligned}
$$

- The Prior: $P(\theta)$
- The Sampling (Data): $P(y|\theta)$
- The Posterior: $P(\theta|y)$
- The Marginal: $P(y)$

Posterior Inference

- What about the reverse?
  - Estimate $P(\theta|f(\cdot), y)$ or, more simply, $P(\theta|y)$
- Apply Bayes' theorem:

$$\begin{aligned} P(\theta|y) &= \frac{P(\theta, y)}{P(y)} \\ &= \frac{P(\theta)P(y|\theta)}{P(y)} \end{aligned}$$

- The Prior: $P(\theta)$
- The Sampling (Data): $P(y|\theta)$
- The Posterior: $P(\theta|y)$
- The Marginal: $P(y)$

The Likelihood Approach

- Like randomization as the "reasoned basis of inference", developed by R.A. Fisher (in his junior year).

The Likelihood Approach

- Like randomization as the "reasoned basis of inference", developed by R.A. Fisher (in his junior year).
- Following frequentist principles: θ is fixed and y is random. Define the likelihood:

$$\begin{aligned} \mathcal{L}(\theta|y) &= \mathcal{K}(y)P(y|\theta) \\ &\propto P(y|\theta) \end{aligned}$$

where $\mathcal{K}(y)$ is unknown, but does not depend on θ.

The Likelihood Approach

- Like randomization as the “reasoned basis of inference”, developed by R.A. Fisher (in his junior year).
- Following frequentist principles: θ is fixed and y is random. Define the likelihood:

$$\begin{aligned} \mathcal{L}(\theta|y) &= \mathcal{K}(y)P(y|\theta) \\ &\propto P(y|\theta) \end{aligned}$$

  where $\mathcal{K}(y)$ is unknown, but does not depend on θ.
- Fixing y at observed values, and assuming a model of the sampling distribution $P(y|\theta)$, $\mathcal{L}(\theta|y)$ measures the “likelihood” that $\theta = \hat{\theta}$

The Likelihood Approach

- Like randomization as the "reasoned basis of inference", developed by R.A. Fisher (in his junior year).

- Following frequentist principles: θ is fixed and y is random. Define the likelihood:

$$\begin{aligned} \mathcal{L}(\theta|y) &= \mathcal{K}(y)P(y|\theta) \\ &\propto P(y|\theta) \end{aligned}$$

  where $\mathcal{K}(y)$ is unknown, but does not depend on θ.

- Fixing y at observed values, and assuming a model of the sampling distribution $P(y|\theta)$, $\mathcal{L}(\theta|y)$ measures the "likelihood" that $\theta = \hat{\theta}$
- Then $\hat{\theta}$ is found at $\arg\max_\theta \mathcal{L}(\theta|y)$
  - Note: $\arg\max_\theta \mathcal{L}(\theta|y) \implies \arg\max_\theta P(y|\theta)$

OLS in MLE

- We randomly sample (iid) n units, and observe y_i and X_i. Assuming $y \sim N(X\beta, \sigma^2)$:

$$
\begin{aligned}
P(y|\theta, X) &= \prod_{i=1}^{n} p(y_i|\theta) \\
\mathcal{L}(y|\beta, \sigma, X) &\propto \prod_{i=1}^{n} N(y_i|X_i\beta, \sigma^2) \\
&\propto \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2\right\}
\end{aligned}
$$

OLS in MLE

- Easier to evaluate the $log(\mathcal{L}) = \ell$. In this case $\ell(y|\beta, \sigma, X)$:

$$\propto \sum_{i=1}^{n} \left\{ -\log(\sigma) - \frac{1}{2}\log(2\pi) - \frac{1}{2\sigma^2}(y_i - X_i\beta)^2 \right\}$$

$$\propto -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2$$

OLS in MLE

- Easier to evaluate the $log(\mathcal{L}) = \ell$. In this case $\ell(y|\beta, \sigma, X)$:

$$\propto \sum_{i=1}^{n} \left\{ -\log(\sigma) - \frac{1}{2}\log(2\pi) - \frac{1}{2\sigma^2}(y_i - X_i\beta)^2 \right\}$$

$$\propto -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2$$

- How do we find $\hat{\beta}$?
  - Maximize $\ell(y|\beta, \sigma, X)$ w/r/t to β
  - Solve $\frac{\partial \ell(y|\beta,\sigma,X)}{\partial \beta} = 0$, to find $\hat{\beta}$ i
  - Solve $\frac{\partial^2 \ell(y|\beta,\sigma,X)}{\partial \beta^2} < 0$, to ensure $\hat{\beta}$ is a maximum
  - IF derivatives exist, can solve analytically

Bivariate Normal with $\sigma = 1$

- Here:

$$\ell(y|\beta, \sigma, x) \quad \propto \quad -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\beta)^2$$

Bivariate Normal with $\sigma = 1$

- Here:

$$\ell(y|\beta, \sigma, x) \quad \propto \quad -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\beta)^2$$

- First derivative:

$$\begin{aligned} \frac{\partial \ell(y|\beta, \sigma, X)}{\partial \beta} &= \sum_{i=1}^{n}\left(X_i y_i - X_i^2\beta\right) \\ &= \sum_{i=1}^{n} X_i y_i - \beta\sum_{i=1}^{n} X_i^2 \end{aligned}$$

Bivariate Normal with $\sigma = 1$

- Here:

$$\ell(y|\beta,\sigma,x) \propto -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\beta)^2$$

- First derivative:

$$\frac{\partial \ell(y|\beta,\sigma,X)}{\partial \beta} = \sum_{i=1}^{n}\left(X_i y_i - X_i^2\beta\right) = \sum_{i=1}^{n} X_i y_i - \beta\sum_{i=1}^{n} X_i^2$$

- Setting to zero and solving gives us (implicitly assuming zero intercept):

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i y_i}{\sum_{i=1}^{n} X_i^2}$$

Bivariate Normal with $\sigma = 1$

- Check the second derivative:

$$\frac{\partial^2 \ell(y|\beta, \sigma, X)}{\partial \beta^2} = \frac{\partial \sum_{i=1}^{n} \left(X_i y_i - X_i^2 \beta \right)}{\partial \beta}$$

$$= -\sum_{i=1}^{n} X_i^2 < 0$$

Multivariate Normal MLE

- Again $\ell(y|\beta, \sigma, x)$:

$$\begin{aligned}
&\propto -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta) \\
&\propto -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\left(\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{y}'\mathbf{x} + \beta'\mathbf{x}'\mathbf{x}\beta\right)
\end{aligned}$$

Multivariate Normal MLE

- Again $\ell(y|\beta, \sigma, x)$:

$$\begin{aligned}
&\propto -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta) \\
&\propto -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\left(\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{y}'\mathbf{x} + \beta'\mathbf{x}'\mathbf{x}\beta\right)
\end{aligned}$$

- First derivative:

$$\frac{\partial \ell(\mathbf{y}|\beta, \sigma, \mathbf{x})}{\partial \beta} = \frac{1}{\sigma^2}\left(\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta\right)$$

Multivariate Normal MLE

- Again $\ell(y|\beta,\sigma,x)$:

$$\begin{aligned}
&\propto -\frac{n}{2}\log(\sigma^2)-\frac{n}{2}\log(2\pi)-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{x}\beta)'(\mathbf{y}-\mathbf{x}\beta)\\
&\propto -\frac{n}{2}\log(\sigma^2)-\frac{n}{2}\log(2\pi)-\frac{1}{2\sigma^2}\left(\mathbf{y}'\mathbf{y}-2\beta'\mathbf{y}'\mathbf{x}+\beta'\mathbf{x}'\mathbf{x}\beta\right)
\end{aligned}$$

- First derivative:

$$\frac{\partial\ell(\mathbf{y}|\beta,\sigma,\mathbf{x})}{\partial\beta} = \frac{1}{\sigma^2}\left(\mathbf{x}'\mathbf{y}-\mathbf{x}'\mathbf{x}\beta\right)$$

- Setting to zero and solving:

$$\begin{aligned}
0 &= \frac{1}{\sigma^2}\left(\mathbf{x}'\mathbf{y}-\mathbf{x}'\mathbf{x}\beta\right)\\
\mathbf{x}'\mathbf{x}\beta &= \mathbf{x}'\mathbf{y}\\
\hat{\beta} &= \left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\mathbf{y}
\end{aligned}$$

Estimating the Variance σ^2

- To estimate σ^2, take the derivative w/r/t to σ^2:

$$\begin{aligned}\ell &\propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}(\sigma^2)^{-1}(\mathbf{y}-\mathbf{x}\beta)'(\mathbf{y}-\mathbf{x}\beta)\\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{x}\beta)'(\mathbf{y}-\mathbf{x}\beta)\end{aligned}$$

Estimating the Variance σ^2

- To estimate σ^2, take the derivative w/r/t to σ^2:

$$
\begin{aligned}
\ell &\propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}(\sigma^2)^{-1}(\mathbf{y}-\mathbf{x}\beta)'(\mathbf{y}-\mathbf{x}\beta) \\
\frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{x}\beta)'(\mathbf{y}-\mathbf{x}\beta)
\end{aligned}
$$

- Setting to zero and solving:

$$
\begin{aligned}
n &= \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{x}\beta)'(\mathbf{y}-\mathbf{x}\beta) \\
n &= \frac{1}{\sigma^2}(\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}}) \\
n &= \frac{1}{\sigma^2}\mathbf{e}'\mathbf{e} \\
\hat{\sigma}^2 &= \frac{\mathbf{e}'\mathbf{e}}{n} \qquad (1)
\end{aligned}
$$

Fisher Information

Fisher Theorem: Suppose $X_1, X_2, ..., X_n$ are IID with probability distribution governed by the parameter θ. Let θ_0 be the true value of θ. Under regularity conditions, the MLE for θ is asymptotically normal. The asymptotic mean of the MLE is θ_0. The asymptotic variance can be computed as follows:

$$\left[-\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial\theta\theta'}\right]^{-1}$$

If $\hat{\theta}$ is the MLE and v_n is the asymptotic variance, the theorem says that $\frac{\hat{\theta}-\theta}{\sqrt{v_n}} \sim N(0,1)$ (nearly), when the sample size n is large

The Hessian

- Let θ be a vector containing the paramters being estimated. For example, in the regression $y = \alpha + \beta x + \epsilon$ with variance σ; $\theta : \{\alpha, \beta, \sigma\}$.
- The *Hessian* is a matrix of second derivatives defined as:

$$\mathbf{H}(\theta) = \frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}$$

$$\mathbf{H}(\theta) = \begin{pmatrix} \frac{\partial^2 \ell(\theta)}{\partial\alpha\partial\alpha} & \frac{\partial^2 \ell(\theta)}{\partial\alpha\partial\beta} & \frac{\partial^2 \ell(\theta)}{\partial\alpha\partial\sigma} \\ \frac{\partial^2 \ell(\theta)}{\partial\beta\partial\alpha} & \frac{\partial^2 \ell(\theta)}{\partial\beta\partial\beta} & \frac{\partial^2 \ell(\theta)}{\partial\beta\partial\sigma} \\ \frac{\partial^2 \ell(\theta)}{\partial\sigma\partial\alpha} & \frac{\partial^2 \ell(\theta)}{\partial\sigma\partial\beta} & \frac{\partial^2 \ell(\theta)}{\partial\sigma\partial\sigma} \end{pmatrix}$$

The Information Matrix

- The information matrix is defined as the negative of the expected value of the Hessian: $-E[\mathbf{H}(\theta)]$. Under very general conditions, the covariance matrix for the ML estimator is the inverse of the information matrix:

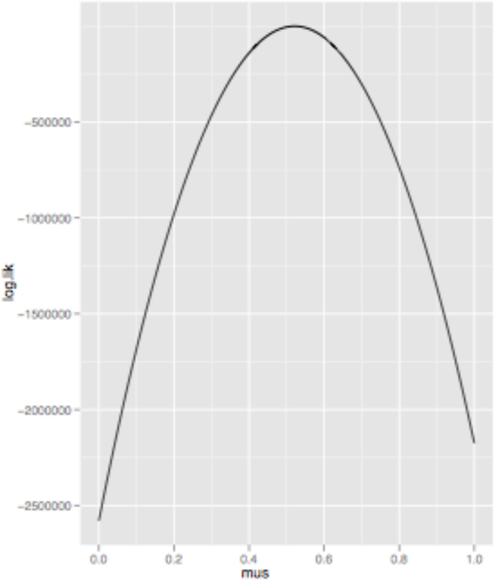
$$Var(\hat{\theta}) = -E[\mathbf{H}(\theta)]^{-1}$$

$$Var(\theta) = \begin{pmatrix} -E\left(\frac{\partial^2 \ell(\theta)}{\partial\alpha\partial\alpha}\right) & -E\left(\frac{\partial^2 \ell(\theta)}{\partial\alpha\partial\beta}\right) & -E\left(\frac{\partial^2 \ell(\theta)}{\partial\alpha\partial\sigma}\right) \\ -E\left(\frac{\partial^2 \ell(\theta)}{\partial\beta\partial\alpha}\right) & -E\left(\frac{\partial^2 \ell(\theta)}{\partial\beta\partial\beta}\right) & -E\left(\frac{\partial^2 \ell(\theta)}{\partial\beta\partial\sigma}\right) \\ -E\left(\frac{\partial^2 \ell(\theta)}{\partial\sigma\partial\alpha}\right) & -E\left(\frac{\partial^2 \ell(\theta)}{\partial\sigma\partial\beta}\right) & -E\left(\frac{\partial^2 \ell(\theta)}{\partial\sigma\partial\sigma}\right) \end{pmatrix}^{-1}$$

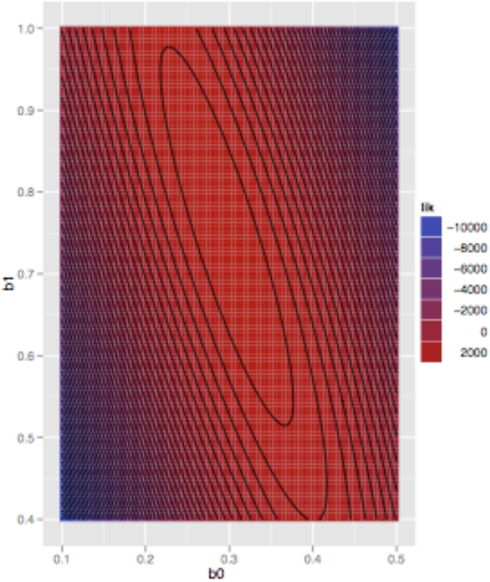
Estimating θ Numerically

- No analytical solution for derivatives to solve for $\hat{\theta}$, due to non-linearities, complex functions, etc
- Brute-force methods:
  - Search over all permissible values of θ

Likelihood for One Parameter



Likelihood for Two Parameters



Estimating θ Numerically

- No analytical solution for derivatives to solve for $\hat{\theta}$, due to non-linearities, complex functions, etc
- Brute-force methods:
  - Search over all permissible values of θ
  - Very inefficient
  - Impossible as dimensions in θ increase

Newton Algorithm

- Recall that maximum at $\mathcal{L}(\beta)$ when $\mathcal{L}'(\beta) = 0$
- Find root of $\mathcal{L}'(\beta)$ using an iterative algorithm to approximate a quadratic function around β_i

Newton Algorithm

- Recall that maximum at $\mathcal{L}(\beta)$ when $\mathcal{L}'(\beta) = 0$
- Find root of $\mathcal{L}'(\beta)$ using an iterative algorithm to approximate a quadratic function around β_i
- Newton Algorithm:
  - Start with initial values β_0

Newton Algorithm

- Recall that maximum at $\mathcal{L}(\beta)$ when $\mathcal{L}'(\beta) = 0$
- Find root of $\mathcal{L}'(\beta)$ using an iterative algorithm to approximate a quadratic function around β_i
- Newton Algorithm:
  - Start with initial values β_0
  - So long as $\mathcal{L}'(\beta) \neq 0$

Newton Algorithm

- Recall that maximum at $\mathcal{L}(\beta)$ when $\mathcal{L}'(\beta) = 0$
- Find root of $\mathcal{L}'(\beta)$ using an iterative algorithm to approximate a quadratic function around β_i
- Newton Algorithm:
  - Start with initial values β_0
  - So long as $\mathcal{L}'(\beta) \neq 0$
    - Improve guess about β by linear approximation at β_i
    - Accordingly: $\beta_{i+1} = \beta_i - \frac{\mathcal{L}'(\beta_i)}{\mathcal{L}''(\beta_i)}$
    - Matrix notation this is: $\beta_{i+1} = \beta_i + [-\mathbf{H}(\beta_i)]^{-1} g(\beta_i)$, where $g(\beta)$ is the matrix of first derivatives

Newton Algorithm

- Recall that maximum at $\mathcal{L}(\beta)$ when $\mathcal{L}'(\beta) = 0$
- Find root of $\mathcal{L}'(\beta)$ using an iterative algorithm to approximate a quadratic function around β_i
- Newton Algorithm:
  - Start with initial values β_0
  - So long as $\mathcal{L}'(\beta) \neq 0$
    - Improve guess about β by linear approximation at β_i
    - Accordingly: $\beta_{i+1} = \beta_i - \frac{\mathcal{L}'(\beta_i)}{\mathcal{L}''(\beta_i)}$
    - Matrix notation this is: $\beta_{i+1} = \beta_i + [-\mathbf{H}(\beta_i)]^{-1} g(\beta_i)$, where $g(\beta)$ is the matrix of first derivatives
  - If approximation $\beta_i - \frac{\mathcal{L}'(\beta_i)}{\mathcal{L}''(\beta_i)} > 0$, decrease β_i
  - If approximation $\beta_i - \frac{\mathcal{L}'(\beta_i)}{\mathcal{L}''(\beta_i)} < 0$, increase β_i

Newton Algorithm

- Recall that maximum at $\mathcal{L}(\beta)$ when $\mathcal{L}'(\beta) = 0$
- Find root of $\mathcal{L}'(\beta)$ using an iterative algorithm to approximate a quadratic function around β_i
- Newton Algorithm:
  - Start with initial values β_0
  - So long as $\mathcal{L}'(\beta) \neq 0$
    - Improve guess about β by linear approximation at β_i
    - Accordingly: $\beta_{i+1} = \beta_i - \frac{\mathcal{L}'(\beta_i)}{\mathcal{L}''(\beta_i)}$
    - Matrix notation this is: $\beta_{i+1} = \beta_i + [-\mathbf{H}(\beta_i)]^{-1} g(\beta_i)$, where $g(\beta)$ is the matrix of first derivatives
  - If approximation $\beta_i - \frac{\mathcal{L}'(\beta_i)}{\mathcal{L}''(\beta_i)} > 0$, decrease β_i
  - If approximation $\beta_i - \frac{\mathcal{L}'(\beta_i)}{\mathcal{L}''(\beta_i)} < 0$, increase β_i
- Often must compute numerical derivatives; very inefficient

Optimization

- Many additional optimization methods:
  - Newton-Rhapson
  - BFGS
  - genoud
- Some of these emphasize efficiency
  - Newton-Rhapson
  - BFGS
- Others emphasize robustness to saddle points and significant irregularities
  - genoud

Binomial Model(s) in MLE

- For n (iid) random trials, y measures whether the ith trial was a 'success,' so $y \sim Binomial(p)$, where p is the probability of success
- Define $m = \sum y_i$ to be the number of successes. Following the binomial:

$$
\begin{aligned}
\mathcal{L} &\propto \prod_{i=1}^{n} p^{y_i}(1 - p^{1-y_i}) \\
\ell &\propto \sum_{i=1}^{n} [y_i log(p) + (1 - y_i) log(1 - p)] \\
&\propto log(p) \sum_{i=1}^{n} y_i + log(1 - p) \sum_{i=1}^{n} (1 - y_i) \\
&\propto log(p)m + log(1 - p)(n - m)
\end{aligned}
$$

Binomial Model(s) in MLE

- First derivative:

$$\begin{aligned}\frac{\partial \ell}{\partial p} &= \frac{\partial log(p)m + log(1-p)(n-m)}{\partial p} \\ &= \frac{m}{p} - \frac{n-m}{1-p}\end{aligned}$$

- Setting to zero and solving:

$$\begin{aligned}\frac{m}{p} &= \frac{n-m}{1-p} \\ \hat{p} &= \frac{m}{n}\end{aligned}$$

Binomial Model(s) in MLE

- Second derivative:

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial p^2} &= \frac{\partial \left(\frac{m}{p} - \frac{n-m}{1-p} \right)}{\partial p} \\
&= -\frac{m}{p^2} - \frac{n-m}{(1-p)^2} < 0
\end{aligned}
$$

The Probit Model

- Let y_i be a binary response variable that is a function of covariates x_i
- Given x_i, the y_i responses are assumed independent random variables where:

$$P(y_i = 1|x_i) \quad = \quad \Phi(x_i\beta)$$

  where Φ is the CDF of the standard normal distribution
- The likelihood:

$$\begin{aligned}\mathcal{L}(x_i, \beta) &= \prod_{i=1}^{n} y_i \cdot \Phi(x_i\beta) \times (1 - y_i) \cdot \{1 - \Phi(x_i\beta)\} \\ \ell(x_i, \beta) &= \sum_{i=1}^{n} (y_i \log[\Phi(x_i\beta)] + (1 - y_i) \log\{1 - \Phi(x_i\beta)\})\end{aligned}$$