Approximate Normality, Newton-Raphson, & Multivariate Delta Method

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 740: Statistical Computing

Statistical models...

...come in all shapes and sizes!

- STAT 704/705: linear models with normal errors, logistic & Poisson regression
- STAT 520/720: time series models, e.g. ARIMA
- STAT 771: longitudinal models with fixed and random effects
- STAT 770: logistic regression, generalized linear mixed models, $r \times k$ tables
- STAT 530/730: multivariate models
- STAT 521/721: stochastic processes
- et cetera, et cetera, et cetera...

Examples from your textbook...

- Censored data (Ex. 1.1, p. 2)

$$Z_i = \min\{X_i, \omega\},\ X_1, \dots, X_n \overset{iid}{\sim} f(x|\alpha, \beta) = \alpha\beta x^{\alpha-1}\exp(-\beta x^{\alpha}).$$

- Finite mixture (Ex. 1.2 & 1.10, p. 3 & 11)

$$X_1, \dots, X_n \overset{iid}{\sim} \sum_{j=1}^{k} p_j N(\mu_j, \sigma_j^2).$$

- Beta data

$$X_1, \dots, X_n \overset{iid}{\sim} f(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

- Logistic regression (Ex. 1.13, p. 15)

$$Y_i \overset{ind.}{\sim} \text{Bern}\left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right).$$

- More: MA(q), normal, gamma, student t, et cetera...

Likelihood

- The likelihood is simply the joint distribution of data, viewed as a function of model parameters.
- $L(\boldsymbol{\theta}|\mathbf{x}) = L(\theta_1, \ldots, \theta_k | x_1, \ldots, x_n) = f(x_1, \ldots, x_n | \theta_1, \ldots, \theta_k)$.
- If data are independent then $L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} f_i(x_i|\boldsymbol{\theta})$ where $f_i(\cdot|\boldsymbol{\theta})$ is the marginal pdf/pmf of X_i.
- Beta data

$$L(\alpha, \beta|\mathbf{x}) = \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1}.$$

- Logistic regression data

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right)^{y_i} \left(\frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right)^{1-y_i}.$$

Two main inferential approaches...

- Section 1.2: Maximum likelihood. Use likelihood “as is” for information on $\boldsymbol{\theta}$.
- Section 1.3: Bayesian. Include additional information on $\boldsymbol{\theta}$ through prior.
- Methods of moments (MOM) and generalized method of moments (GMOM) are simple, direct methods for estimating model parameters that match population moments to sample moments. Sometimes easier than MLE, e.g. beta data, gamma data.
- Your text introduces the Bayesian approach in Chapter 1; we will first consider large-sample approximations.

First derivative vector and second derivative matrix

The gradient vector (or "score vector") of the log-likelihood are the partial derivatives:

$$\nabla \log L(\boldsymbol{\theta}|\mathbf{x}) = \begin{bmatrix} \frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k} \end{bmatrix},$$

and the Hessian, i.e. matrix of second partial derivatives is denoted:

$$\nabla^2 \log L(\boldsymbol{\theta}|\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1^2} & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k^2} \end{bmatrix}.$$

Maximum likelihood & asymptotic normality

- $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}|\mathbf{x})$.
- Finds $\boldsymbol{\theta}$ that makes data $\mathbf{x}$ as "likely" as possible.
- Under regularity conditions,

$$\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N_k(\boldsymbol{\theta}, \mathbf{V}_{\boldsymbol{\theta}}),\ \mathbf{V}_{\boldsymbol{\theta}} = [-E\{\nabla^2 \log L(\boldsymbol{\theta}|\mathbf{X})\}]^{-1}.$$

- Normal approximation historically extremely important; crux of countless approaches and papers.
- Regularity conditions have to do with differentiability of $L(\boldsymbol{\theta}|\mathbf{x})$ w.r.t. $\boldsymbol{\theta}$, support of x_i not depending on $\boldsymbol{\theta}$, invertibility of Fisher information matrix, and certain expectations are finite. Most models we'll consider satisfy the necessary conditions.

How to maximize likelihood?

- Problem is to maximize $L(\boldsymbol{\theta}|\mathbf{x})$, or equivalently $\log L(\boldsymbol{\theta}|\mathbf{x})$.
- Any extrema $\hat{\boldsymbol{\theta}}$ satisfies $\nabla \log L(\boldsymbol{\theta}|\mathbf{x}) = \mathbf{0}$.
- In one dimension this is $\frac{\partial}{\partial\theta} L(\theta|\mathbf{x}) = 0$.
- Global max if $\frac{\partial^2}{\partial\theta^2} \log L(\theta|\mathbf{x}) < 0$ for $\theta \in \Theta$.
- Only in simple problems can we solve this directly.
- Various iterative approaches: steepest descent, conjugate gradient, Newton-Raphson, etc.
- Iterative approaches work fine for moderate number of parameters k and/or sample sizes n, but can break down when problem gets "too big."

Classic examples with closed-form MLEs

- $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bern}(\pi)$.
- $Y = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, \pi)$.
- $X_i \overset{ind.}{\sim} \text{Pois}(\lambda t_i)$, t_i is an "offset" or "exposure time."
- $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.
- $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

About your textbook...

- Main focus: simulation-based approaches to obtaining inference. More broadly useful than traditional deterministic methods, especially for large data sets and/or complicated models.
- We will discuss non-simulation based approaches as well in a bit more detail than your book; still used *a lot* and quite important. See Gentle (2009) for complete treatment.
- Your text has lots of technical points, pathological examples, convergence theory, alternative methods (e.g. profile likelihood), etc. We may briefly touch on these but omit details. Please read your book!
- Main focus of course: gain experience and develop intuition on numerical methods for obtaining inference, i.e. build your "toolbox."

Deterministic versus stochastic numerical methods

- Deterministic methods use a prescribed algorithm to arrive at a solution. Given the same starting values the solution they arrive at the same answer each time.
- Stochastic methods introduce randomness into an algorithm. Different "runs" typically produce (slightly) different solutions.
- We will first examine a deterministic approach to maximizing the likelihood and obtain inference based on approximate normality.

First derivative vector and second derivative matrix

The gradient vector (or "score vector") of the log-likelihood are the partial derivatives:

$$\nabla \log L(\boldsymbol{\theta}|\mathbf{x}) = \begin{bmatrix} \frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k} \end{bmatrix},$$

and the Hessian, i.e. matrix of second partial derivatives is denoted:

$$\nabla^2 \log L(\boldsymbol{\theta}|\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1^2} & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k^2} \end{bmatrix}.$$

Maximum likelihood...

- The *likelihood* or *score* equations are

  $$\nabla \log L(\boldsymbol{\theta}|\mathbf{x}) = \mathbf{0}.$$

  The MLE $\hat{\boldsymbol{\theta}}$ satisfies these. However, a solution to these equations may not be the MLE.
- $\hat{\boldsymbol{\theta}}$ is a local max if $\nabla \log L(\boldsymbol{\theta}|\mathbf{x}) = \mathbf{0}$ and $\nabla^2 \log L(\boldsymbol{\theta}|\mathbf{x})$ is negative-semidefinite (has non-positive eigvenvalues).
- The expected Fisher information is

  $$\mathbf{I}_n(\boldsymbol{\theta}) = E\{[\nabla \log L(\boldsymbol{\theta}|\mathbf{X})][\nabla \log L(\boldsymbol{\theta}|\mathbf{X})]'\} = -E\{\nabla^2 \log L(\boldsymbol{\theta}|\mathbf{x})\}.$$

- The observed Fisher information is this quantity without the expectation

  $$\mathbf{J}_n(\boldsymbol{\theta}) = -\nabla^2 \log L(\boldsymbol{\theta}|\mathbf{x}).$$

Maximum likelihood...

- In large samples under regularity conditions,

$$\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N_k(\boldsymbol{\theta}, \mathbf{I}_n(\boldsymbol{\theta})^{-1}) \text{ and } \hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N_k(\boldsymbol{\theta}, \mathbf{J}_n(\boldsymbol{\theta})^{-1}).$$

- Just replace the unknown $\boldsymbol{\theta}$ by the MLE $\hat{\boldsymbol{\theta}}$ in applications and the approximations still hold.

$$\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N_k(\boldsymbol{\theta}, \mathbf{I}_n(\hat{\boldsymbol{\theta}})^{-1}) \text{ and } \hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N_k(\boldsymbol{\theta}, \mathbf{J}_n(\hat{\boldsymbol{\theta}})^{-1}).$$

- The observed Fisher information $\mathbf{J}_n(\hat{\boldsymbol{\theta}})$ is easier to compute and recommended by Efron and Hinkley (1978) when using the normal approximation above.

Multivariate delta method

Once the MLE $\hat{\boldsymbol{\theta}}$ and its approximate covariance $\mathbf{J}_n(\hat{\boldsymbol{\theta}})$ is obtained, we may be interested in functions of $\boldsymbol{\theta}$, e.g. the vector

$$\mathbf{g}(\boldsymbol{\theta}) = \begin{bmatrix} g_1(\boldsymbol{\theta}) \\ \vdots \\ g_m(\boldsymbol{\theta}) \end{bmatrix}.$$

The multivariate delta method gives

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) \overset{\bullet}{\sim} N_m(\mathbf{g}(\boldsymbol{\theta}), [\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]' \mathbf{J}(\hat{\boldsymbol{\theta}})^{-1} [\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]).$$

This stems from approximate normality and the first-order Taylors approximation

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) \approx \mathbf{g}(\boldsymbol{\theta}) + [\nabla \mathbf{g}(\boldsymbol{\theta})]'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

Multivariate delta method

As before,

$$\nabla \mathbf{g}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial g_m(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_2} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial g_m(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_k} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_k} & \cdots & \frac{\partial g_m(\boldsymbol{\theta})}{\partial \theta_k} \end{bmatrix}.$$

For univariate functions $g(\boldsymbol{\theta}) : \mathbb{R}^k \to \mathbb{R}$

$$\nabla g(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_k} \end{bmatrix}.$$

Example: $g(\mu, \sigma) = \mu/\sigma$, the signal-to-noise ratio.

Multivariate delta method: testing

We have

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) \overset{\bullet}{\sim} N_m(\mathbf{g}(\boldsymbol{\theta}), [\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]' \mathbf{J}(\hat{\boldsymbol{\theta}})^{-1} [\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]).$$

Let $\mathbf{g}_0 \in \mathbb{R}^m$ be a fixed, known vector.

An approximate Wald test of $H_0 : \mathbf{g}(\boldsymbol{\theta}) = \mathbf{g}_0$ has test statistic

$$W = (\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}_0)' \{[\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]' \mathbf{J}(\hat{\boldsymbol{\theta}})^{-1} [\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]\}^{-1} (\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}_0).$$

In large samples

$$W \overset{\bullet}{\sim} \chi^2_m.$$

Simple confidence intervals and tests

- Along the diagonal of $\mathbf{J}(\hat{\boldsymbol{\theta}})^{-1}$ are the estimated variances of $\hat{\theta}_1, \ldots, \hat{\theta}_k$. Approximate 95% confidence interval for θ_j is $\hat{\theta}_j \pm 1.96\sqrt{[\mathbf{J}(\hat{\boldsymbol{\theta}})^{-1}]_{jj}}$.
- Can test $H_0 : \theta_j = t$ from test statistic $z = \frac{\hat{\theta}_j - t}{\sqrt{[\mathbf{J}(\hat{\boldsymbol{\theta}})^{-1}]_{jj}}}$.
- For any function $g(\boldsymbol{\theta})$ can obtain CI and test in similar manner using result on previous slide. For example, 95% CI is $g(\hat{\boldsymbol{\theta}}) \pm 1.96\sqrt{[\nabla g(\hat{\boldsymbol{\theta}})]' \mathbf{J}(\hat{\boldsymbol{\theta}})^{-1} [\nabla g(\hat{\boldsymbol{\theta}})]}$.

Univariate Newton-Raphson

In general, computing the MLE, i.e. the solution to $\nabla \log L(\boldsymbol{\theta}|\mathbf{x}) = \mathbf{0}$, is non-trivial except in very simple models. An iterative method for finding the MLE is Newton-Raphson.

Want to find a zero of some univariate function $g(\cdot)$, i.e. an x s.t. $g(x) = 0$. A first-order Taylors approximation gives $g(x_1) \approx g(x_0) + g'(x_0)(x_1 - x_0)$. Setting this to zero and solving gives $x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}$. The iterative version is

$$x_{j+1} = x_j - \frac{g(x_j)}{g'(x_j)}.$$

I will show on the board how this works in practice. For univariate θ, solve $\frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x}) = 0$ via $\theta_{j+1} = \theta_j - \frac{\frac{\partial}{\partial\theta} \log L(\theta_j|\mathbf{x})}{\frac{\partial^2}{\partial\theta^2} \log L(\theta_j|\mathbf{x})}$.

R example

Bernoulli data

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{Bern}(\pi),$$

which is the same as

$$Y = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, \pi).$$

- MLE $\hat{\pi}$ directly...
- MLE via Newton-Raphson...
- Starting values?
- Multiple local maxima? Example 1.9 (p. 10); Example 1.17 (p.19).

Multivariate Newton-Raphson

The multivariate version applied to $\nabla \log L(\boldsymbol{\theta}|\mathbf{x}) = \mathbf{0}$ is

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - [\nabla^2 \log L(\boldsymbol{\theta}_j|\mathbf{x})]^{-1}[\nabla \log L(\boldsymbol{\theta}_j|\mathbf{x})].$$

Taking inverses in a terribly inefficient way to solve a linear system of equations and instead one solves

$$[\nabla^2 \log L(\boldsymbol{\theta}_j|\mathbf{x})]\boldsymbol{\theta}_{j+1} = [\nabla^2 \log L(\boldsymbol{\theta}_j|\mathbf{x})]\boldsymbol{\theta}_j - [\nabla \log L(\boldsymbol{\theta}_j|\mathbf{x})]$$

for $\boldsymbol{\theta}_{j+1}$ through either a direct decomposition (Cholesky) or iterative (conjugate gradient) method.

Stop when $||\hat{\boldsymbol{\theta}}_{j+1} - \hat{\boldsymbol{\theta}}_j|| < \epsilon$ for some norm $||\cdot||$ and tolerance ϵ. $||\boldsymbol{\theta}||_1 = \sum_{i=1}^k |\theta_i|$, $||\boldsymbol{\theta}||_2 = \sqrt{\sum_{i=1}^k \theta_i^2}$, and $||\boldsymbol{\theta}||_\infty = \max_{i=1,\ldots,k} |\theta_i|$. This is absolute criterion; can also use relative.

Notes on Newton-Raphson

- Computing the Hessian can be computationally demanding; finite difference approximations can reduce this burden.
- Steepest descent only uses the gradient $\nabla \log L(\boldsymbol{\theta}|\mathbf{x})$, ignoring how the gradient is changing (the Hessian). May be easier to program.
- Quasi-Newton methods use approximations that satisfy "the secant condition"; an approximation to the Hessian is built up as the algorithm progresses using low-rank updates. Popular version is Broyden-Fletcher-Goldfarb-Shanno (BFGS). Details beyond scope of this course (see Sec. 6.2 in Gentle, 2009).

Useful R functions

- optim in R is powerful optimization function that performs Nelder-Mead (only function values $\log L(\boldsymbol{\theta}|\mathbf{x})$ are used!), conjugate gradient, BFGS (including constrained), and simulated annealing. The ucminf package also has the stable function ucminf for optimization and uses the same syntax as optim. The Hessian at the maximum is provided for most approaches.
- Maple, Mathematica (http://www.wolframalpha.com/calculators/derivative-calculator/) can symbolically differentiate functions, so can R! deriv is built-in and Deriv is in the Deriv package.
- jacobian and hessian are numerical estimates from the numDeriv package. The first gives the gradient, the second the matrix of second partial derivatives, both evaluated at a vector.

Censored data likelihood

Consider right censored data. Event times are

$$T_1, \ldots, T_n \overset{iid}{\sim} f(t),$$

independent of censoring times

$$C_1, \ldots, C_n \overset{iid}{\sim} g(t).$$

We observe $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I\{T_i \leq C_i\}$: $\delta_i = 1$ if the ith observation is uncensored.

Censored data likelihood

This is not obvious, but the the joint distribution of $\{(Y_i, \delta_i)\}_{i=1}^n$ is

$$\prod_{i=1}^n \{f(y_i)[1 - G(y_i)]\}^{\delta_i} \{g(y_i)[1 - F(y_i)]\}^{1-\delta_i}.$$

If we have time later we will derive this formally.

For now, note that if $f(\cdot)$ and $g(\cdot)$ do not share any parameters, and $f(\cdot)$ has parameters $\boldsymbol{\theta}$, then

$$L(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\delta}) \propto \prod_{i=1}^n f(y_i|\boldsymbol{\theta})^{\delta_i}[1 - F(y_i|\boldsymbol{\theta})]^{1-\delta_i}.$$

Multivariate Newton-Raphson & delta method examples

- Censored gamma data (p. 24)...
- Logistic regression data (p. 15)...

Note: Tim has used `optim` many, many times fitting models and/or obtaining starting values and initial covariance matrices for random-walk Markov chain Monte Carlo (later).

Usually not necessary to program your own iterative optimization method, but it might come up...

Crude starting values for logistic regression...

Let's find first-order Taylor's approximation to logistic function. Let $g(x) = \frac{e^x}{1+e^x}$; then $g'(x) = \frac{e^x}{(1+e^x)^2}$ and

$$g(x) \approx g(0) + g'(0)(x-0) \Leftrightarrow \frac{e^x}{1+e^x} \approx \tfrac{1}{2} + \tfrac{1}{4}x.$$

In other words,

$$\frac{e^{4\mathbf{x}_i'\boldsymbol{\theta}-\frac{1}{2}}}{1+e^{4\mathbf{x}_i'\boldsymbol{\theta}-\frac{1}{2}}} \approx \mathbf{x}_i'\boldsymbol{\theta}.$$

Therefore can fit usual l.s. to get $\hat{E}(Y_i) = \hat{\theta}_1 + \hat{\theta}_2 t_i$ and take initial guess as $\theta_1 = 4\hat{\theta}_1 - \frac{1}{2}$, $\theta_2 = 4\hat{\theta}_2$.

Taylors theorem very useful!

The Bayesian approach...

1. ...considers $\boldsymbol{\theta}$ to be random, and
2. augments the model with additional information about $\boldsymbol{\theta}$ called the "prior" $\pi(\boldsymbol{\theta})$.

Bayes' rule gives the posterior distribution

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(x_1, \ldots, x_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(x_1, \ldots, x_n)}.$$

Here,

$$f(x_1, \ldots, x_n) = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \underbrace{f(x_1, \ldots, x_n|\boldsymbol{\theta})}_{L(\boldsymbol{\theta}|\mathbf{x})} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

is the marginal density of $\mathbf{x}$ integrating over the prior through the model.

Normalizing constant of posterior density $f(\mathbf{x})$

$f(\mathbf{x}) = f(x_1, \ldots, x_n)$ is often practically impossible to compute, but thankfully most information for $\boldsymbol{\theta}$ is carried by *the shape* of the density w.r.t. $\boldsymbol{\theta}$

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \underbrace{f(x_1, \ldots, x_n|\boldsymbol{\theta})}_{f(\mathbf{x}|\boldsymbol{\theta})=L(\boldsymbol{\theta}|\mathbf{x})} \pi(\boldsymbol{\theta}).$$

$f(\mathbf{x})$ is an ingredient in a Bayes' factor for comparing two models.

Instead of using the likelihood for inference, we use the posterior density, which is proportional to the likelihood times the prior $\pi(\boldsymbol{\theta}|\mathbf{x}) \propto L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})$. Note that if $\pi(\boldsymbol{\theta})$ is constant then we obtain the likelihood back!

There is more *similar* than *different* between Bayesian and likelihood-based inference. Both specify a probability model and use $L(\boldsymbol{\theta}|\mathbf{x})$. Bayes' justs adds $\pi(\boldsymbol{\theta})$.

Posterior inference

A common estimate of $\boldsymbol{\theta}$ is the posterior mean

$$\tilde{\boldsymbol{\theta}} = \int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \boldsymbol{\theta} \pi(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta}.$$

This is the Bayes estimate w.r.t. squared error loss (p. 13). For $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ the estimate of θ_j w.r.t. to absolute loss is the posterior median $\tilde{\theta}_j$, satisfying

$$\int_{-\infty}^{\tilde{\theta}_j} \pi(\theta_j | \mathbf{x}) d\theta_j = 0.5.$$

A 95% credible interval (L, U) satisfies $P(L \leq \theta_j \leq U | \mathbf{x}) = 0.95$:

$$\int_L^U \pi(\theta_j | \mathbf{x}) d\theta_j = 0.95.$$

Functions of parameters $\boldsymbol{\theta}$

We are often interested in functions of model parameters $g(\boldsymbol{\theta})$, in which case the posterior mean is

$$\widetilde{g(\boldsymbol{\theta})} = \int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} g(\boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta},$$

and the posterior median satisfies...*something difficult to even write down!* Need to derive the distribution of $g(\boldsymbol{\theta})$ from $\pi(\boldsymbol{\theta}|\mathbf{x})$.

Note that the distribution of $g(\boldsymbol{\theta})$ *can be derived* if $\boldsymbol{\theta}|\mathbf{x}$ is multivariate normal or approximately multivariate normal via the multivariate delta method.

Bayesian inference was at a bottleneck for years...

Parameter estimates, as well as parameter intervals, are impossible to compute directly except for very simple problems!

We need ways to approximate integrals and quantiles to move forward!

One broadly useful approach is Markov chain Monte Carlo, which requires simulating random variables from different distributions, depending on the model and approach.

Simulation is also necessary to determine the frequentist operating characteristics (MSE, bias, coverage probability) of estimators under known conditions; many statistical papers have a "simulations" section!

But first we'll consider a Bayesian normal approximation...

Gradient and Hessian of log posterior...

$$\log \pi(\boldsymbol{\theta}|\mathbf{x}) = \log\{L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})\} - \log f(\mathbf{x}).$$

Last term not a function of $\boldsymbol{\theta}$, so...

$$\nabla \log \pi(\boldsymbol{\theta}|\mathbf{x}) = \begin{bmatrix} \frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_k} \end{bmatrix} = \begin{bmatrix} \frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1} + \frac{\partial \log \pi(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k} + \frac{\partial \log \pi(\boldsymbol{\theta})}{\partial \theta_k} \end{bmatrix},$$

and

$$\nabla^2 \log \pi(\boldsymbol{\theta}|\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\partial \theta_k^2} \end{bmatrix}.$$

Gradient and Hessian of log posterior...

$$\nabla \log \pi(\boldsymbol{\theta}|\mathbf{x}) = \nabla \log L(\boldsymbol{\theta}|\mathbf{x}) + \nabla \log \pi(\boldsymbol{\theta}),$$

and

$$\nabla^2 \log \pi(\boldsymbol{\theta}|\mathbf{x}) = \nabla^2 \log L(\boldsymbol{\theta}|\mathbf{x}) + \nabla^2 \log \pi(\boldsymbol{\theta}).$$

Note first part is function of data $\mathbf{x}$ and n, second is not.

Normal approximation to posterior

The normal approximation for MLEs works almost *exactly the same for posterior distributions*!

Let $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \pi(\boldsymbol{\theta}|\mathbf{x})$ be the *posterior mode* (found via Newton-Raphson or by other means). Then

$$\boldsymbol{\theta}|\mathbf{x} \overset{\bullet}{\sim} N_k(\hat{\boldsymbol{\theta}}, [-\nabla^2 \log \pi(\boldsymbol{\theta}|\mathbf{x})]^{-1}).$$

Here, $\hat{\boldsymbol{\theta}}$ is the posterior mode, but also the (approximate) posterior mean and (componentwise) has the posterior medians.

Poisson example

Let $X_i \overset{ind.}{\sim} \text{Pois}(\lambda t_i)$, where t_i is exposure time. Take the prior $\lambda \sim \Gamma(\alpha, \beta)$ where α and β are given. Then

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^{n} \frac{e^{-\lambda t_i}(\lambda t_i)^{x_i}}{x_i!} \propto e^{-\lambda \sum_{i=1}^{n} t_i} \lambda^{\sum_{i=1}^{n} x_i},$$

and

$$\pi(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda},$$

leading to

$$\lambda|\mathbf{x} \sim \Gamma\left(\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} t_i\right).$$

Here, using a *conjugate prior*, we have an exact, closed-form solution. Exact Bayesian solutions are *very, very rare*. However, this allows us to compare the exact posterior density to the normal approximation for real data.

Ache monkey hunting

Garnett McMillan spent a year with the Ache tribe in Paraguay. While living with them he ate the same foods they did including armadillo, honey, tucans, anteaters, snakes, lizards, giant grubs, and capuchin monkeys (including their brains). Garnett says "*Lizard eggs were a real treat*."

Garnett would go on extended hunting trips in the rain forest with hunters; we consider data from $n = 11$ Ache hunters aged 50-59 years followed over several jungle hunting treks. X_i is the number of capuchin monkeys killed over t_i days. We'll assume

$$X_i \overset{ind.}{\sim} \text{Pois}(\lambda t_i), \ i = 1, \ldots, 11.$$

Garnett's prior for an average number of monkeys killed over 10 days is 1, with a 95% interval of 0.25 to 4.

Normal approximation to Challenger data

Recall the logistic regression likelihood

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)} \right)^{y_i} \left(\frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)} \right)^{1-y_i} .$$

Hanson, Branscum, and Johnson (2014) develop the prior

$$\boldsymbol{\beta} \sim N_2 \left(\left[\begin{array}{c} 0 \\ 0 \end{array} \right], \left[\begin{array}{cc} 168.75 & -2.402 \\ -2.402 & 0.03453 \end{array} \right] \right)$$

for the Challenger data.

What's next?

Normal approximations are fairly easy to compute and wildly useful! Newton-Raphson is an approach to finding a MLE or posterior mode; the normal approximation also makes use of the observed Fisher information matrix. The multivariate delta method allows us to obtain inference for functions $g(\boldsymbol{\theta})$.

However, normal approximations break down with large numbers of parameters in $\boldsymbol{\theta}$ and/or complex models. They are also too crude for small sample sizes and do not work for some models. We need an all-purpose, widely applicable approach to obtained posterior inference for Bayesian models; our main tool will be Markov chain Monte Carlo (MCMC).

To understand MCMC we need to first discuss how to generate random variables, then consider Monte Carlo approximations to means and quantiles...these are our next topics.