Maximum Likelihood Part Two[1]

STA 312 Fall 2023

[1] See last slide for copyright information.

Background Reading

Maximum likelihood handout (see course home page)

Overview

Two more issues

- Maximum likelihood estimates are often not available in closed form.
- Multiple parameters.

Most real-world problems have both these features.

No formula for the MLE

All we need is one example to see the problem.

Let $X_1, \ldots, X_n$ be independent observations from a distribution with density

$$f(x|\theta) = \begin{cases} \frac{1}{\Gamma(\theta)} e^{-x} x^{\theta-1} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Where the parameter $\theta > 0$. This is a gamma with $\alpha = \theta$ and $\lambda = 1$.

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ell(\theta) &= \frac{\partial}{\partial\theta} \log \left(\prod_{i=1}^{n} \frac{1}{\Gamma(\theta)} e^{-x_i} x_i^{\theta-1} \right) \\
&= \frac{\partial}{\partial\theta} \log \left(\Gamma(\theta)^{-n} e^{-\sum_{i=1}^n x_i} \left(\prod_{i=1}^{n} x_i \right)^{\theta-1} \right) \\
&= \frac{\partial}{\partial\theta} \left(-n \log \Gamma(\theta) - \sum_{i=1}^{n} x_i + (\theta-1) \sum_{i=1}^{n} \log x_i \right) \\
&= -\frac{n\Gamma'(\theta)}{\Gamma(\theta)} - 0 + \sum_{i=1}^{n} \log x_i \stackrel{set}{=} 0
\end{aligned}$$

Numerical MLE

By computer

- The log likelihood defines a surface sitting over the parameter space.
- It could have hills and valleys and mountains.
- The value of the log likelihood is easy to compute for any given set of parameter values.
- This tells you the height of the surface at that point.
- Take a step uphill (blindfolded).
- Are you at the top? Compute the slopes of some secant lines.
- Take another step uphill.
- How big a step? Good question.
- Most numerical routines *minimize* a function of several variables.
- So minimize the minus log likelihood.

Multiple parameters

Most real-world problems have a *vector* of parameters.

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with expected value μ and variance σ^2.
  The parameters μ and σ^2 are unknown.
- For $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$, where

  $\beta_0, \ldots, \beta_{p-1}$ are unknown constants.
  $x_{i,j}$ are known constants.
  $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.
  σ^2 is an unknown constant.
  $y_1, \ldots, y_n$ are observable random variables.

  The parameters $\beta_0, \ldots, \beta_{p-1}, \sigma^2$ are unknown.

Multi-parameter MLE

You know most of this.

- Suppose there are k parameters.
- The plane tangent to the log likelihood should be horizontal at the MLE.
- Partially differentiate the log likelihood (or minus log likelihood) with respect to each of the parameters.
- Set the partial derivatives to zero, obtaining k equations in k unknowns.
- Solve for the parameters, if you can.
- Is it really a maximum?
- There is a multivariate second derivative test.

The Hessian matrix

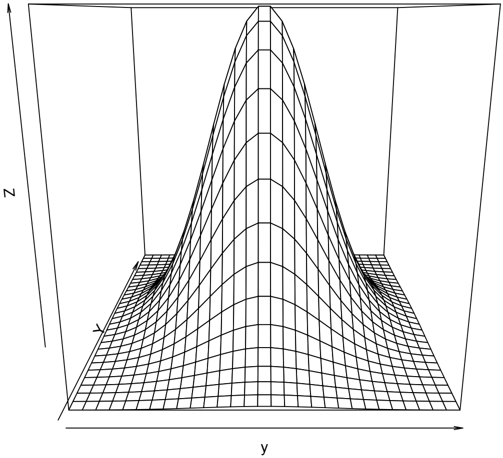
$$\mathbf{H} = \left[\frac{\partial^2(-\ell)}{\partial\theta_i \partial\theta_j} \right]$$

- If there are k parameters, the Hessian is a $k \times k$ matrix whose (i, j) element is $\frac{\partial^2}{\partial\theta_i \partial\theta_j}(-\ell(\boldsymbol{\theta}))$.
- If the second derivatives are continuous, $\mathbf{H}$ is symmetric.
- If the gradient is zero at a point and $|\mathbf{H}| \neq 0$, then
  - If all eigenvalues are positive at the point, local minimum.
  - If all eigenvalues are negative at the point, local maximum.
  - If there are both positive and negative eigenvalues at the point, saddle point.

Large-sample Theory

Earlier results generalize to the multivariate case

The vector of MLEs is asymptotically normal. That is, multivariate normal.



The Multivariate Normal

The multivariate normal distribution has many nice features. For us, the important ones are:

- It is characterized by a $k \times 1$ vector of expected values and a $k \times k$ variance-covariance matrix.
- Write $\mathbf{y} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- $\boldsymbol{\Sigma} = [\sigma_{i,j}]$ is a symmetric matrix with variances on the main diagonal and covariances on the off-diagonals.
- All the marginals are normal. $y_j \sim N(\mu_j, \sigma_{j,j})$.

The vector of MLEs is asymptotically multivariate normal. (Thank you, Mr. Wald)

$$\widehat{\boldsymbol{\theta}}_n \stackrel{\cdot}{\sim} N_k\left(\boldsymbol{\theta}, \frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}\right)$$

- Compare $\widehat{\theta}_n \stackrel{\cdot}{\sim} N(\theta, \frac{1}{n\,I(\theta)})$.
- $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ is the Fisher information matrix.
- Specifically, the Fisher information in one observation.
- A $k \times k$ matrix

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \left[-E\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j}\log f(Y;\boldsymbol{\theta})\right)\right]$$

- The Fisher Information in the whole sample is $n\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$.

$\widehat{\boldsymbol{\theta}}_n$ is asymptotically $N_k\left(\boldsymbol{\theta}, \frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}\right)$

- Asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}_n$ is $\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}$, and of course we don't know $\boldsymbol{\theta}$.
- For tests and confidence intervals, we need a good *approximate* asymptotic covariance matrix,
- Based on a good estimate of the Fisher information matrix.
- $\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_n)$ would do.
- But it's inconvenient: Need to compute partial derivatives and expected values in

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \left[E[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y;\boldsymbol{\theta})]\right]$$

and then substitute $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$.

The observed Fisher information

Approximate

$$\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1} = \left[n\, E[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y;\boldsymbol{\theta})]\right]^{-1}$$

with

$$\widehat{\mathbf{V}}_n = \left(\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\boldsymbol{\theta},\mathbf{Y})\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}\right)^{-1}$$

As in the univariate case, substitute the MLE for the parameter instead of taking the expected value.

Compare the Hessian and (Estimated) Asymptotic Covariance Matrix

- $\widehat{\mathbf{V}}_n = \left(\left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\boldsymbol{\theta}, \mathbf{Y}) \right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \right)^{-1}$
- Hessian at MLE is $\mathbf{H} = \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\boldsymbol{\theta}, \mathbf{Y}) \right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$
- So to estimate the asymptotic covariance matrix of $\boldsymbol{\theta}$, just invert the Hessian.
- The Hessian is usually available as a by-product of a numerical search for the MLE.
- Because it's needed for the second derivative test.

Connection to Numerical Optimization

- Suppose we are minimizing the minus log likelihood by a direct search.
- We have reached a point where the gradient is close to zero. Is this point a minimum?
- The Hessian is a matrix of mixed partial derivatives. If all its eigenvalues are positive at a point, the function is concave up there.
- Partial derivatives are usually approximated by the slopes of secant lines – no need to calculate them symbolically.
- It's *the* multivariable second derivative test.

So to find the estimated asymptotic covariance matrix

- Minimize the minus log likelihood numerically.
- The Hessian at the place where the search stops is usually available.
- Invert it to get $\widehat{\mathbf{V}}_n$.
- This is so handy that sometimes we do it even when a closed-form expression for the MLE is available.

Estimated Asymptotic Covariance Matrix $\widehat{\mathbf{V}}_n$ is Useful

- Asymptotic standard error of $\widehat{\theta}_j$ is the square root of the jth diagonal element.
- Denote the asymptotic standard error of $\widehat{\theta}_j$ by $S_{\widehat{\theta}_j}$.
- Thus

$$Z_j = \frac{\widehat{\theta}_j - \theta_j}{S_{\widehat{\theta}_j}}$$

  is approximately standard normal.

Confidence Intervals and Z-tests

Have $Z_j = \frac{\widehat{\theta}_j - \theta_j}{S_{\widehat{\theta}_j}}$ approximately standard normal, yielding

- Confidence intervals: $\widehat{\theta}_j \pm S_{\widehat{\theta}_j} z_{\alpha/2}$
- Test $H_0 : \theta_j = \theta_0$ using

$$Z = \frac{\widehat{\theta}_j - \theta_0}{S_{\widehat{\theta}_j}}$$

Some null hypotheses involve multiple parameters

For example,

$$H_0 : \qquad \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_0 : \qquad \frac{1}{3}(\theta_1 + \theta_2 + \theta_3) = \frac{1}{3}(\theta_4 + \theta_5 + \theta_6) = \frac{1}{2}(\theta_7 + \theta_8)$$

Two hypothesis tests for multi-parameter problems

They also apply to single-parameter models

- Wald tests and likelihood ratio tests.
- They both apply to linear null hypotheses of the form $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$
- Where $\mathbf{L}$ is an r by k matrix with linearly independent rows.
- This kind of null hypothesis is familiar from linear regression (STA302).

Example

Linear regression with 4 explanatory variables

- $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \sigma^2)$
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{0}$

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Another example of $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$

A collection of linear constraints on the parameter $\boldsymbol{\theta}$

Example with $k = 7$ parameters: H_0 has three parts

- $\theta_1 = \theta_2$ and
- $\theta_6 = \theta_7$ and
- $\frac{1}{3}(\theta_1 + \theta_2 + \theta_3) = \frac{1}{3}(\theta_4 + \theta_5 + \theta_6)$

$$
\left(\begin{array}{rrrrrrr}
1 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 \\
1 & 1 & 1 & -1 & -1 & -1 & 0
\end{array} \right)
\left(\begin{array}{c}
\theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7
\end{array} \right)
=
\left(\begin{array}{c}
0 \\ 0 \\ 0
\end{array} \right)
$$

Notice the number of rows in $\mathbf{L}$ is the number of $=$ signs in H_0.

Wald Test for $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$

Based on $(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2(p)$

$$W_n = (\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})^\top \left(\mathbf{L}\widehat{\mathbf{V}}_n\mathbf{L}^\top\right)^{-1} (\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})$$

- Looks like the formula for the general linear F-test in regression.
- Asymptotically chi-squared under H_0.
- Reject for large values of W_n.
- df = number of rows in $\mathbf{L}$.
- Number of linear constraints specified by H_0.

The Wtest Function

$W_n = (\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})^\top \left(\mathbf{L}\widehat{\mathbf{V}}_n\mathbf{L}^\top\right)^{-1} (\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})$

```
Wtest = function(L,Tn,Vn,h=0) # H0: L theta = h
# For Wald tests based on numerical MLEs, Tn = theta-hat,
# and Vn is the inverse of the Hessian.
     {
     value = numeric(3)
     names(value) = c("W","df","p-value")
     r = dim(L)[1]
     W = t(L%*%Tn-h) %*% solve(L%*%Vn%*%t(L)) %*%
          (L%*%Tn-h)
     W = as.numeric(W)
     pval = 1-pchisq(W,r)
     value[1] = W; value[2] = r; value[3] = pval
     return(value)
     }
```

Likelihood ratio tests

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_\theta,\ \theta \in \Theta$
- $H_0 : \theta \in \Theta_0$ v.s. $H_1 : \theta \in \Theta \cap \Theta_0^c$

$$
\begin{aligned}
G^2 &= -2 \log \left(\frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} \right) = -2 \log \frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})} \\
&= 2 \left(\ell(\widehat{\theta}) - \ell(\widehat{\theta}_0) \right)
\end{aligned}
$$

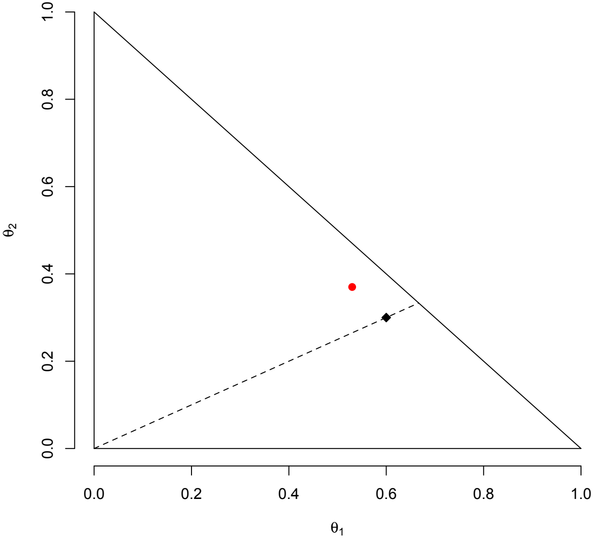
- Under H_0, G^2 has an approximate chi-squared distribution for large n.
- Degrees of freedom = number of (non-redundant, linear) equalities specified by H_0.
- Reject when G^2 is large.

Example: Multinomial with 3 categories

- Parameter space is 2-dimensional.
- Unrestricted MLE is (p_1, p_2): Sample proportions.
- $H_0 : \theta_1 = 2\theta_2$

Parameter space for $H_0 : \theta_1 = 2\theta_2$

Red dot is unrestricted MLE, Black square is restricted MLE



Comparing Likelihood Ratio and Wald tests

- Asymptotically equivalent under H_0, meaning $(W_n - G_n^2) \stackrel{p}{\rightarrow} 0$
- Under H_1,
  - Both have the same approximate distribution (non-central chi-square).
  - Both go to infinity as $n \rightarrow \infty$.
  - But values are not necessarily close for the same data set.
- Likelihood ratio test tends to get closer to the right Type I error probability for small samples.
- Wald can be more convenient when testing lots of hypotheses, because you only need to fit the model once.
- Wald can be more convenient if it's a lot of work to write the restricted likelihood.

Non-linear functions of the parameter vector

- Most tests are about linear combinations of the model parameters.
- Sometimes we want tests and confidence intervals for *non-linear* functions of $\boldsymbol{\theta} \in \mathbb{R}^k$.
- Like $\frac{\alpha}{\lambda^2}$ (variance of a gamma).
- Fortunately, smooth functions of an asymptotically multivariate normal random vector are asymptotically normal.

Theorem based on the delta method of Cramér

The delta method is more general than this.

Let $\boldsymbol{\theta} \in \mathbb{R}^k$. Under the conditions for which $\widehat{\boldsymbol{\theta}}_n$ is asymptotically $N_k\left(\boldsymbol{\theta}, \mathbf{V}_n\right)$ with $\mathbf{V}_n = \frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}$, let the function $g : \mathbb{R}^k \to \mathbb{R}$ be such that the elements of $\dot{\mathrm{g}}(\boldsymbol{\theta}) = \left(\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_k}\right)$ are continuous in a neighbourhood of the true parameter vector $\boldsymbol{\theta}$. Then

$$g(\widehat{\boldsymbol{\theta}}) \mathrel{\dot\sim} N\left(g(\boldsymbol{\theta}), \dot{\mathrm{g}}(\boldsymbol{\theta})\mathbf{V}_n\, \dot{\mathrm{g}}(\boldsymbol{\theta})^\top\right).$$

Note that the asymptotic variance $\dot{\mathrm{g}}(\boldsymbol{\theta})\mathbf{V}_n\, \dot{\mathrm{g}}(\boldsymbol{\theta})^\top$ is a matrix product: $(1 \times k)$ times $(k \times k)$ times $(k \times 1)$.

The standard error of $g(\widehat{\boldsymbol{\theta}})$ is $\sqrt{\dot{\mathrm{g}}(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{V}}_n\, \dot{\mathrm{g}}(\widehat{\boldsymbol{\theta}})^\top}$.

Example of $\dot{\mathrm{g}}(\boldsymbol{\theta}) = \left(\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_k} \right)$

- Variance of gamma is $g(\alpha, \lambda) = \frac{\alpha}{\lambda^2}$.
- $\theta_1 = \alpha$, $\theta_2 = \lambda$, $k = 2$,
- So $\dot{\mathrm{g}}(\boldsymbol{\theta})$ is 1×2.

$$
\begin{aligned}
\dot{\mathrm{g}} &= \left(\frac{\partial g}{\partial \alpha}, \frac{\partial g}{\partial \lambda} \right) \\
&= \left(\frac{1}{\lambda^2}, \; \alpha(-2)\lambda^{-3} \right) \\
&= \left(\frac{1}{\lambda^2}, \; \frac{-2\alpha}{\lambda^3} \right)
\end{aligned}
$$

Then, $\dot{\mathrm{g}}(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{V}}_n \, \dot{\mathrm{g}}(\widehat{\boldsymbol{\theta}})^\top$ is easy if you have $\widehat{\mathbf{V}}_n$.

Copyright Information