Maximum Likelihood

Introduction:

The technique of maximum likelihood (ML) is a method to: (1) estimate the parameters of a model; and (2) test hypotheses about those parameters. There have been books written on the topic (a good one is *Likelihood* by A.W.F. Edwards, New York: Cambridge University Press, 1972), so this handout will serve only as a simple introduction to help understand the process.

There are two essential elements in maximum likelihood. They are (1) a set of data (which, of course, is necessary for all analyses); (2) a mathematical model that describes the *distribution of the variables* in the data set. The model that describes the distribution of the variables will have certain unknown quantities in it. These are called *parameters*. The purpose of maximum likelihood is to *find the parameters of the model that best explain the data in the sense of yielding the largest probability or likelihood of explaining the data*. In least squares (see the class notes chapter on Least Squares Estimation), one finds the parameter of the model that yield the minimum sum of squared prediction errors. Thus, maximum likelihood differs from least squares mostly in terms of the *criterion for estimating parameters*. In least squares, one minimizes the sum of squared errors; in maximum likelihood, one maximizes the probability of a model fitting the data. A second difference is that in using maximum likelihood, one must always make some assumption about the distribution of the data. Sometimes these distributional assumptions are made in least squares (e.g., MANOVA), but at other times they are not necessary (e.g., estimating regression parameters).

Estimation:

It is perhaps easiest to view the estimation part of maximum likelihood by starting with an example. Suppose that we wanted to estimate the mean and the standard deviation for a single variable. Let X_i denote the score of the variable for the ith observation and let n denote the total number of observations. Let us further assume that the scores are normally distributed. Then the likelihood of the ith observation is simply the ordinate of the normal curve for that observation, or

$$L(X_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X_i - \mu}{\sigma}\right)^2\right] \quad (1)$$

Here, μ and σ are, respectively, the mean and the standard deviation of the distribution. Because the observations are all independent (recall the assumption of the independence of the rows of the data matrix), the likelihood for any two observations is simply the product of their respective likelihoods. For example, the joint likelihood of the ith and jth observation is simply $L(x_i) \diamond L(x_j)$. Again, because of independence, the joint likelihood of any three observations will be the product of their individual likelihoods. Continuing with this logic, we see that the joint likelihood of the *vector* of observations (i.e., the vector **x**) is

$$L(\mathbf{x}) = \prod_{i=1}^{n} L(X_i) \quad (2)$$

Hence, the problem of maximum likelihood is to find μ and σ so that L(**x**) is maximized.

The problem of finding the maximum (or minimum) of a function with respect to the unknowns is a classic problem in calculus. That is, we want to differentiate the function L(**x**) with respect to μ and σ, then set these first partials to 0, and finally rework the equations so that we solve for μ and σ. That is, we want to find μ and σ such that

$$\frac{\partial L(\mathbf{x})}{\partial \mu} = 0, \quad \text{and} \quad \frac{\partial L(\mathbf{x})}{\partial \sigma} = 0 \tag{3}$$

The resulting estimates, usually denoted as and , are then known as the *maximum likelihood* estimators of μ and σ.

The process of maximum likelihood is almost always performed on the *natural logs* of the likelihood function. That is, in terms of the example, we want to find the μ and σ that maximize the log likelihood of the data given the model. The log likelihood is gotten by simply taking the log of both sides of Equation (2) or

$$\log(L(\mathbf{x})) = \sum_{i=1}^{n} \log(L(X_i)) \tag{4}$$

The log likelihood of X_i is simply the log of the right side of Equation (1), or

$$\log(L(X_i)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma) - \frac{1}{2}\left(\frac{X_i - \mu}{\sigma}\right)^2 \tag{5}$$

Substituting Equation (5) into Equation (4) and reducing gives the log likelihood of the sample as

$$\log(L(\mathbf{x})) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \tag{6}$$

The same calculus technique is used with the log likelihoods. That is, we want to find μ and σ that satisfy the equations

$$\frac{\partial \log[L(\mathbf{x})]}{\partial \mu} = 0, \quad \text{and} \quad \frac{\partial \log[L(\mathbf{x})]}{\partial \sigma} = 0 \tag{7}$$

There are two reasons for working with log likelihoods rather than with the likelihoods themselves. First, likelihoods are often (but not always) quantities between 0 and 1. Hence, taking the products of a large number of fractions can yield extreme rounding and truncation error, even with the most modern computers. Second, it is considerably easier to calculate the derivatives of the log likelihood than it is to obtain them for the likelihood because the log likelihood is a series of *sums* whereas the likelihood is a series of *products*. Those with experience

with calculus will recognize the ease in differentiating sums relative to differentiating a series of products.

Without going through the mathematics of differentiating Equation 6 with respect to μ and σ, setting the partials to 0, and then solving for μ and σ, we merely state the results. According to this example, the estimator or μ is

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n} \tag{8}$$

or the sample mean. Similarly, the estimator of σ is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}} \tag{9}$$

or the *sample* variance. Note that the maximum likelihood estimator of σ is a *biased* estimate of the population standard deviation because the divisor is *n* and not (*n* - 1). The extent of this bias becomes smaller as sample size increases.

In this example, it is easy to arrive at analytical expressions for the estimators μ and σ. This is actually unusual. Typically, one cannot solve the differential equation, or in many other cases, the equations can be solved but it would take considerable effort to do so. Here, numerical methods are used to find the maximum likelihood estimators. There are many different numerical methods, many of which are geared to a specific class of problems. They are too numerous to detail them here. Instead, one may look at them as highly sophisticated "trial and error" procedures. That is, they begin with some initial estimates of the parameters and evaluate the log likelihood for these estimates. They then find improved estimates based on the log likelihood surface of the original estimates and continue with this procedure in an iterative fashion until the log likelihood cannot be maximized any further.

To illustrate how this method works, Example 1 includes the SAS code and the output for estimating the mean and standard deviation for 50 scores selected from a normal distribution. The maximum likelihood technique used here is simply called a *grid* search. That is, it varies the value of the mean from -.10 to 0 in increments of .01 and the value of the standard deviation from .95 to 1.0, also in increments of .01. The value of the log likelihood is then plotted for each combination of the mean and the standard deviation. You can see how the largest value of the log likelihood (-22.937) occurs with a mean of -.07 and a standard deviation of .96, agreeing with the analytical estimates.

Example 1. SAS code and output for performing a grid search for maximum likelihood parameter estimates.

```
/* ----------------------------------------------------------
  Example of estimating parameters using maximum likelihood
  File of ~carey/p7291dir/maxlik.ex1.sas
  Problem: Given 50 observations what is the mean and
           what is the standard deviation, if the distribution
           can be assumed to be normal?
  See Class notes titled Maximum Likelihood Estimation
  ---------------------------------------------------------- */
title Example of Maximum Likelihood Estimation;
title2 Mean and standard deviation of a normal distribution;
data maxlik;
* generate 50 data random values from a normal distribution;
   array x x1-x50;
   do over x; x = rannor(91748); end;
* calculate the mean and the standard deviation of the
  50 scores;
   mean = mean(of x1-x50);
   css = css(of x1-x50);
   std1 = sqrt(css/49);
   std2 = sqrt(css/50);
   file print;
   put 'The mean is' (mean) (7.3) /
       'The corrected sum of squares is' (css) (7.3) /
       'The standard deviation dividing by (n-1) is'
          (std1)  (7.3)
       'The standard deviation dividing by   n   is'
          (std2)  (7.3);
/*
  print the headings for the log likelihood values
*/
   put // 'Grid of the log likelihoods' /;
   put  @33 'st. dev. =';
   array logl [6] logl1-logl6;
   do i=1 to 6; logl[i]=.94+.01*i; end;
   put (logl1-logl6) (' Mean' 6*10.3) /;
   do m=-.1 to 0 by .01;
      s=.94;
      do i=1 to 6;
```

```
          s=s+.01;
          var = s*s;
          logl[i]=0;
          do over x;
             z = (x - m)/s;
             loglx = -.5*log(var) -.5*z*z;
             logl[i]=logl[i] + loglx;
          end;
       end;
       put (m logl1-logl6) (5.2 6*10.3);
   end;
run;

Example of Maximum Likelihood Estimation
Mean and standard deviation of a normal distribution

The mean is -0.071
The corrected sum of squares is 46.040
The standard deviation dividing by (n-1) is  0.969
The standard deviation dividing by   n   is  0.960

Grid of the log likelihoods
                                st. dev. =
 Mean     0.950     0.960     0.970     0.980     0.990     1.000

-0.10   -22.965   -22.959   -22.964   -22.980   -23.005   -23.040
-0.09   -22.952   -22.946   -22.952   -22.968   -22.993   -23.028
-0.08   -22.944   -22.939   -22.945   -22.961   -22.987   -23.022
-0.07   -22.942   -22.937   -22.943   -22.959   -22.985   -23.020
-0.06   -22.946   -22.941   -22.946   -22.962   -22.988   -23.023
-0.05   -22.955   -22.950   -22.955   -22.971   -22.996   -23.031
-0.04   -22.970   -22.964   -22.969   -22.985   -23.010   -23.045
-0.03   -22.990   -22.984   -22.988   -23.004   -23.029   -23.063
-0.02   -23.015   -23.009   -23.013   -23.028   -23.052   -23.086
-0.01   -23.047   -23.040   -23.043   -23.057   -23.081   -23.114
-0.00   -23.084   -23.076   -23.079   -23.092   -23.115   -23.148
```

In reality the grid search would then be repeated around the area for a mean of -.07 and a standard deviation of .96 using a smaller increment. This would refine the estimates. This procedure could then be repeated until the estimates reach a predetermined degree of accuracy.

Hypothesis Testing:

Tests of hypothesis in ML usually involve a *likelihood ratio test*. The likelihood ratio test compares the log likelihoods of two models. The first model is a general model. The second model is a constrained model. The constrained model must be nested within the general model. That is, it uses the *same* parameters as the fixed model but sets some of these parameters to fixed values according to the hypothesis to be tested. The parameters that are estimated are called *free* parameters and those that are set to a predetermined/hypothesized value are called *fixed* or *constrained* parameters. Let L_G denote the log likelihood of the general model with k free parameters. Let L_C denote the log likelihood of the constrained model with $(k - j)$ free parameters, j being the number of constrained parameters. Then the likelihood ratio test is simply $2(L_G - L_C)$ or twice the difference between the log likelihoods. In large samples, this will be distributed as a χ^2 with j degrees of freedom. The degrees of freedom will always equal the number of free parameters in the general model less the number of free parameters in the constrained model. If the χ^2 is large and significant, then the hypothesis that generated the constrained model is rejected.

To illustrate the likelihood ratio test, consider a study of a single, two allele genotype in a wild population of flowers. The research question is whether the genotypic frequencies are in Hardy-Weinberg equilibrium or whether some factor such as natural selection or assortative mating perturb the genotypic frequencies away from the equilibrium values. The observed data (hypothetical) are given in the right hand side of Table 1.

Table 1. Distribution of genotypes and expected frequencies from a general and constrained (Hardy-Weinberg) model. p = frequency of allele A; x = conditional probability that the second allele is A given that the first allele is A.

		Predicted Frequency:	
Genotype	Observed Number	General	Constrained
AA	427	px	p^2
Aa	332	$2p(1 - x)$	$2pq$
aa	235	$q - p(1 - x)$	q^2

Under ordinary circumstances, the test of this hypothesis is not difficult. Let p denote the frequency of allele A and $q = (1 - p)$ denote the frequency of allele a. The estimate of p may be derived from the observed data as the number of *AA* flowers plus one-half the number of *Aa* divided by the total number of flowers or $(427 + 332/2)/994 = .597$. One could then construct predicted numbers for the three genotypes from the Hardy-Weinberg equations. For example, the predicted frequency of *AA* would be p^2 times the total number of plants or $(.597)^2 994 = 353.9$. The Pearson χ^2 would then test whether the genotypic frequencies depart from the Hardy-Weinberg expectations. The formula for the Pearson χ^2 is

$$\chi^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} \tag{10}$$

where O_i is the observed number and E_i is the predicted number for the ith genotype.

We can now develop a likelihood approach to this problem. First, we wish to develop a general model. Again, let p and $q = (1 - p)$ denote the frequencies of respectively alleles *A* and *a*. Let x denote the conditional probability that the second allele in a genotype is *A* given that the first allele in the genotype is *A*. The general model can then be written in terms of two free parameters, p and x. The Hardy-Weinberg model is a constrained model where $x = p$. (In Hardy-Weinberg, the probability of a second allele in a genotype given the first allele is simply the probability of picking that second allele from the population at large.) The expected genotypic frequencies for the general and constrained models are given in the right-hand side of Table 1.

The probability distribution for this case is very simple. In the general model, if a genotype is *AA* then is probability or "likelihood" is px. Furthermore, this is the probability associated with every *AA* flower in the sample. Hence, the joint likelihood of all of the *AA* flowers is simply (px) raised to the n_{AA} power where n_{AA} is the number of *AA* flowers. The log likelihood for all the *AA* flowers then becomes $n_{AA}\log(px)$. If we repeat this for genotypes *Aa* and *aa*, we can come to the formula for the general model as

$$L_G = n_{AA}\log(px) + n_{Aa}\log[2p(1-x)q] + n_{aa}\log[q - p(1-x)] \tag{11}$$

In the constrained, Hardy-Weinberg model $x = p$. Hence, the log likelihood for the constrained model becomes

$$L_C = n_{AA}\log(p^2) + n_{Aa}\log(2pq) + n_{aa}\log(q^2) \tag{12}$$

A SAS program to estimate the maximum likelihood parameters for both the general and the constrained models is given in Example 2. The program is somewhat complicated but it illustrates a more sophisticated method of performing a grid search for two parameters, x and p, in the general model and one parameter, p, in the constrained model. The output from the program is given immediately after the SAS code.

Example 2. SAS code and output for maximum likelihood analysis of allele frequencies.

```
/* ------------------------------------------------
  Example of a maximum likelihood analysis
  testing for Hardy-Weinberg equilibrium for
  genotypes of three flowers
  See handout (class notes) on maximum likelihood
```

```
  for background.
  this program is on ~carey/p7291dir/maxlik.3x2.sas
 ------------------------------------------------- */
title  Example of Maximum Likelihood Hypothesis Testing;
title2  Genotypic frequencies for flowers ;
data flowers;
  input geno1 $ geno2 $ geno3 $ n1 n2 n3;
cards;
AA Aa aa  427  332  235
;
data general;
   set flowers;
/* --- arrays --- */
   array geno geno1-geno3;
   array n n1-n3;
   array prob prob1-prob3;
   ntot=n1+n2+n3;
   array obs obs1-obs3;
   do over obs; obs=n/ntot; end;
/* --- Pearson chi square --- */
   p = (n1 + n2/2)/ntot;
   q=1-p;
   prob1=p*p; prob2=2*p*q; prob3=q*q;
   pearchi=0;
   do over prob;
      con = ((obs - prob)**2)/prob;
      pearchi = pearchi + con;
   end;
   pearchi = ntot*pearchi;
   pearprob = 1 - probchi(pearchi,1);


/* ------------------------------------ *
 * MAXIMUM LIKELIHOOD: two parameter model *
 * ------------------------------------ */
file print;
put '------------------' /
    '   General Model'/
    '------------------'/;
link init;
loop2:
        do p = plow to phigh by delta;
            do x = xlow to xhigh by delta;
                 link getlogl;
```

```
           end;
        end;
        link deltait;
        if delta > .0001 then goto loop2;
link printit;
lg=maxlogl;

/* ------------------------------------- *
 * MAXIMUM LIKELIHOOD: one parameter model *
 * ------------------------------------- */

put ///'-------------------' /
       ' Constrained Model'/
       '-------------------'/;
link init;
loop1:
        do p = plow to phigh by delta;
              x=p;
              link getlogl;
        end;
        link deltait;
        if delta > .0001 then goto loop1;
link printit;
lc=maxlogl;
lrchi = 2*(lg - lc);
lrprob = 1 - probchi(lrchi,1);
put /  'Likelihood ratio chi square' (lrchi) (7.3)
       +3 'Probability' (lrprob) (7.4);
put   '     Pearson     chi square' (pearchi) (7.3)
       +3 'Probability' (pearprob) (7.4);
/* stop the program */
stop;
* initialization section;
init:
      maxlogl=-10**10;
      plow=.01;
      phigh=.99;
      xlow=.01;
      xhigh=.99;
      delta=.01;
      return;
/* ----------------------------------
   --- calculate the log likelihood ---
   ---------------------------------- */
```

```
getlogl:
          prob1=p*x;
          prob2=2*p*(1-x);
          prob3=(1-p) - p*(1-x);
          logl=0;
          do over geno;
             if prob le 0 or prob ge 1 then logl=-10**10;
             else logl=logl + n * log(prob);
          end;
          if logl > maxlogl then do;
             pmax=p;
             xmax=x;
             maxlogl=logl;
          end;
          return;
/* ------------------------------
   --- change the delta value ---
   ------------------------------ */
deltait:
          delta = .1*delta;
          plow = pmax - 2*delta;
          phigh = pmax + 2*delta;
          xlow = xmax - 2*delta;
          xhigh= xmax + 2*delta;
          return;
/* --- print out the results --- */
printit:
          put 'Log likelihood is ' (maxlogl) (8.4);
          put 'Estimate of p is ' (pmax) (7.4);
          put 'Estimate of x is ' (xmax) (7.4);
          p = pmax; x=xmax;
          link getlogl;
          put 'Geno     Obs     Pre';
          do over prob;
              put (geno obs prob) (@3 $2. 2*8.4);
          end;
          return;
run;


Example of Maximum Likelihood Hypothesis Testing
Genotypic frequencies for flowers

------------------
```

```
   General Model
-------------------

Log likelihood is -1063.78
Estimate of p is  0.5980
Estimate of x is  0.7210
Geno     Obs     Pre
  AA  0.4296  0.4312
  Aa  0.3340  0.3337
  aa  0.2364  0.2352

-------------------
 Constrained Model
-------------------

Log likelihood is -1110.54
Estimate of p is  0.5980
Estimate of x is  0.5980
Geno     Obs     Pre
  AA  0.4296  0.3576
  Aa  0.3340  0.4808
  aa  0.2364  0.1616

Likelihood ratio chi square 93.521   Probability 0.0000
     Pearson     chi square 93.136   Probability 0.0000
```

Note first the bottom of the output on Example 3. This shows both the likelihood-ratio χ^2 and the Pearson χ^2 for these same data. Both have one degree of freedom and are highly significant. Both are also very similar in magnitude. Generally when sample size is large--as it is in this example--the differences between the two approaches will be trivial. The difference, however, generally favors the likelihood approach.

The estimate of *p* is the same in both the general and the constrained model. Such similarity in parameter estimates across models is very unusual, however. Most often estimates of the same parameter will differ as one goes from a general to a constrained model. They are the same here only because of the particular nature of this problem.