Maximum Likelihood Estimation

Readings:

Haddon 2011 (Section 3.4)

Acknowledgements: M. Wilberg, R. Latour

Announcements

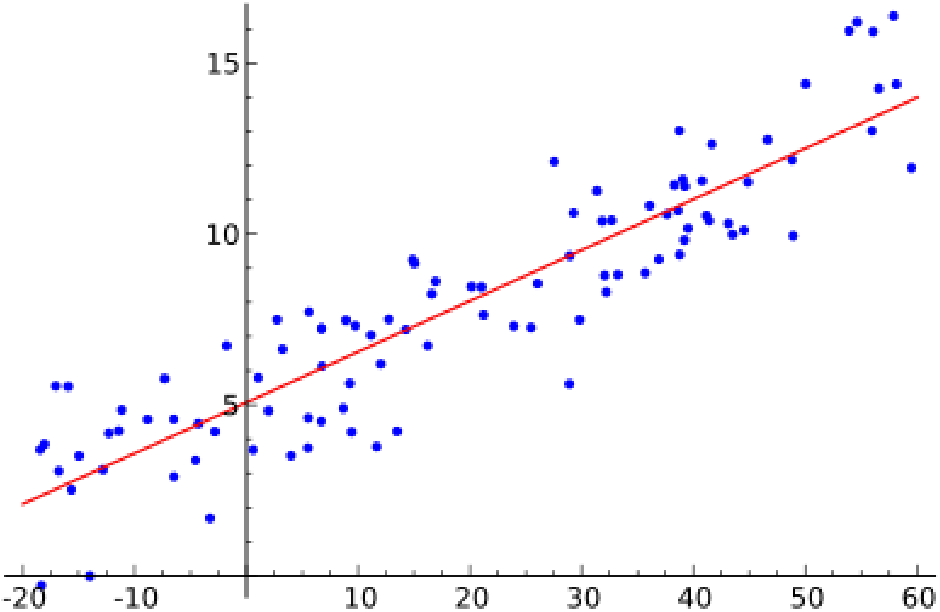
- 558 project synopses → Due Wednesday, 3/18 (11:59pm)

How do we fit models to data?

- Least squares (see refresher below)
- Maximum likelihood
- Bayesian methods (not for this class)

$$residual = \varepsilon_i = Y_i - \widehat{Y}_i$$

$$RSS = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$



Maximum Likelihood

- Alternative way of fitting models and getting estimates…
- Need a quick review of probability

- Simple intro/explanation:
  - https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

Probability

Probability – Frequency View

- Probability is long-run relative frequency (from a random process)
- Same as relative frequency in the population

- Examples
  - Dice toss $p(1) = p(2) = \ldots = p(6) = 1/6$
  - Coin flip $p(\text{Head}) = p(\text{Tail}) = .5$

Probability

- Some characteristics
  - Between 0 and 1
  - Probabilities must be non-negative.
  - The sum of probabilities over all possible mutually exclusive outcomes must equal one (e.g., dice)
  - If two events, A and B, are mutually exclusive, the probability of observing either of the events is the sum of their individual probabilities.

$$P(A \cup B) = P(A) + P(B)$$

"Union"
P(A or B)

Mutually exclusive: related in such a way that each thing makes the other thing impossible; not able to be true at the same time or to exist together

E.g., what is P(1 or 2) when rolling a die?

Independence

- **Joint probability** is the probability that two (or more) different events will occur
- Statistical **independence** means that knowledge of one event provides no information about the probability that another event will occur

$$P(A,B) = P(A)P(B)$$

intersection

P(A,B) is the same as P(A and B) or $\mathrm{P}(A \cap B)$

E.g., what is P(1 and 2) when rolling two die?

Independence

- We usually assume that observations are in some way independent of one another when we fit models

$$P(y_1, y_2, \ldots, y_n) = P(y_1)P(y_2) \ldots P(y_n) = \prod_{i=1}^{n} P(y_i)$$

- E.g., in linear regression, we assume that the errors are independent

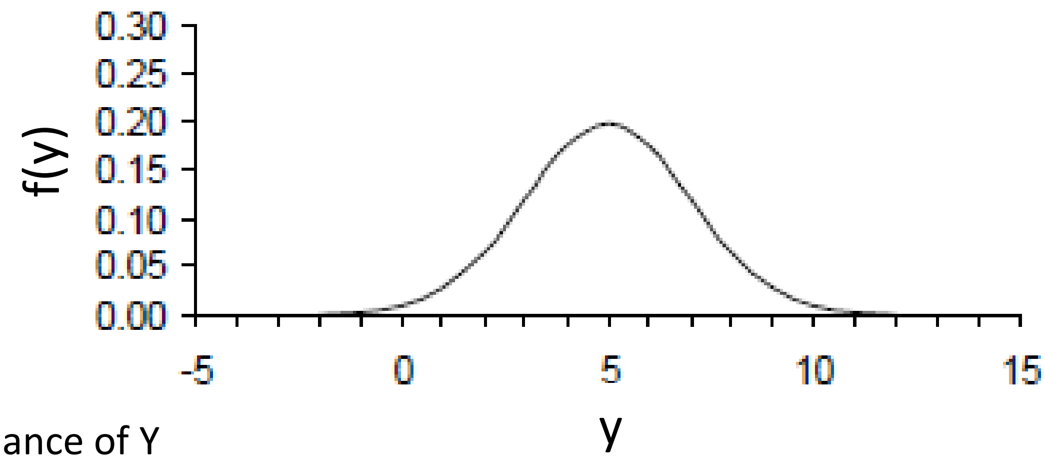
pdf for the normal distribution

- **Probability density function (pdf)** describes the probability of an event occurring for a *continuous* distribution.
  - The total area under the curve = 1
- The normal distribution is one of the most common pdfs:

$$f(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5\left(\frac{y-\mu}{\sigma}\right)^2}$$

$$E(Y) = \mu$$

$$Var(Y) = \sigma^2$$



E(Y) – “expected value of Y”; Var(Y) – variance of Y

Maximum Likelihood

Maximum Likelihood Estimation (MLE)

- ML techniques are a powerful and flexible method for parameter estimation
- Alternative to least squares methods
- **This technique involves finding the parameters that maximize the probability of generating the observed data** (this is how we define "best fit" with ML)
- Likelihood ≠ probability

ML - Developed by R. A. Fischer (published this while a junior in college!)

Maximum Likelihood Parameter Estimation

- ML parameter estimates maximize the probability of observing the data
- Reverses the role of parameters and data (compared to probability)
- **Treat the data as fixed and find parameters that maximize the probability of observing those data**

$$L(parameters \mid data) = P(data \mid parameters)$$

Likelihood (L) is conditioned on the data

Likelihood for continuous distributions

- For a continuous distribution, the likelihood is calculated as:

Probability density function

$$L(\theta \mid Y_i) = \prod_{i=1}^{n} \mathrm{pdf}(Y_i \mid \theta)$$

Parameters

Product

Data

- "the likelihood of the parameter(s) θ (theta) is the product of the pdf values for each of the n observations Y_i given the parameter(s) θ"

Haddon 2011, p. 79

Example 1 – Normal distribution

- To get probability of whole data set (given parameters):
  - Multiply prob. of each data point together b/c independent
  - $P(y_{all}) = P(y_1) x P(y_2) x \ldots x P(y_n)$
- Use normal PDF for getting probabilities.
- Parameters for normal are μ and σ.

Data

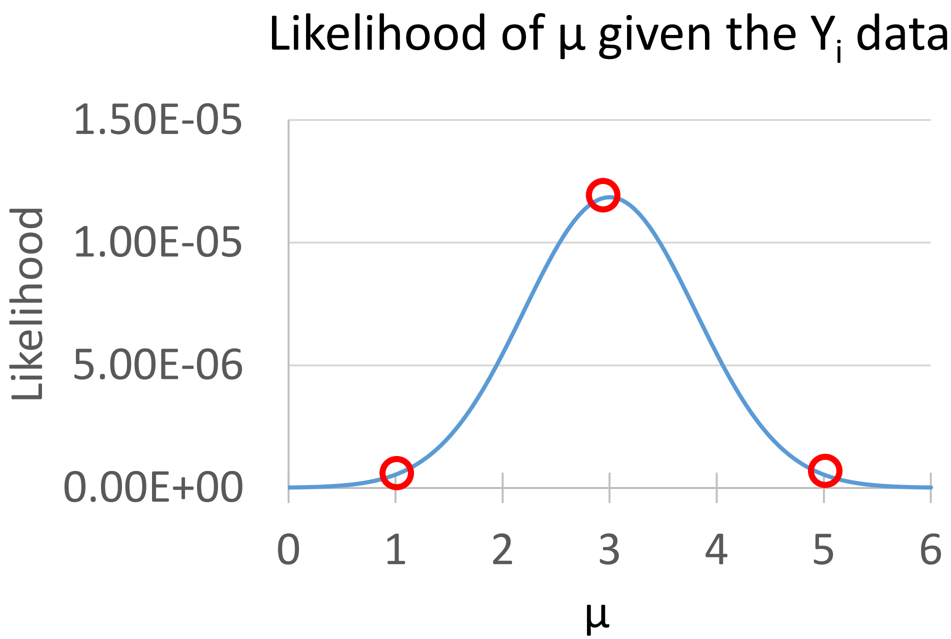
$$L(\mu, \sigma \mid y_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{(y_i - \mu)^2}{\sigma}\right)}$$

Parameters

Maximum Likelihood Example 1

- You take a sample (n=4) from a population and you want to estimate the population average (μ) using ML.
- Samples of Y_i = 0, 2, 4, 6
- Calculate L for different values of μ (here, using sample SD as σ=2.6). Find the value of μ that would "maximize the likelihood".

	$P(Y_i \mid \mu)$		
Y_i	If μ=1	If μ=3	If μ=5
0	0.127	0.027	0.001
2	0.127	0.127	0.027
4	0.027	0.127	0.127
6	0.001	0.027	0.127
L	5.35E-07	1.19E-05	5.35E-07
log(L)	-14.44	-11.34	-14.44
-log(L)	14.44	11.34	14.44



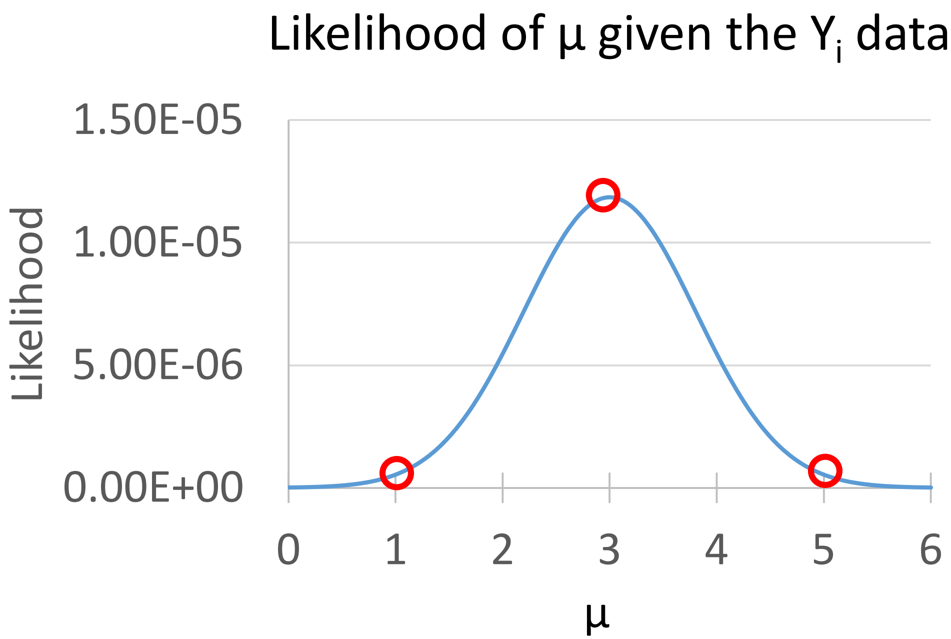
$$L(\mu, \sigma \mid y_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{(y_i - \mu)^2}{\sigma}\right)} = P(Y_1)xP(Y_2)xP(Y_3)xP(Y_4)$$

Maximum Likelihood Example 1

- You take a sample (n=4) from a population and you want to estimate the population average (μ) using ML.
- Samples of Y_i = 0, 2, 4, 6
- Calculate L for different values of μ (here, using sample SD as σ=2.6). Find the value of μ that would "maximize the likelihood".

Our ML estimate of μ is 3.

- This is the value that maximizes the likelihood on the graph
- note this matches our sample mean ($\bar{Y}$ = (0+2+4+6)/4 =3), which is an unbiased estimator of μ.



$$L(\mu,\sigma \mid y_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{(y_i-\mu)^2}{\sigma}\right)}$$ = $P(Y_1)xP(Y_2)xP(Y_3)xP(Y_4)$

Properties of Likelihoods

- Likelihoods do not need to sum to one
  - They are NOT probabilities

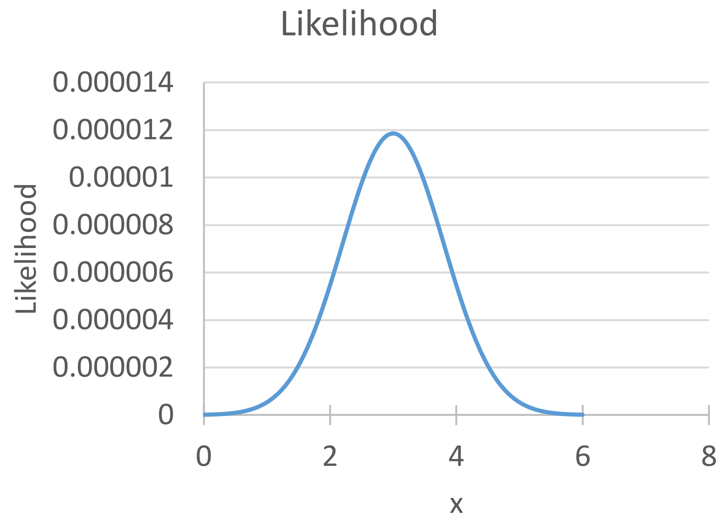
- Likelihoods are a *relative* (not absolute) measure of model fit

- Calculating products is difficult, so logs are often used

Likelihood Nomenclature

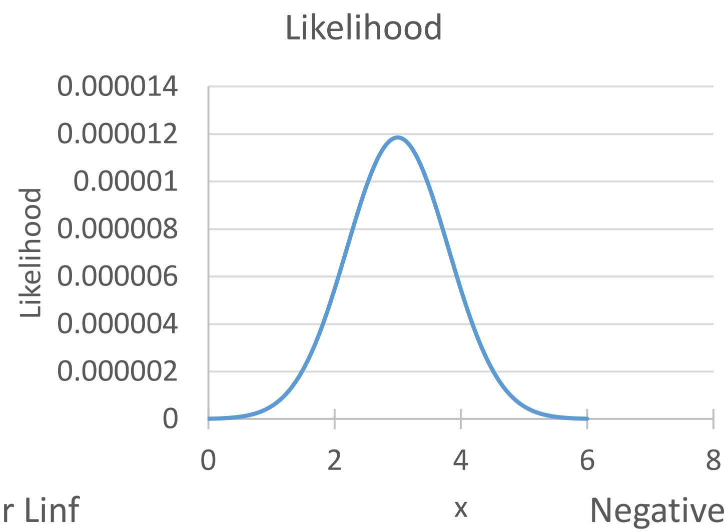
- Terms used to identify likelihood methods can be confusing:
- **Likelihood** $L(\theta|\text{data})$ [where θ are parameters]
  - Remember: goal is to maximize likelihood
- **Log likelihood** $\log(L(\theta|\text{data}))$ or LL
  - Used because easier to deal with sums than with products.
- **Negative log likelihood** $-\log(L(\theta\ |\text{data}))$ or –LL or NLL
  - Used for historical reasons (e.g., folks used to minimizing Sums of Squares), and software more common previously for minimizing functions

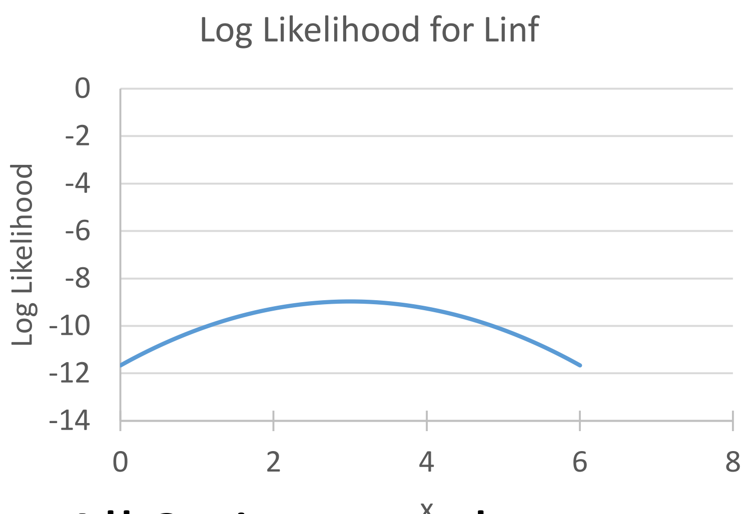
ML Example 1, n=4

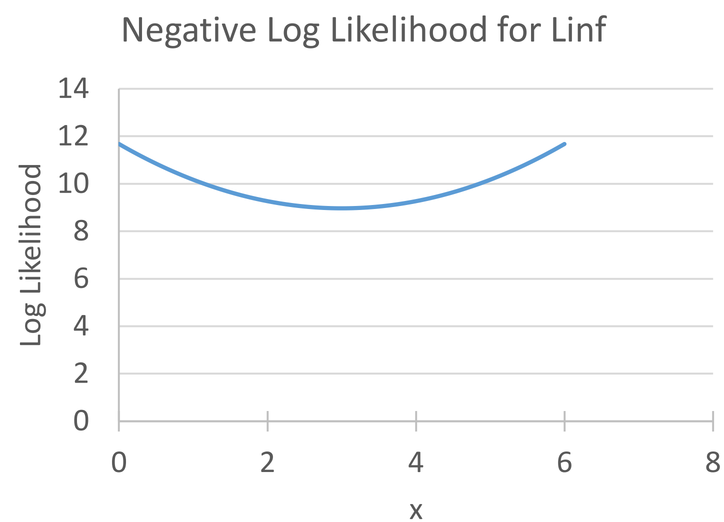


- What would LL and –LL look like?

ML Example 1, n=4

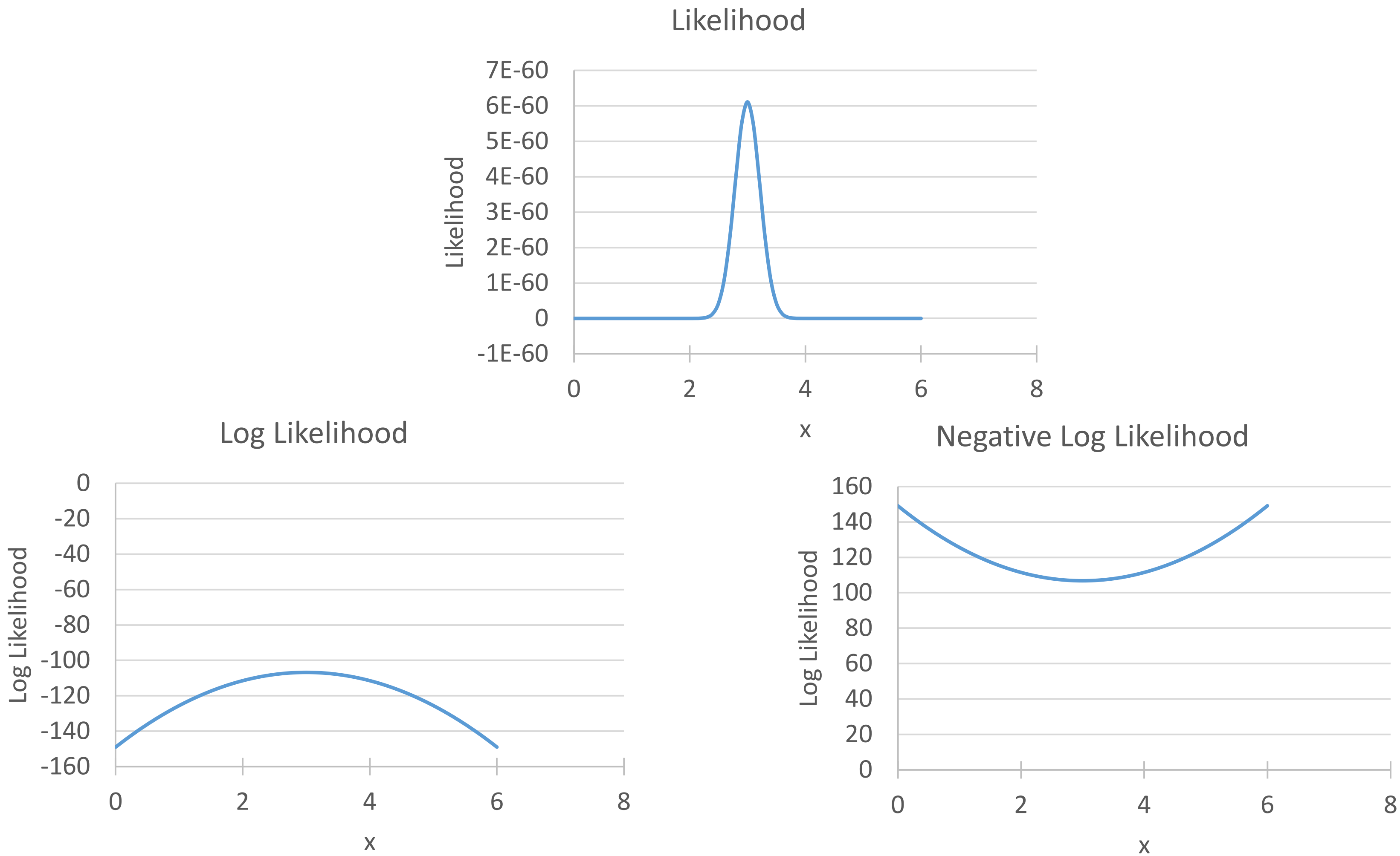






- All 3 give you the same result.
- Profile shapes give you some information of the relative likelihood of different results

ML Example 2, n=48



- As sample size increases, the amount of information increases, so you have a narrower range of values near the "best estimate" → **greater confidence in value**

-LL eqns. – Normal distribution

- Negative log likelihood (negLL, -LL) [to be minimized!]
  - Done to facilitate calculations

$$\boxed{\text{negLL} = \sum_{i=1}^{n}\left(\frac{1}{2}\log(2\pi) + \log(\sigma) + \frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right)}$$

- Negative log likelihood with constants removed:

$$\boxed{\text{negLL} = \sum_{i=1}^{n}\left(\log(\sigma) + \frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right)}$$

μ would be whatever your model is!
e.g., $\mu = \beta_0 + \beta_1 X$

Properties of MLEs

- **Invariant to transformations**
- Asymptotically efficient (**lowest possible variance**)
- Asymptotically **normally distributed**
  - Useful for getting CI intervals
- Asymptotically **unbiased** (expected value of the estimated parameter equals the true value)
- **ML estimates will be the same as least squares estimates if errors are normal, additive, and with constant variance**

"Asymptotic" - deals with the conditions at infinitely large sample sizes.

Potential Challenges of ML Methods

- Programming may be required because usually problem specific
- Likelihood equations need to be worked out for a given problem
- Numerical techniques are often required to find MLE
- MLEs may be biased for small samples and asymptotic benefits may not apply to small samples

Summary 1: Maximum likelihood

- Goal: find the parameter values that make the observed data most likely
  - *Maximize* likelihood (L), *maximize* log-likelihood (LL), or *minimize* negative log-likelihood (-LL) → give same result
- Likelihood is different than probability
  - Probability: Knowing parameters → Prediction of data
  - Likelihood: Observation of data → Estimation of parameters
- ML is an alternative to least squares for fitting models
  - ML and Least Squares give same results if errors are normal and additive with constant variance
- Writing a likelihood function relies on the equation for your distribution
  - See Haddon 2011 for examples, or lecture for eqns.

Model fitting with nonlinear optimization

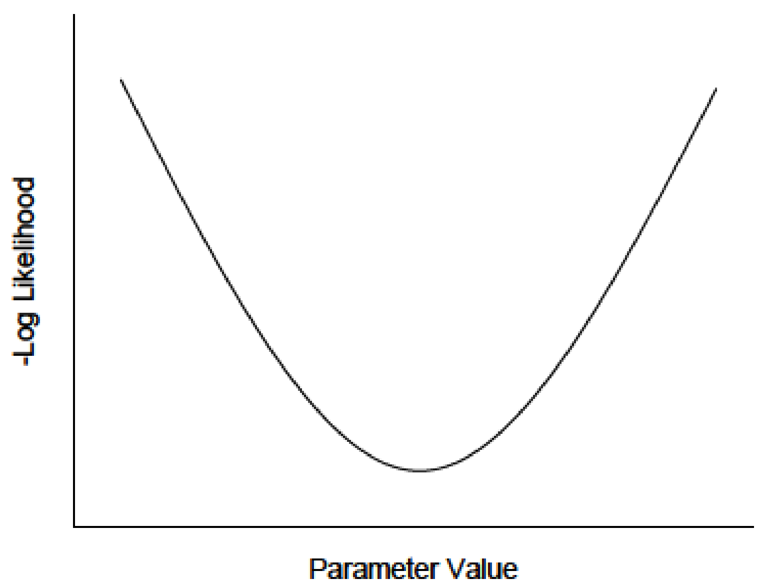
Readings:

Haddon 2011 (Section 3.4)

Acknowledgements: M. Wilberg, R. Latour

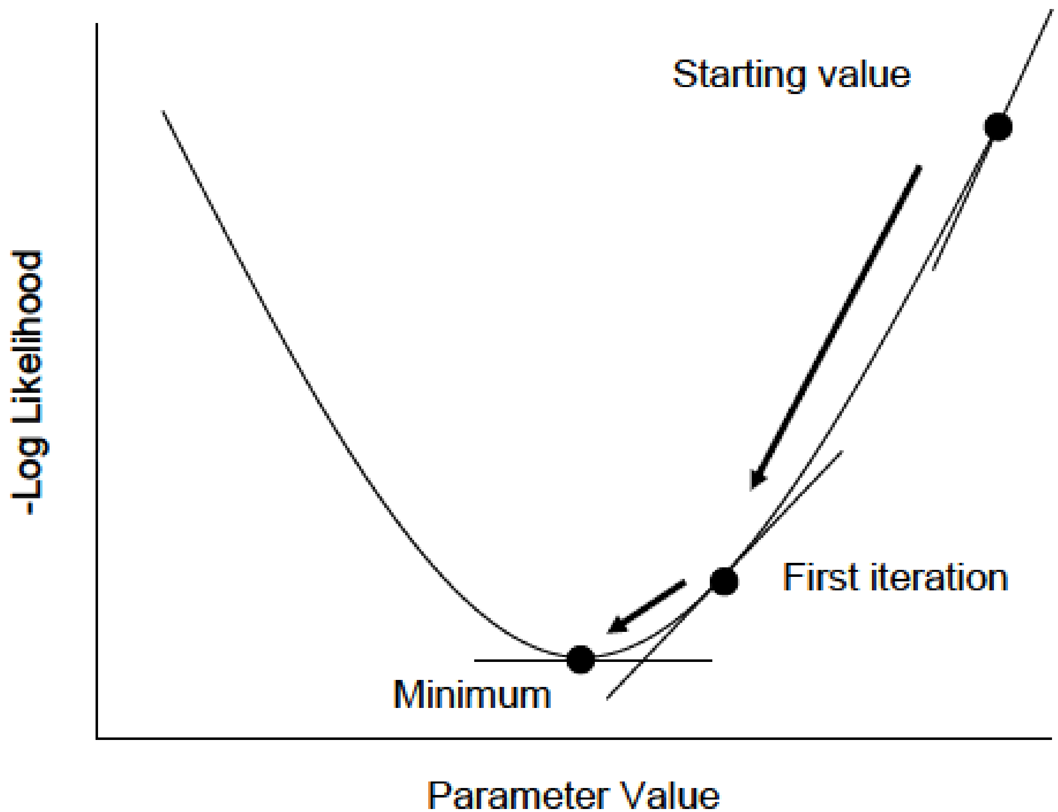
How to find minimum of -LL? Nonlinear Optimization

- **Nonlinear optimization** is the general term for trying to find the maximum or minimum of a function (e.g., likelihood fxn)
- Numerical solution (vs. Analytical solution) – approximate solution found through iterative searching

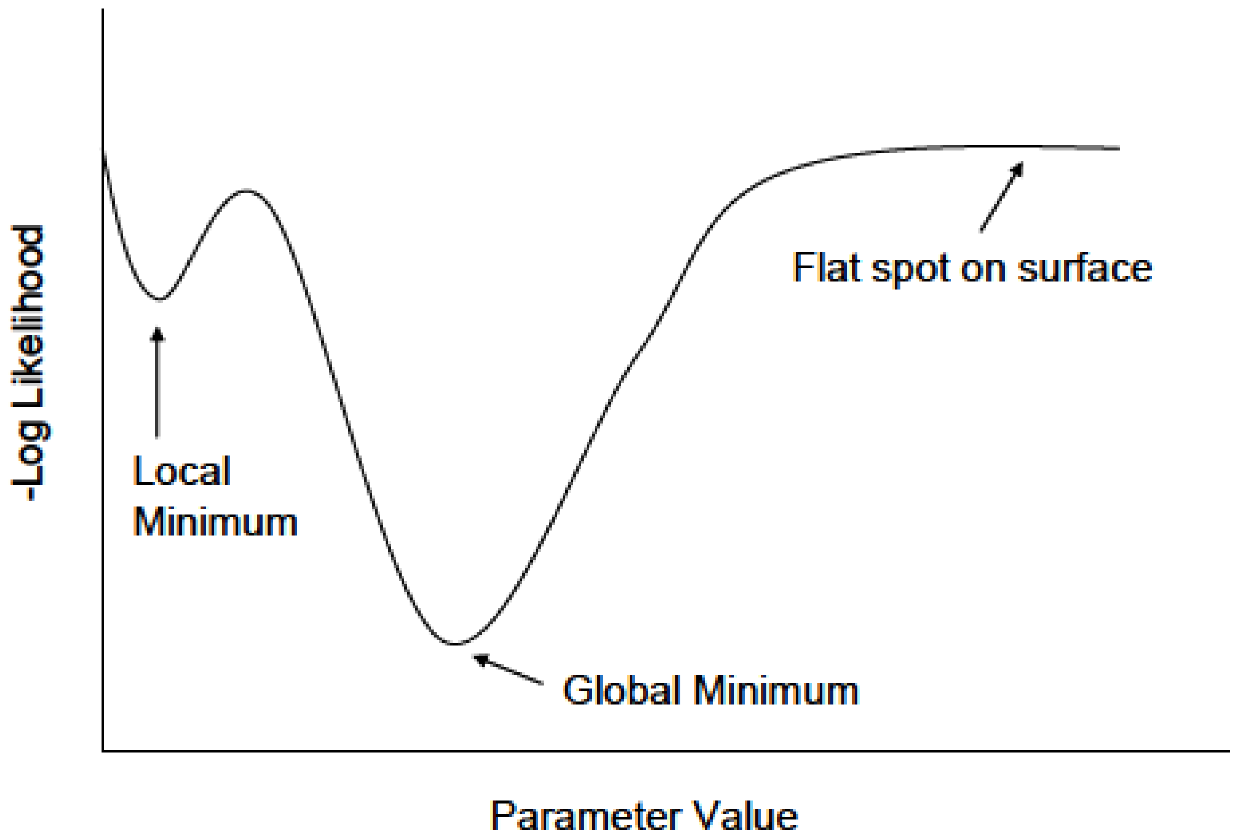


Minimization

- R has functions to do this: e.g., optim()



Potential Nonlinear Optimization Problems



optim() in R

- optim() uses an algorithm that relies on derivatives to find the optimal solution
- Some common outcomes:
  - it finds the “best” solution
  - it reaches its limit of iterations (can change *maxit*)
  - Misc. errors
- Check that convergence criteria were met:
  - **$convergence = 0 → SUCCESS!**
  - $convergence = 1 → max iterations reached

Potential Nonlinear Optimization Problems

1. Improper starting values
2. Model specification or coding error
3. Negative log likelihood function may be undefined for some parameter values
4. Parameterization problems (scaling, correlated parameters)
5. Uninformative data

1. Starting values

- Parameters need to be provided "good" starting values for all nonlinear optimization routines
  - The starting value should be: of the same sign and similar magnitude the expected solution
- Obtain starting values by eye
- Use several sets of starting values
- Can increase max number of iterations (using *maxit*) to allow longer search
  - optim(…, control=list(maxit=10000))

2. Model Specification and Coding Errors

- R cannot detect mistakes in your model
- Make sure that model predictions make sense
  - Do the numbers make sense?
  - Graph predictions over the data
- Try specific sets of parameter values for which you know the correct answer
- Fit model to simulated data

3. negLL function may be undefined for some parameter values

- Common example:
  - Sigma (σ) can’t be negative – can’t have a negative SD of residuals
  - Warning

```
Warning messages:
1: In log(sigma) : NaNs produced
```

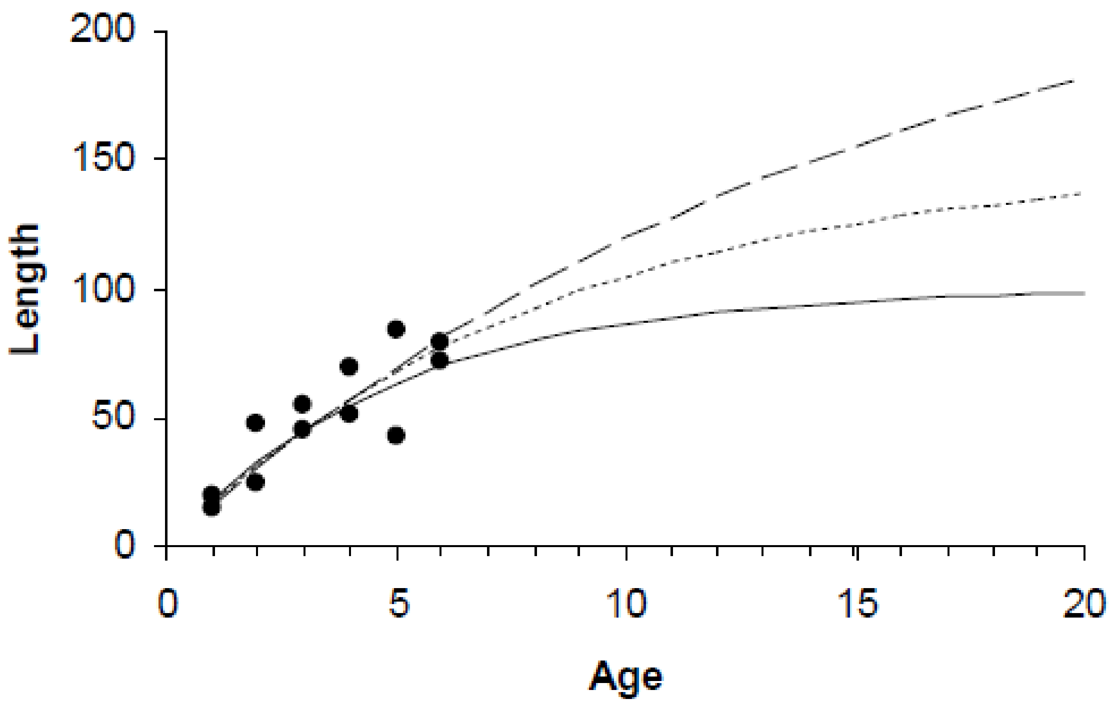
  - Possible solutions:
    - Ignore
    - Use a different method that allows constraints (e.g., "L-BFGS-B“)
    - Estimate log(σ) which can be negative, then backtransform

4. Parameterization Problems

- Many models can be written in several forms, each called a parameterization
- Parameters should have low correlations with one another
- Parameters should be of similar magnitude (e.g., avoid 0.000000529 and 3.54)
    - → estimating parameters in log space can help (e.g., ln(5.3E-07) = -14.45

5. Uninformative data

- The data may have little or no information about one or several parameters
- Only able to statistically estimate n-1 parameters



Summary 2 - Nonlinear Optimization

- **Nonlinear optimization** - general term for trying to find the maximum or minimum of a function (e.g., likelihood function).
  - Numerical solution used when can't Analytical solution)

Potential problem with nonlinear optimization:

1. Improper starting values
2. Model specification or coding error
3. Negative log likelihood function may be undefined for some parameter values
4. Parameterization problems (scaling, correlated parameters)
5. Uninformative data