Professor Ross L. Matsueda

LECTURE 12: MAXIMUM LIKELIHOOD ESTIMATION

I. BRIEF REVIEW OF DERIVATIVES.
II. INTRODUCTION TO MAXIMUM LIKELIHOOD ESTIMATION.
III. ML ESTIMATION OF A UNIVARIATE DISTRIBUTION AND BIVARIATE REGRESSION.
IV. ML ESTIMATION IN COVARIANCE STRUCTURE MODELS.
A. MAXIMIZING THE LIKELIHOOD FUNCTION
B. INFORMATION MATRIX AND LIKELIHOOD RATIO TEST.

In this lecture, we will introduce maximum likelihood estimation (ML), which is the most common estimation procedure used in covariance structure analysis. We'll begin with a brief review of derivatives, and then discuss the need and rationale of maximum likelihood estimation. I'll then apply the principle to estimating the mean and variance of a univariate distribution and then to a bivariate regression. We'll conclude by applying ML to the covariance structure (LISREL) model. To understand the maximum likelihood estimation principle and likelihood testing method, it will be useful to know a little calculus.

I. BRIEF REVIEW OF DERIVATIVES.

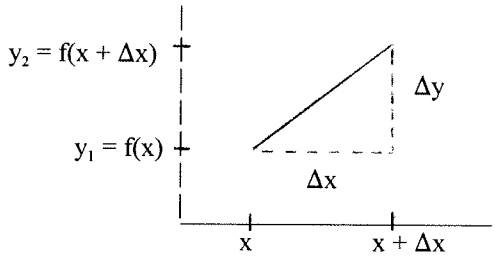
A derivative is an equation for a slope of a line. For two variables, Y and X, we first express Y as some function of X, $Y = f(X)$, then define the derivative as:

$dY/dX = d[f(X)]/dX$ which is read "the derivative of Y with respect to X"

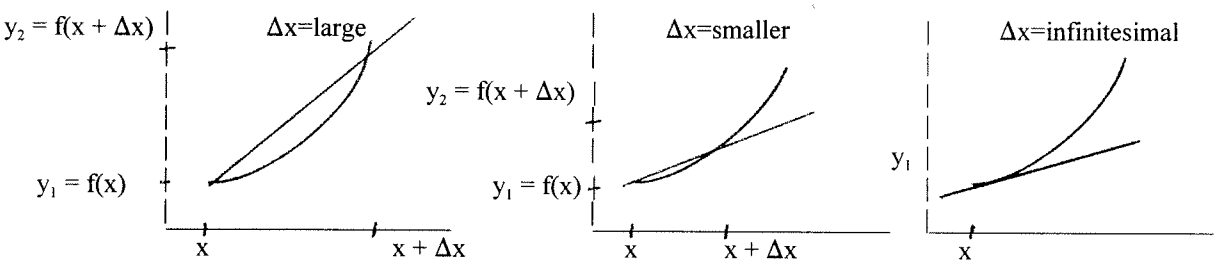
The nature of the slope depends on the nature of the function relating Y and X. For example, if the function is linear, the slope would be constant for values of X, as in a linear regression. But if the function is nonlinear, the slope will vary depending on the value of X. Here derivatives are very handy. But how does one determine the slope of a nonlinear function, since the slope varies by where you are on X? Begin with the logic of a linear function: to determine the slope, take any two points on the line and determine the change in Y, call it ΔY associated with a given change in X, call it ΔX. So $b = \Delta Y/\Delta X$. Another way of expressing this is in terms of the general function $Y = f(X)$. Then the change in Y for a given change (ΔX) in X becomes $\Delta Y = f(X + \Delta X) - f(X)$. Then our slope becomes:

$$b = \frac{\Delta Y}{\Delta X} = \frac{f(X + \Delta X) - f(X)}{\Delta X}$$

$$\frac{\text{rise}}{\text{run}}$$



Now, what about a curved line? We could use the same strategy, and add ΔX to X and compute the slope of our function f(X). Notice that if we add a big ΔX to our function, we'll get an inaccurate approximation of the slope of a curve; in fact, as we reduce ΔX, we get a better and better approximation:



The derivative is defined as the limiting slope of the function as ΔX approaches zero:

$$\frac{dY}{dX} = \lim_{\Delta x \to 0} \frac{f(X + \Delta X) - f(X)}{\Delta X}$$

For a curved function, it gives the slope of a tangent line at a given value of X. Here's an example: Suppose we are interested in the quadratic function $Y = f(X) = 3 + X^2$. We can express the derivative:

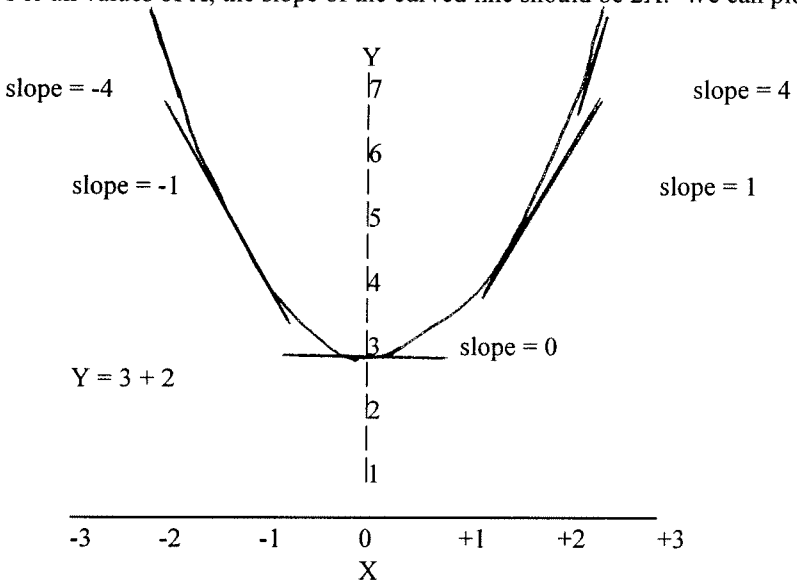
$$\frac{dY}{dX} = \lim_{\Delta x \to 0} \frac{f(X + \Delta X) - f(X)}{\Delta X} = \frac{[3 + (X + \Delta X)^2 - (3 + X^2)}{\Delta X} = \frac{3 + X^2 + 2X\Delta X + \Delta X^2 - (3 + X^2)}{\Delta X}$$

$$= \frac{2X\Delta X + \Delta X^2}{\Delta X} = \frac{\Delta X(2X + \Delta X)}{\Delta X} = \lim_{\Delta x \to 0} (2X + \Delta X) = 2X$$

Since ΔX becomes zero at the limit, we are left with the derivative of Y with respect to X for the function $Y = 3 + X^2$ is 2X:

$$\frac{dY}{dX} = 2X$$

For all values of X, the slope of the curved line should be 2X. We can plot this and visualize the slopes:



The derivative can also tell us where the minimum of the function is. In the above graph, the minimum of the function occurs when X = 0 and the slope (derivative) is zero. In general, we can find the minimum of a function by setting the derivative (slope) equal to zero and then solving for the value of X:

$$\frac{dY}{dX} = 2X = 0 \qquad X = 0 \qquad \text{(Divide both sides by 2)}$$

We can also take the second derivative, which is the derivative of the derivative, and gives us the change in the change, or how much the slope is changing at a given value of X. Since our first derivative of f(X) with respect to X was 2X, the derivative of 2X is 2:

$$\frac{d^2Y}{dX^2} = 2$$

The second derivative tells us whether we our solution to our equation is a maximum or a minimum. If the second derivative is positive, it means the curve is concave up (as above) and we have a minimum; if the second derivative is negative, the curve is concave down and we have a maximum. The figure above corresponds to a positive second derivative, and thus a minimum.

In general, for complicated functions, we follow rules for obtaining derivatives. Here are a few:

1.	$dc/dX = 0$	for Y = c = constant. Derivative of a constant is zero.
2.	$dcX/dX = c$	for Y = cX. Derivative of a constant times X is the constant.
3.	$d(cu)/dX = c\ du/dX$	for Y = cu, where c is a constant and u is a function of X, u(X).
4.	$d(u + v)/dX = du/dX + dv/dX$	for Y = v + u, where v and u are functions of X, v(X) and u(X)
5.	$d(X^n)/dX = nX^{n-1}$	for Y = X^n, where n is a power of X.
6.	$d(u\ v)/dX = v\ du/dX + u\ dv/dX$	for v(X) and u(X) both functions of X.
7.	$de^x/dX = e^x$	for Y = e^x, exponential of X.
8.	$de^v/dX = e^v\ dv/dX$	for Y = e^v, exponential of v, a function of x v(X).
9.	$d(\ln X)/dX = 1/X$	for Y = ln X, the natural logarithm of X.
10.	$d(\ln v)/dX = 1/v\ dv/dx$	for Y = ln v, the natural logarithm of v, a function of X, v(X).

It will also be handy to use partial derivatives, which specify derivatives of multivariate functions. Suppose $Y = f(X_1, X_2)$, then $\partial Y/\partial X_1$ denotes the partial derivative of Y with respect to X_1. This gives the slope of X_1 while treating X_2 (and all other Xs) as constants. Note that this corresponds to the partial slopes of a multivariate regression. To find the minimum of a function, we would want to find the values of X_1 and X_2 that give zero partial derivatives. For example, if

$$Y = 3 + 2X_1^2 + X_1 X_2 + 5X_2$$

$\partial Y/\partial X_1 = 4X_1 + X_2$ since $d(X^n)/dX = nX^{n-1}$ and $dcX/dX = c$

$\partial Y/\partial X_2 = X_1 + 5$ since $dcX/dX = c$

We can also take the partial derivative of partial derivatives, which are called second-order derivatives (making initial derivatives first-order). The second-order derivative of Y with respect to X is designated $\partial^2 Y/\partial X^2$. This gives us the change in the change, telling us how much the partial slope (first derivative) is changing at a given value of X. To calculate second derivatives, use the same rules as above, but apply them to the first derivatives. Thus for our example:

$$Y = 3 + 2X_1^2 + X_1 X_2 + 5X_2$$

$\partial Y/\partial X_1 = 4X_1 + X_2$ $\partial^2 Y/\partial X_1^2 = 4$ $\partial^2 Y/\partial X_1 X_2 = 1$

$\partial Y/\partial X_2 = X_1 + 5$ $\partial Y^2/\partial X_2^2 = 0$ $\partial Y^2/\partial X_2 X_1 = 1$

Second derivatives are important in maximum likelihood theory because (1) they are related to the covariance matrix of estimates, and thus, standard errors; and (2) they help in minimizing functions: at the value of X where the derivative is zero (horizontal slope), if the second derivative is negative, the derivative has located a maximum of the function; if it is positive, the derivative has located a minimum of the function.

It will be convenient to place first and second partial derivatives in vectors and matrices. For example, we can express the derivatives of an equation $Y = b_1 X_1 + b_2 X_2$, where $b_1 = \partial Y/\partial X_1$ and $b_2 = \partial Y/\partial X_2$:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \qquad \mathbf{b}' = [b_1 \quad b_2] \quad = \quad [\partial Y/\partial X_1 \;\; \partial Y/\partial X_2] =$$

$$Y = \mathbf{b}'\mathbf{X}$$

$$\mathbf{b}' = \partial Y/\partial \mathbf{X} = [\partial Y/\partial X_1 \;\; \partial Y/\partial X_2]$$

The matrix of second-order partial derivatives would be:

$$\partial Y^2/\partial \mathbf{X}\mathbf{X}' = \begin{bmatrix} \partial Y^2/\partial X_1^2 & \partial Y^2/\partial X_1 X_2 \\ \partial Y^2/\partial X_2 X_1 & \partial Y^2/\partial X_2^2 \end{bmatrix}$$

II. INTRODUCTION TO MAXIMUM LIKELIHOOD ESTIMATION.

Recall that when we discussed the walking dog model, we found that in an overidentified model, there is more than one way of computing parameters from observable population moments, which implies that when in the sample, there is more than one way of estimating parameters from observable sample moments. At that point, we lacked a principle by which to weight the different estimators in some optimal way. Maximum likelihood is such a principle. It ends up weighting the estimators in a way that yields an asymptotically unbiased (consistent) and efficient estimator. For just-identified models, there is only one way of estimating parameters from sample moments, and in that case, the method of moments and maximum likelihood (under the assumption of multivariate normality) are identical.

A. ADVANTAGES AND LIMITATIONS OF MAXIMUM LIKELIHOOD.

When applied to the general covariance structure (LISREL) model, maximum likelihood has the following advantages:

1. Asymptotic normal estimator. As the sample size approaches infinity, the sampling distribution of the ML estimator approaches a normal distribution, allowing us to use statistical theory based on the normal distribution to construct test statistics:

   $\theta_{ML} \rightarrow N[\theta, AVAR(\theta_{ML})]$, where $AVAR(\theta_{ML})$ is the asymptotic variance of θ_{ML}

2. Consistent estimator. As the sample size approaches infinity, the sampling distribution of the ML estimator has a mean of the population parameter:

   $\underset{n\rightarrow\infty}{\text{plim}}\, E(\theta_{ML}) = \theta.$

3. Best (efficient) asymptotic normal estimator. As the sample size approaches infinity, the sampling distribution of the estimator has a variance that is as small or smaller than any other consistent normal estimator. It reaches the Cramer-Rao lower bound for consistent estimators.

   $AVAR(\theta_{ML}) = - E[\partial^2 \ln(\mathcal{L})/\partial\theta\partial\theta']^{-1} = E[\partial \ln(\mathcal{L})/\partial\theta][\partial \ln(\mathcal{L})/\partial\theta']^{-1}$

4. A general intuitively appealing principle that can be applied to interpret or evaluate other ad hoc estimators.

Limitations of maximum likelihood in covariance structure analysis:

1. The general properties (consistency and asymptotic efficiency) are asymptotic; thus, very large samples are needed to capitalize on such properties. Finite sample properties of the estimator are unknown in the general case; we know of finite properties only for a small number of specific classes of models. (Some Monte Carlo evidence is useful here.)

2. The desirable properties of ML are predicated on specific assumptions of the distribution of the variables. In general, all variables are assumed to be continuous and normally-distributed. (This can be relaxed in a few specific cases -- e.g., perfectly-measured exogenous variables.) We don't have analytical results on the robustness of the estimator to departures from such assumptions, but there are some useful Monte Carlo results.

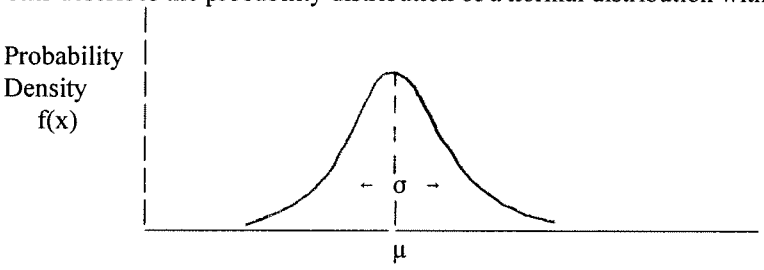
3. Because ML is a full-information (system) estimator, misspecification in one portion of the model can spill over and bias estimates in another.

B. LOGIC OF MAXIMUM LIKELIHOOD.

The principle of maximum likelihood estimation begins with the assumption that our model is the true population model that generated the sample data we observe. *Given the sample data*, ML then chooses values of a model's parameters that maximizes the likelihood that the sample data was indeed generated from the specified model. We can apply this logic to three examples of estimation: the mean, the slope of a bivariate regression, and the parameters of the LISREL model. Suppose we're interested in estimating μ, the population mean of a normally-distributed random variable drawn from a random sample. We begin with the probability density function of a normally-distributed variable:

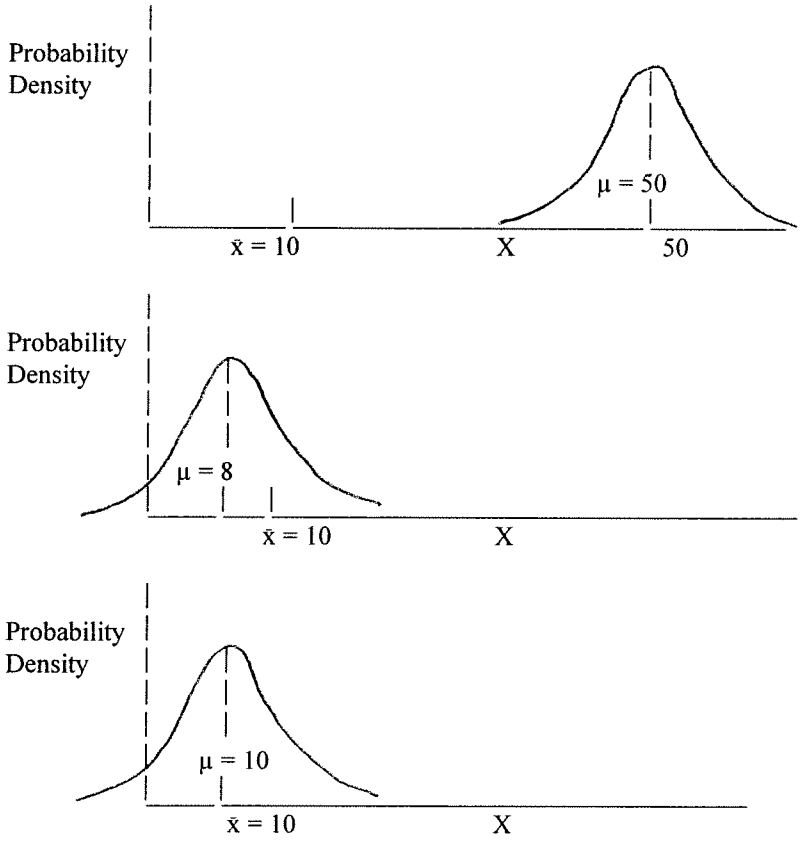
$$f(x_i | \mu, \sigma^2) = [1/(2\pi\sigma^2)]^{1/2} \exp\{[-1/(2\sigma^2)](x_i - \mu)^2\}$$

This describes the probability distribution of a normal distribution with mean μ and variance σ^2:



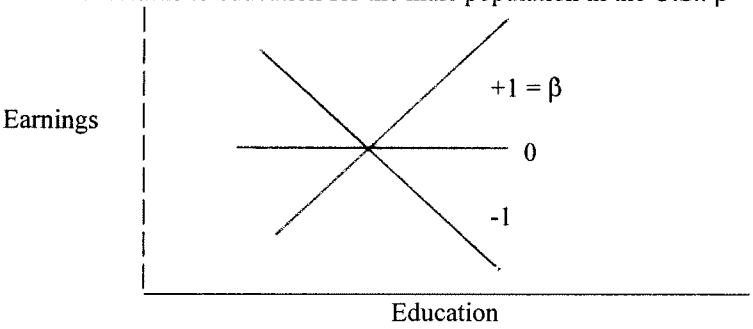
The probability density function f(x) above will trace a normal distribution if we graphed frequency (probability) distributions of x.

Now suppose we have three candidates for the population mean: $\mu_1 = 10$, $\mu_2 = 8$, and $\mu_3 = 50$. For each of these population means, suppose we were able to run a sampling experiment in which we drew numerous random samples, and computed the sample mean for each sample. We can graph the *sampling distribution of the estimator of the mean* -- not to be confused with the distribution of x -- given these three population values (and known sampling variances). Now suppose we go out and collect data on some random sample, and compute the sample mean, which we know is an unbiased estimate, and find that $\bar{x} = 10$. From which population is our sample most likely to have come?

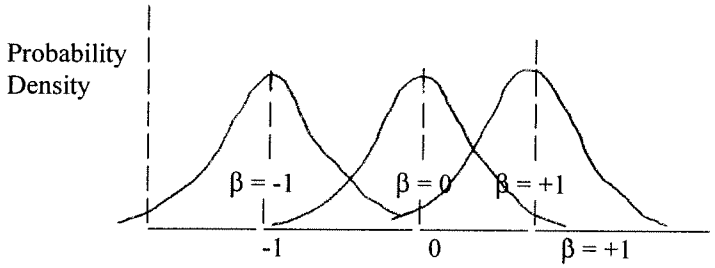


Given the sample mean of 10, for $\mu = 50$ to hold, we would have to attribute the difference between 10 and 50 to sampling variability, which is quite unlikely. It is more likely that $\mu = 8$ is the population parameter, since the probability that the sample estimate 10 differs from 8 is more likely due solely to sampling variability. But clearly, the "best guess" estimate of μ is 10, since that value maximizes the likelihood that the sample mean was generated by a given population μ. It turns out that the sample mean is the maximum likelihood estimator of the mean.

We can apply the same logic to a regression coefficient in a bivariate regression. Suppose we have three candidate values of returns to education for the male population in the U.S.: $\beta = 0$, $\beta = 1$, $\beta = -1$.



And suppose we drew a sample and used the OLS estimator $\beta = +1$. Again, given the sample, the most likely value of β that could have generated the sample is $\beta = +1$. The least likely is -1, since it is least likely that it differs from β by sampling variability alone. It turns out that the OLS estimator of β in a regression on observables is the maximum likelihood estimator. We can graph the sampling distributions that correspond to the three parameters (on the same graph):



Finally, we can apply the same logic to estimate $\Sigma(\theta)$. Let's assume we're estimating our walking dog model with four variables, ten moments, and nine parameters to be estimated. Suppose we have three candidate population models that differ in only one parameter, γ (all other parameter values are identical). This gives rise to three different population models, each implying a different covariance matrix implied by the population parameters $\Sigma(\theta)$:

Model 1: $\gamma_1 = 1$

$$\Sigma_1(\theta) = \begin{bmatrix} 45 & & & \\ 18 & 45 & & \\ 9 & 9 & 36 & \\ 9 & 9 & 9 & 36 \end{bmatrix}$$

Model 2: $\gamma_2 = 2$

$$\Sigma_2(\theta) = \begin{bmatrix} 72 & & & \\ 45 & 72 & & \\ 18 & 18 & 36 & \\ 18 & 18 & 9 & 36 \end{bmatrix}$$

Model 3: $\gamma_3 = 5$

$$\Sigma_3(\theta) = \begin{bmatrix} 261 & & & \\ 234 & 261 & & \\ 45 & 45 & 36 & \\ 45 & 45 & 9 & 36 \end{bmatrix}$$

Now, suppose we drew a sample and estimated our sample moments, **S**:

$$\mathbf{S} = \begin{bmatrix} 45 & & & \\ 18 & 45 & & \\ 9 & 9 & 36 & \\ 9 & 9 & 9 & 36 \end{bmatrix}$$

Clearly, the model that is least likely to have generated the sample moments is Model 3, since the discrepancy between sample covariance matrix and covariance matrix implied by the model, $\mathbf{S} - \Sigma(\theta)$, is very large and unlikely to be due solely to sampling variability. Model 2's discrepancy matrix has entries that are smaller, so it is more consistent with the sample data. But clearly, Model 1 is most likely to have generated the sample moment matrix, since the discrepancy matrix $\mathbf{S} - \Sigma_1(\theta) = \mathbf{0}$ is smaller than that of Models 2 and 3. If that discrepancy matrix is smaller than that of any other possible set of parameter values, then it would be the maximum likelihood estimator of γ. This example was contrived such that the sample data fit the true model precisely; in general, we would not expect the overidentifying restriction to hold exactly in the sample (because of sampling variability), even if the model is correctly-specified.

So, the principle of maximum likelihood estimation chooses parameter values in a way that maximizes probability that the model generated the sample data. By doing so, it provides consistent and asymptotically efficient estimates for identified models. But exactly how are those parameter values chosen?

III. ML ESTIMATION OF A UNIVARIATE DISTRIBUTION AND BIVARIATE REGRESSION.

To derive maximum likelihood estimators, we need to follow four steps:

1. **Probability Density Function:** Specify the distribution of variables in the population (joint probability density function). We will assume a multinormal density function.

2. **Likelihood Function:** Specify a likelihood function based on the probability density function.

3. **Maximize Log-Likelihood:** Maximize the log-likelihood function by taking partial derivatives with respect to each parameter (placed in a vector), setting each of the resulting equations equal to zero, and then solving for parameters.

4. **Asymptotic Covariance Matrix:** Obtain the asymptotic covariance matrix of the estimates by taking the second partial derivatives of the likelihood function with respect to each parameter, take the expected value of minus one times the resulting matrix (which gives the "information matrix", and then invert the matrix.

A. ML ESTIMATION OF THE MEAN AND VARIANCE OF A NORMAL DISTRIBUTION.

To illustrate the method, let's apply it to a simple case, and run through the above four steps. We'll begin with a univariate normal distribution.

1. Probability Density Function:
Let X be a random variable such that $X \sim N(\mu, \sigma^2)$. The probability density function for a normally distributed variable is, for *one* observation:

$$f(x \mid \mu, \sigma^2) = [1/(2\pi\sigma^2)]^{1/2} \exp\{[-1/(2\sigma^2)](x - \mu)^2\}$$

For n independent observations $(x_1, x_2, \ldots x_n)$, the joint distribution is simply the product of their univariate distributions: $f(x_1, x_2, \ldots x_n \mid \mu, \sigma^2) = f(x_1 \mid \mu, \sigma^2) \times f(x_2 \mid \mu, \sigma^2) \times \ldots \times f(x_n \mid \mu, \sigma^2) = \prod_i f(x_i \mid \mu, \sigma^2)$

$$f(x_1, x_2, \ldots x_n \mid \mu, \sigma^2) = [1/(2\pi\sigma^2)]^{n/2} \exp\{[-1/(2\sigma^2)] \sum_{i=1}^{n}(x_i - \mu)^2\}$$

which is the joint density of the sample given μ and σ^2. Usually, we read this as expressing the sample as a function of μ and σ^2.

2. Likelihood Function:
Mathematically, the above equation can also be read as expressing the parameters μ and σ^2 as a function of the sample. The likelihood of the sample, given the sample, expresses the parameters as a function of the sample: When expressed this way, the joint density is called a likelihood function $\mathcal{L}$:

$$\mathcal{L}(\mu, \sigma^2 \mid x_1, x_2, \ldots x_n) = [1/(2\pi\sigma^2)]^{n/2} \exp\{[-1/(2\sigma^2)] \sum_{i=1}^{n}(x_i - \mu)^2\}$$

We want to maximize $\mathcal{L}$: given the observed sample values of $X(x_1, x_2, \ldots x_n)$, what values of μ and σ^2 will maximize $\mathcal{L}$, the likelihood of the sample? It is computationally easier to maximize the natural logarithm of $\mathcal{L}$ than $\mathcal{L}$ itself:

$$\ln \mathcal{L}(\mu, \sigma^2 \mid x_1, x_2, \ldots x_n) = \ln ([1/(2\pi\sigma^2)]^{n/2} \exp\{[-1/(2\sigma^2)] \sum_{i=1}^{n}(x_i - \mu)^2\})$$

Let's simplify this ($\ln \mathcal{L}$ refers to the left-hand side above):

$\ln \mathcal{L} = \ln ([(2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\{[-1/(2\sigma^2)] \sum_{i=1}^{n}(x_i - \mu)^2\}])$ since $(1/xy)^{1/2} = x^{-1/2} y^{-1/2}$

$\ln \mathcal{L} = \ln(2\pi)^{-n/2} + \ln(\sigma^2)^{-n/2} + \ln \exp[-1/(2\sigma^2)\sum_{i=1}^{n}(x_i - \mu)^2]$ $\ln(ab) = (\ln a)(\ln b) = \ln a + \ln b$

$\ln \mathcal{L} = -(n/2) \ln 2\pi - (n/2) \ln \sigma^2 - 1/(2\sigma^2) \sum_{i=1}^{n}(x_i - \mu)^2$ $\ln X^a = a \ln X$; $\ln[\exp(Xa)] = Xa$

3. Maximize Log-Likelihood:
Now to find the value of μ that maximizes the likelihood of the sample, we maximize the log-likelihood function, by taking the partial derivative with respect to μ and setting the equation equal to zero:

$\ln \mathcal{L} = -(n/2) \ln 2\pi - (n/2) \ln \sigma^2 - 1/(2\sigma^2) \sum_{i=1}^{n}(x_i - \mu)^2$ (note first term is a constant - no μ)

$\partial(\ln \mathcal{L})/\partial\mu = (-1/2\sigma^2)\, 2 \sum_{i=1}^{n}(x - \mu)\,(-1)$ $du^n/dx = n\, u^{n-1}\, du/dx$

$\partial(\ln \mathcal{L})/\partial\mu = 2/(2\sigma^2) \sum_{i=1}^{n}(x_i - \mu)$ set this equal to zero and solve for μ

$0 = 1/\sigma^2 \sum_{i=1}^{n}(x_i - \mu)$ multiply both sides by σ^2

$0 = \sum_{i=1}^{n}(x_i - \mu)$ distribute the summation:

$0 = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\mu$ add $\sum_{i=1}^{n} \mu = n\,\mu$ to both sides

$n\,\mu = \sum_{i=1}^{n}x_i$ divide both sides by n

$\hat{\mu}_{ML} = \sum_{i=1}^{n}x_i/n$

The maximum likelihood estimator of the mean is exactly the sample mean, which we already knew is unbiased and efficient in finite samples. This shows it also has ML properties -- consistent and asymptotically efficient.

Do the same for σ^2:

$\ln \mathcal{L} = -(n/2) \ln 2\pi - (n/2) \ln \sigma^2 - 1/(2\sigma^2) \sum_{i=1}^{n}(x_i - \mu)^2$ $1/(2\sigma^2) = (1/2)(1/\sigma^2) = (1/2)(\sigma^2)^{-1}$

$\ln \mathcal{L} = -(n/2) \ln 2\pi - (n/2) \ln \sigma^2 - (1/2)(\sigma^2)^{-1} \sum_{i=1}^{n}(x_i - \mu)^2$ $d(c \ln a)/dx = c\ 1/a$

$\partial(\ln \mathcal{L})/\partial\sigma^2 = -(n/2)(1/\sigma^2) - (-1)(1/2)(\sigma^2)^{-2} \sum_{i=1}^{n}(x_i - \mu)^2$ $d(c\, a^{-1})/dx = c\,(-1)\, a^{-2}$

$\partial(\ln \mathcal{L})/\partial\sigma^2 = -\,n/(2\sigma^2) + 1/[2(\sigma^2)^2] \sum_{i=1}^{n}(x_i - \mu)^2$ $(\sigma^2)^{-2} = 1/(\sigma^2)^2$

Now set the equation equal to zero and solve for σ^2:

$0 = -\,n/(2\sigma^2) + 1/[2(\sigma^2)^2] \sum_{i=1}^{n}(x_i - \mu)^2$

$n/(2\sigma^2) = 1/[2(\sigma^2)^2] \sum_{i=1}^{n}(x_i - \mu)^2$ add $n/(2\sigma^2)$ to both sides

$n = 1/\sigma^2 \sum_{i=1}^{n}(x_i - \mu)^2$ multiply both sides by $2\sigma^2$

$n\,\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2$ multiply both side by σ^2

$\hat{\sigma}^2_{ML} = [\sum_{i=1}^{n}(x_i - \mu)^2]/n$ divide both sides by n

The maximum likelihood estimator of the variance is biased in finite samples, since the unbiased estimator uses n - 1 in the denominator. We know, however, that as the sample size increases, the difference between n and n - 1 decreases, until they are equivalent when $n = \infty$. This is another way of saying the ML estimator of the variance is consistent.

Aside: After differentiating the likelihood function with respect to parameters, we don't know if we've arrived at a maximum or minimum of the function. To determine this, take the second derivative, and if it is negative, we have a maximum; if it is positive, we have a minimum. For example, for the mean, we'll show below that:

$\partial(\ln \mathcal{L})/\partial\mu = 1/\sigma^2 \sum_{i=1}^{n}(x_i - \mu)$ $\partial^2(\ln \mathcal{L})/\partial\mu^2 = -n/\sigma^2$ We have a maximum for the mean.

4. Asymptotic Covariance Matrix:

But what about the standard errors? Maximum likelihood provides a way of getting asymptotic standard errors of parameter estimates. Let θ be an r x 1 vector of r parameters to be estimated. Then $\partial \ln \mathcal{L}(\theta)/\partial\theta$ is an r x 1 vector or first-order partial derivatives of ln $\mathcal{L}$ with respect to each of the parameters to be estimated (called the "score vector"). Furthermore, let $\partial^2 \ln \mathcal{L}(\theta)/\partial\theta\partial\theta'$ = an r x r matrix of second-order partial derivatives of the log likelihood function with respect to each parameter (called the "Hessian matrix"). The negative of the expected value of this matrix is called the "information matrix" in maximum likelihood theory:

$$\mathbf{I}(\theta) = - E[\partial^2 \ln \mathcal{L}(\theta)/\partial\theta\partial\theta']$$

where $\mathbf{I}(\theta)$ is the information matrix (*not* an identity matrix!). Moreover, in large samples, the inverse of the information matrix $\mathbf{I}(\theta)$ is the covariance matrix of the maximum likelihood estimates:

$$ACOV(\theta_{ML}) = \mathbf{I}(\theta)^{-1} = - E[\partial^2 \ln \mathcal{L}(\theta)/\partial\theta\partial\theta']^{-1}$$

The asymptotic standard errors of the ML estimator θ_{ML} are the square roots of the diagonal elements of $ACOV(\theta_{ML})$. When applied to our univariate example, we would take the first order partial derivatives, and then differentiate each with respect to μ and σ^2:

$$\begin{bmatrix} \partial^2 \ln \mathcal{L}/\partial\mu^2 & \partial^2 \ln \mathcal{L}/\partial\mu\partial\sigma^2 \\ \partial^2 \ln \mathcal{L}/\partial\sigma^2\partial\mu & \partial^2 \ln \mathcal{L}/\partial(\sigma^2)^2 \end{bmatrix} = \begin{bmatrix} -n/\sigma^2 & -1/\sigma^4 \sum_{i=1}^n (x_i - \mu) \\ -1/\sigma^4 \sum_{i=1}^n (x_i - \mu) & n/2\sigma^4 - 1/\sigma^6 \sum_{i=1}^n (x_i - \mu)^2 \end{bmatrix}$$

For example to get the first diagonal element, take the first partial derivative of $\mathcal{L}$ with respect to μ, $\partial \ln \mathcal{L}/\partial\mu$, and differentiate it again with respect to μ:

$$\partial \ln \mathcal{L}/\partial\mu = 1/\sigma^2 \sum_{i=1}^n (x_i - \mu) = 1/\sigma^2 (\sum_{i=1}^n x_i) - 1/\sigma^2 (\sum_{i=1}^n \mu),$$

$$\partial^2 \ln \mathcal{L}/\partial\mu^2 = - 1/\sigma^2 (\sum_{i=1}^n 1) = - 1/\sigma^2 \, n = - n/\sigma^2$$

Aside: Note that the negative second partial derivative means we've identified a maximum solution to our equation.

For the second diagonal element, take the first partial derivative of $\mathcal{L}$ with respect to σ^2, $\partial \ln \mathcal{L}/\partial\sigma^2$, and differentiate it again with respect to σ^2.

$$\partial \ln \mathcal{L}/\partial\sigma^2 = - (n/2)(\sigma^2)^{-1} + (1/2)(\sigma^2)^{-2} \sum_{i=1}^n (x_i - \mu)^2$$

$$\partial^2 \ln \mathcal{L}/\partial(\sigma^2)^2 = (-1)(-n/2)(\sigma^2)^{-2} + (-2)(1/2)(\sigma^2)^{-3} \sum_{i=1}^n (x_i - \mu)^2$$

$$= n/2(\sigma^{-4}) - \sigma^{-6} \sum_{i=1}^n (x_i - \mu)^2 \qquad \text{(this is shown above; but we can simplify this)}$$

$$= n/2\sigma^4 - \sigma^{-6} \sum_{i=1}^n (x_i - \mu)^2 \qquad \text{but } \sum_{i=1}^n (x_i - \mu)^2 = n\,\sigma^2 \text{ because } (x_i - \mu)^2 = \sigma^2 \text{ and there are n of them}$$

$$= n/2\sigma^4 - n\sigma^2/\sigma^6 \quad = n/2\sigma^4 - n/\sigma^4$$

$$= n/2\sigma^4 - 2n\sigma/2\sigma^4 \quad = -n/2\sigma^4$$

Aside: Note again that the negative second partial derivative means we've identified a maximum solution to our equation.

Now take *minus* the expected value of the above matrix to obtain the information matrix. The off-diagonals are zero because $E(x_i - \mu) = 0$. The first diagonal element is $-E(-n/\sigma^2) = n/\sigma^2$, which is a constant. The second diagonal element is also nonstochastic, so $-E(-n/2\sigma^4) = n/2\sigma^4$

Collecting terms we have:

$$\mathbf{I}(\theta) = -E[\partial^2 \ln \mathcal{L}(\theta)/\partial\theta\partial\theta'] = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{bmatrix}$$

Invert the information matrix and obtain the asymptotic covariance matrix:

$$ACOV(\theta_{ML}) = \mathbf{I}(\theta)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}$$

So, the asymptotic standard error of $\hat{\mu}_{ML}$ is $\sigma/\sqrt{n}$, and of $\hat{\sigma}^2_{ML}$ is $\sigma^2\sqrt{2}/\sqrt{n}$. Note also that the zero off-diagonals indicates that the mean and variance are uncorrelated.

> Aside: Note that these are asymptotic standard errors, which are unbiased in large samples, but biased in small samples. Standard errors that are unbiased in finite samples use n - 1 in the denominator: $\sigma/\sqrt{(n-1)}$ and $\sigma^2\sqrt{2}/\sqrt{(n-1)}$

B. ML ESTIMATION OF A BIVARIATE REGRESSION MODEL.

We can apply the same logic and four steps to a bivariate regression model:

$$Y = \beta X + \epsilon$$

1. Probability Density Function:

Begin with the density function for one observation:

$$f(y \mid x\beta, \sigma_\epsilon^2) = [1/(2\pi\sigma_\epsilon^2)]^{1/2} \exp\{[-1/(2\sigma_\epsilon^2)](y - \beta x)^2\}$$

Then define the joint density function for n independent randomly sampled observations:

$$f(y_1, y_2, \ldots y_n \mid x_i \beta, \sigma_\epsilon^2) = [1/(2\pi\sigma_\epsilon^2)]^{n/2} \exp\{[-1/(2\sigma_\epsilon^2)] \textstyle\sum_{i=1}^{n}(y_i - \beta x_i)^2\}$$

2. Likelihood Function:

This can be used to define the likelihood function of the sample:

$$\mathcal{L}(x_i \beta, \sigma_\epsilon^2 \mid y_1, y_2, \ldots y_n) = [1/(2\pi\sigma_\epsilon^2)]^{n/2} \exp\{[-1/(2\sigma_\epsilon^2)] \textstyle\sum_{i=1}^{n}(y_i - \beta x_i)^2\}$$

Take the log of $\mathcal{L}$:

$$\ln \mathcal{L}(x_i \beta, \sigma_\epsilon^2 \mid y_1, y_2, \ldots y_n) = -n/2 \ln 2\pi - n/2 \ln \sigma_\epsilon^2 - 1/(2\sigma_\epsilon^2) \textstyle\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

3. Maximize Log-Likelihood:

To find values of β and σ_ϵ^2 that maximize the likelihood of the sample, we can maximize $\ln \mathcal{L}$ by differentiating with respect to β and σ_ϵ^2, set the equations equal to zero, and solve for the parameters:

$$\ln \mathcal{L} = -n/2 \ln 2\pi - n/2 \ln \sigma_\epsilon^2 - 1/(2\sigma_\epsilon^2) \textstyle\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

$\partial(\ln \mathcal{L})/\partial\beta = -1/(2\sigma_\epsilon^2) \sum_{i=1}^n 2(y_i - \beta x_i)(-x_i) = -1/\sigma_\epsilon^2 \sum_{i=1}^n (y_i - \beta x_i)(-x_i) = 1/\sigma_\epsilon^2 \sum_{i=1}^n x_i y_i - \beta x_i^2$

$1/\sigma_\epsilon^2 \sum_{i=1}^n x_i y_i - \beta x_i^2 = 0$

$\sum_{i=1}^n x_i y_i - \beta x_i^2 = 0$

$\sum_{i=1}^n x_i y_i - \beta \sum_{i=1}^n x_i^2 = 0$

$\sum_{i=1}^n x_i y_i = \beta \sum_{i=1}^n x_i^2$

$\sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2 = \beta$

$\beta = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2 = s_{xy}/s_x^2$, which is the sample analog to σ_{xy}/σ_x^2

Aside: by maximizing $\ln \mathcal{L}$, we are minimizing $+1/(2\sigma_\epsilon^2) \sum_{i=1}^n (y_i - \beta x_i)^2$. But since $\sum_{i=1}^n (y_i - \beta x_i)^2 = \sum_{i=1}^n \epsilon_i^2$, we are minimizing the sum of squared residuals, which is the OLS method. In the case of regression on observables, OLS and ML give identical results.

4. Asymptotic Covariance Matrix:

What about the standard error of β? Let's take the second-order partial derivatives of $\ln \mathcal{L}$ with respect to β:

$\partial(\ln \mathcal{L})/\partial\beta = 1/\sigma_\epsilon^2 \sum_{i=1}^n x_i y_i - \beta x_i^2 = 1/\sigma_\epsilon^2 \sum_{i=1}^n x_i y_i - \beta/\sigma_\epsilon^2 \sum_{i=1}^n x_i^2$ Here's the first derivative.

$\partial^2(\ln \mathcal{L})/\partial\beta^2 = -1/\sigma_\epsilon^2 \sum_{i=1}^n x_i^2$

This corresponds to the 1,1 element of the 2 x 2 matrix $\partial^2 \ln \mathcal{L}(\theta)/\partial\theta\partial\theta'$. To get the information matrix, we take one minus the expected value of the second partial derivative:

$-E[\partial^2(\ln \mathcal{L})/\partial\beta^2] = -E(-1/\sigma_\epsilon^2 \sum_{i=1}^n x_i^2) = 1/\sigma_\epsilon^2 \sum_{i=1}^n x_i^2$ since each term is nonstochastic

Then the asymptotic variance of β is

$AVAR(\beta) = 1/\{-E[\partial^2(\ln \mathcal{L})/\partial\beta^2]\} = 1/(1/\sigma_\epsilon^2 \sum_{i=1}^n x_i^2) = \sigma_\epsilon^2/\sum_{i=1}^n x_i^2$

This is the 1,1 element of $ACOV(\theta_{ML}) = I(\theta_{ML})^{-1}$. The standard error of $\beta = (\sigma_\epsilon^2/\sum_{i=1}^n x_i^2)^{1/2}$.

Theil (1971, p. 391) shows that the ML estimate of $\sigma_\epsilon = \epsilon_i^2/n$ and inverse of the information matrix $\mathbf{I}(\theta_{ML})^{-1}$ is:

$$ACOV(\theta_{ML}) = \begin{bmatrix} \sigma_\epsilon^2/\sum_{i=1}^n x_i^2 & 0 \\ 0 & 2\sigma_\epsilon^4/n \end{bmatrix}$$

where $\theta_{ML}' = [\beta \quad \sigma_\epsilon^2]$

IV. ML ESTIMATION IN COVARIANCE STRUCTURE MODELS.

A. MAXIMIZING THE LIKELIHOOD FUNCTION

There are two ways of deriving the likelihood function for covariance structure analysis: (1) begin with the assumption that xs and ys are distributed multivariate normal; or (2) begin with the assumption that the sample

covariance matrix S follows a Wishart distribution. We'll follow the former (for the latter, see Bollen 1989, pp. 134-5 or Hayduk 1987, pp. 133-8.

1. Probability Density Function:
Begin by assuming the joint distribution of p ys and q xs is multivariate normal. As before, we have a (p + q) x 1 vector **Z**, such that **Z'** = **[Y' X']** = $[y_1, y_2 \ldots y_p; x_1, x_2 \ldots x_q]$, and the covariance matrix of **Z**, E(**ZZ'**) is a (p + q) x (p + q) matrix Σ. The multivariate density function of one observation of **Z** is:

$$f[\mathbf{Z} \mid \Sigma(\theta)] = 2\pi^{-(p+q)/2} [\mid \Sigma(\theta) \mid]^{-1/2} \exp [-\tfrac{1}{2} \mathbf{Z}' \Sigma(\theta)^{-1} \mathbf{Z}]$$

where $\Sigma(\theta)$ refers to the parameters of our covariance structure model placed in one long vector. For a random sample of n independent observations, the joint density function is the product of the individual observations densities:

$$f[\mathbf{Z}_1, \mathbf{Z}_2 \ldots \mathbf{Z}_n \mid \Sigma(\theta)] = 2\pi^{-n(p+q)/2} [\mid \Sigma(\theta) \mid]^{-n/2} \exp [-\tfrac{1}{2} \textstyle\sum_{i=1}^{n} \mathbf{Z}_i' \Sigma(\theta)^{-1} \mathbf{Z}_i]$$

2. Likelihood Function:
This joint density function is mathematically equivalent to the likelihood function, expressing the likelihood of the sample given we've observed the sample values of our variables:

$$\mathcal{L}[\Sigma(\theta) \mid \mathbf{Z}_1, \mathbf{Z}_2 \ldots \mathbf{Z}_n] = 2\pi^{-n(p+q)/2} [\mid \Sigma(\theta) \mid]^{-n/2} \exp [-\tfrac{1}{2} \textstyle\sum_{i=1}^{n} \mathbf{Z}_i' \Sigma(\theta)^{-1} \mathbf{Z}_i]$$

Given the sample, we want to find values of the parameters $\Sigma(\theta)$ that maximize the likelihood of the sample. Let's simplify by using $\mathcal{L}$ to refer to the left-hand side above. It'll be easier to maximize the log-likelihood:

$$\ln\mathcal{L} = n(p+q)/2 \ln(2\pi) - n/2 \ln \mid \Sigma(\theta) \mid - (\tfrac{1}{2}) \textstyle\sum_{i=1}^{n} \mathbf{Z}_i' \Sigma(\theta)^{-1} \mathbf{Z}_i$$

The last term (a scalar) can be rewritten:

$- (\tfrac{1}{2}) \sum_{i=1}^{n} \mathbf{Z}_i' \Sigma(\theta)^{-1} \mathbf{Z}_i = - (\tfrac{1}{2}) \sum_{i=1}^{n} tr[\mathbf{Z}_i' \Sigma(\theta)^{-1} \mathbf{Z}_i]$ since the trace of a scalar is the scalar

$= - (n/2) \sum_{i=1}^{n} tr[n^{-1} \mathbf{Z}_i \mathbf{Z}_i' \Sigma(\theta)^{-1}]$ because tr(ABC) = tr(CAB)

$= - (n/2) tr[\mathbf{S}_{ML} \Sigma(\theta)^{-1}]$ since $\mathbf{S}_{ML} = n^{-1} \mathbf{Z}_i \mathbf{Z}_i'$ where $\mathbf{S}_{ML}$ is the biased ML estimator of the covariance of matrix of **Z**s.

Therefore,

$\ln\mathcal{L} = n(p+q)/2 \ln(2\pi) - n/2 \ln \mid \Sigma(\theta) \mid - (n/2) tr[\mathbf{S}_{ML} \Sigma(\theta)^{-1}]$ or rearranging,

$= - n/2 \{\ln \mid \Sigma(\theta) \mid + tr[\mathbf{S}_{ML} \Sigma(\theta)^{-1}]\} + n(p+q)/2 \ln(2\pi)$

Therefore, maximizing ln $\mathcal{L}$ is equivalent to minimizing F = -ln $\mathcal{L}$, where F is the fitting function:

$$F = n/2 \{\ln \mid \Sigma(\theta) \mid + tr[\mathbf{S}_{ML} \Sigma(\theta)^{-1}]\} + n(p+q)/2 \ln(2\pi)$$

Because n, 2, p, q, and π are not functions of parameters (are constant once the sample is drawn), they will not affect the choice of θ. Therefore, we can simplify the fitting function:

$$F = \ln \mid \Sigma(\theta) \mid + tr[\mathbf{S}_{ML} \Sigma(\theta)^{-1}]$$

The function minimized by the LISREL program (Jöreskog and Sörbom 1989, p. 21) is derived from a Wishart distribution for S (see Hayduk 1987, pp. 136-8):

$$F = \ln|\Sigma(\theta)| + tr[\mathbf{S}\,\Sigma(\theta)^{-1}] - \ln|\mathbf{S}| - (p + q)$$

There are two differences between LISREL's fitting function and that derived from the likelihood function of a multivariate normal distribution: the last two terms and **S**, the sample covariance matrix. Given the sample, the terms, $\ln|\mathbf{S}|$ and $(p + q)$, are constant, and therefore will not affect the choice of θ. The sample covariance matrix in maximum likelihood theory, $\mathbf{S}_{ML}$, uses n in the denominator, whereas, Jöreskog and Sörbom begin with the Wishart distribution for **S**, the unbiased estimator of Σ, using n - 1 in the denominator of each element s_{ij}. Asymptotically, n converges to n - 1 and therefore, $\mathbf{S}_{ML}$ converges to **S**. From here on, we'll rely on Jöreskog and Sörbom's fitting function.

For linear systems of equations, minimizing F gives asymptotically optimal estimates for a wide range of applications (and for some classes of models, finite sample properties are known). Most linear models become special cases. For example:

1. In a single-equation model in observables, minimizing F gives OLS estimates (which are unbiased and efficient).

2. In recursive models in observables, minimizing F gives equation-by-equation OLS estimates.

3. In seemingly-unrelated regression equations -- two or more equations that have different regressors and are related by correlated disturbances, minimizing F gives estimates that are asymptotically equivalent to Zellner's GLS estimator (consistent and asymptotically efficient).

4. In non-recursive models (simultaneous equations), minimizing F gives estimates that are asymptotically equivalent to 3SLS estimates (consistent and asymptotically efficient). Note the gain in asymptotic efficiency due to applying Zellner's GLS to 2SLS estimates.

5. In factor analysis models, in which there are only xs, ξs, and δs, minimizing F is equivalent to ML factor analysis (Lawley and Maxwell 1967).

6. In multi-equation models of latent variables with multiple indicators, minimizing F gives consistent and asymptotically efficient estimators. Thus, for overidentified models, they provide a gain in asymptotic efficiency over the method of moments.

3. Maximize Log-Likelihood:
To maximize the fitting function F, we begin by taking the partial derivatives of F with respect to each parameter, set the equation equal to zero and solve for the parameters. The derivatives have been worked out but they are very complicated, and therefore, not too revealing for our purposes. For the curious, here's what they look like (see Jöreskog 1973, 1977, or Hayduk 1987). First, a couple of simplifications. If we take $\mathbf{S} - \Sigma(\theta)$ to be the residual matrix, we can define Ω, a partitioned matrix with the same dimensions as $\Sigma(\theta)$ and **S**,

$$\Omega = \Sigma(\theta)^{-1}\,[\Sigma(\theta) - \mathbf{S}]\,\Sigma(\theta)^{-1} = \begin{bmatrix} \Omega_{yy} & \Omega_{yx} \\ \Omega_{xy} & \Omega_{xx} \end{bmatrix}$$

Also define the covariance matrix of η as $\mathbf{C} = E(\eta\eta') = (\mathbf{I} - \mathbf{B})^{-1}\,(\Gamma\,\Phi\,\Gamma + \Psi)(\mathbf{I} - \mathbf{B})^{-1}$, then the partial derivatives of the fitting function with respect to each free parameter is:

$$\partial F/\partial\Theta_{\epsilon} = \Omega_{yy}$$

$$\partial F/\partial\Theta_{\delta} = \Omega_{xx}$$

$$\partial F/\partial\Lambda_{y} = \Omega_{yy}\,\Lambda_{y}\,\mathbf{C} + \Omega_{xy}'\Lambda_{x}\,\Phi\,\Gamma'(\mathbf{I} - \mathbf{B})^{-1'}$$

$\partial F/\partial \Lambda_x = \Omega_{yy} \Lambda_y (\mathbf{I} - \mathbf{B})^{-1} \Phi \Gamma' + \Omega_{xx}' \Lambda_x \Phi$

$\partial F/\partial \mathbf{B} = -(\mathbf{I} - \mathbf{B})^{-1'} \Lambda_y' (\Omega_{yy} \Lambda_y \mathbf{C} + \Omega_{xy}' \Lambda_x \Phi \Gamma' (\mathbf{I} - \mathbf{B})^{-1'})$

$\partial F/\partial \Gamma = -(\mathbf{I} - \mathbf{B})^{-1'} \Lambda_y' [\Omega_{yy} \Lambda_y (\mathbf{I} - \mathbf{B})^{-1} \Gamma + \Omega_{xy}' \Lambda_x] \Phi$

$\partial F/\partial \Phi = \Gamma' (\mathbf{I} - \mathbf{B})^{-1'} \Lambda_y' \Omega_{yy} \Lambda_y (\mathbf{I} - \mathbf{B})^{-1} \Gamma + \Lambda_x' \Omega_{xy} \Lambda_y (\mathbf{I} - \mathbf{B})^{-1} \Gamma + \Gamma' (\mathbf{I} - \mathbf{B})^{-1'} \Lambda_y' \Omega_{xy}' \Lambda_x + \Lambda_x' \Omega_{xx}' \Lambda_x$

$\partial F/\partial \Psi = (\mathbf{I} - \mathbf{B})^{-1'} \Lambda_y' \Omega_{yy} \Lambda_y (\mathbf{I} - \mathbf{B})^{-1'}$

In the example of maximum likelihood estimation of the mean, variance, and regression, the solutions for parameter values that maximize the likelihood function exist as closed expressions. We can also obtain closed expressions for some special cases of the general covariance structure model -- e.g., the first four submodels noted above. For example, in a regression model, minimizing the sum of squared residuals gives the maximized value of the likelihood function. This can be done because the fitting function is a linear function of parameters; therefore, there is only one slope, which is given by first-order derivatives. In general, however, the fitting function of the covariance structure model is a very complicated nonlinear function of the parameters, and no closed expression exists that would allow us to express the maximized fitting function in terms of parameter values. Therefore, in empirical applications, we have to use a numerical method of maximizing the fitting function. The way this works, is we begin with some starting values for our parameter vector θ. Call the starting value $\theta^{(1)}$. We then compute the value of F_{ML}, the fitting function. Then we move to a new value of θ, $\theta^{(2)}$, and recompute F_{ML}. If this value is smaller than the previous value, we continue to modify θ in this direction. The next value we try $\theta^{(3)}$, then, should look more like $\theta^{(2)}$ than $\theta^{(1)}$. We again compute F_{ML}, and modify our choice of θ accordingly. Two critical questions arise: (1) How do we determine the movement from one $\theta^{(i)}$ to the next $\theta^{(i+1)}$? (2) How do we know when to stop?

Typically, the movement between trial values follows the following equation:

$$\theta^{(i+1)} = \theta^{(i)} - \mathbf{C}^{(i)} \mathbf{g}^{(i)}$$

where $\mathbf{g}^{(i)} = \partial F_{ML}/\partial \theta$ is a vector of partial derivatives of the likelihood function with respect to each parameter estimate evaluated at $\theta^{(i)}$, and $\mathbf{C}^{(i)}$ is some positive definite matrix (see Bollen 1989, pp. 136-43). (Note that the superscript *i* indexes the iterations.) The choice of $\mathbf{C}^{(i)}$ determines the iteration method:

	$\mathbf{C}^{(i)}$	Method	Comment
1.	Identity Matrix	Steepest Descent	Very slow.
2.	$[\partial^2 F_{ML}/\partial\theta\partial\theta']^{-1}$	Newton-Raphson	Computation-intensive, but takes few iterations.
3.	Above, modified	Davidson-Fletcher-Powell	Faster and inexpensive.
4.	$\mathbf{C}^{(i)} = \mathbf{C}^{(i+1)}$	$\mathbf{C}^{(i)}$ never changes	Very fast, but takes many iterations.

The Fletcher-Powell procedure is the default for LISREL 8. The Newton-Raphson method requires heavy computations because it requires computing the inverse of the information matrix at each iteration. The Fletcher-Powell method tries to build up this matrix over iterations. See page 311 in the LISREL 8 Manual for a discussion of options on the iteration procedure. Typically one would not need to change the method. (EQS uses something similar, called a modified Gauss-Newton method.)

The iteration procedure ends when a convergence criterion is met. This criterion specifies that changes in each and every parameter would not change the fitting function substantially. The first-order derivatives of the fitting function with regard to each parameter gives this information. If a derivative of a given parameter is large, it implies that changing the parameter would change the fitting function substantially, and therefore the fitting

function has not reached a minimal value. Conversely, if all derivatives are small, it implies that a minimum has been reached. LISREL 8 uses the following convergence criterion:

$|\partial F_{ML}/\partial\theta_i| < \text{EPS}$ if $|\theta_i| \leq 1$ and

$|(\partial F_{ML}/\partial\theta_i)/\theta_i| < \text{EPS}$ if $|\theta_i| > 1$

and uses as a default value EPS = .000001, which typically means that the solution is accurate to three digits. Also, LISREL 8 limits the number of iterations, *i*, to three times the number of free parameters to be estimated. This can be modified by the IT parameter on the OU card.

Aside: The LISREL and EQS programs offers two other fit functions:

1. $F_{GLS} = (\frac{1}{2})\ \text{tr}\ (\ \{[(\mathbf{S} - \Sigma(\theta)]\ \mathbf{W}^{-1}\}^2\ )$

where W^{-1} is a weight matrix such that $\underset{n\to\infty}{\text{plim}}\ W^{-1} = \Sigma(\theta)$

Depending on the selection of $\mathbf{W}^{-1}$, this fitting function can provide estimates with optimal asymptotic properties when the observed variables depart from normality, but do not have excessive kurtosis. Although a variety of weight matrices are possible, LISREL and EQS use $\mathbf{W}^{-1} = \mathbf{S}^{-1}$ as a weight matrix which yields what they call GLS (generalized least squares estimates). Note that when $\mathbf{W}^{-1} = \Sigma(\theta_{ML})$, then minimization of F_{GLS} gives maximum likelihood estimates.

2. $F_{ULS} = (\frac{1}{2})\ \text{tr}\{[(\mathbf{S} - \Sigma(\theta)]^2\}$

This is a special case of F_{GLS} where $\mathbf{W}^{-1} = \mathbf{I}$, an identity matrix (thus the term, "unweighted least squares"). Relative to F_{ML}, F_{ULS} gives more weight to the covariances than variances, since the covariances appear twice in $\text{tr}\{[(\mathbf{S} - \Sigma(\theta)]^2\}$. ULS estimates are consistent but asymptotically inefficient. Later, when we discuss asymptotic distribution-free estimators, we will present a more general fit function, in which each of the above fit functions are special cases.

B. INFORMATION MATRIX AND LIKELIHOOD RATIO TEST.

We can define the information matrix for the LISREL model. Again, let θ be an r x 1 vector of r parameters (λ_x, λ_y, σ_δ^2, σ_ϵ^2, σ_ξ^2, γ, β, σ_ζ^2) to be estimated. Then $\partial \ln \mathcal{L}/\partial\theta$ is an r x 1 vector or first-order partial derivatives of $\ln \mathcal{L}$ with respect to each of the parameters to be estimated (where $\mathcal{L} = \mathcal{L}[\Sigma(\theta)|z_1, z_2 \ldots z_n]$). Furthermore, let $\partial^2 \ln \mathcal{L}/\partial\theta\partial\theta' =$ an r x r matrix of second-order partial derivatives of the log likelihood function with respect to each parameter. The negative of the expected value of this matrix is called the "information matrix" in maximum likelihood theory:

$$\mathbf{I}(\theta) = -\ E[\partial^2 \ln \mathcal{L}(\theta)/\partial\theta\partial\theta']$$

Moreover, in large samples, the inverse of the information matrix $I(\theta)$ is the covariance matrix of the maximum likelihood estimates:

$$\text{ACOV}(\theta_{ML}) = \mathbf{I}(\theta)^{-1} = -\ E[\partial^2 \ln \mathcal{L}(\theta)/\partial\theta\partial\theta']^{-1}$$

This is an r x r matrix, which corresponds to the ordering of parameters in θ (r x 1), of covariances of parameter estimates. The asymptotic standard errors of the ML estimator θ_{ML} are the square roots of the diagonal elements of $\text{ACOV}(\theta_{ML})$. And the parameter estimate divided by the standard error has a z-distribution (asymptotic t-distribution) with one degree of freedom:

$$\theta_{ML}/[VAR(\theta_{ML})]^{1/2} \sim Z_{(1)}$$

Thus, we can perform tests of hypotheses about parameter estimates in the usual way.

> Aside: Note that if a model is underidentified, at least two of it's parameters are a linear combination of each other, which means that $I(\theta)$ is not singular, has a zero determinant, and cannot be inverted. This is the basis of the empirical identification check in LISREL 7. If $I(\theta)$ cannot be inverted to obtain the asymptotic covariance matrix of estimates, the program prints out an error message.

We can also define the likelihood ratio statistic. Suppose we want to test a models overidentifying restrictions. Let Model H be the hypothesized model we want to test, and let Model A be the alternative model within which Model H is nested. We obtain Model H by constraining k parameters of Model A. Let $\Sigma_H(\theta)$ be the implied covariance matrix of Model H, and F_H be its fitting function. Then let $F_H(\theta_{ML})$ be the values of the *minimized* fitting function for the model:

$$F_H(\theta_{ML}) = \ln|\Sigma_H(\theta_{ML})| + tr[S\,\Sigma_H(\theta_{ML})^{-1}] - \ln|S| - (p+q)$$

$F_H(\theta_{ML})$ can be obtained from F_H by replacing $\Sigma_H(\theta)$ with $\Sigma_H(\theta_{ML})$. Let $\Sigma_A(\theta)$ be the implied covariance matrix for the alternative less-restrictive Model A. Then let F_A be the values of the *minimized* fitting function for Model A:

$$F_A(\theta_{ML}) = \ln|\Sigma_A(\theta_{ML})| + tr[S\,\Sigma_A(\theta_{ML})^{-1}] - \ln|S| - (p+q)$$

Then -2 times the likelihood ratio statistic is defined as:

$$\nu = -2 \ln \lambda = n - 1\ [F_H(\theta_{ML}) - F_A(\theta_{ML})]$$

Moreover, ν is asymptotically distributed as a χ^2 variate with degrees of freedom equal to r, the number of constraints on Model A (or, in other words, the difference in numbers of parameters to be estimated between Models H and A). The LISREL program calls ν "chi-square." Thus, to test any nested hypothesis, simply run the less-constrained model as well as the constrained model, and subtract the χ^2 of the former from the χ^2 of the latter and find (from tabled probability values of the χ^2) if it is statistically significant. If so, one can reject the null hypothesis of the constrained model.

We can apply the likelihood ratio method to test the overall goodness-of-fit of the model -- that is, the simultaneous test of all of the model's overidentifying restrictions against some (unspecified) just-identified alternative. Since the alternative model is just-identified, it reproduces the *sample* moment matrix perfectly, $\Sigma_A(\theta_{ML}) = S$

$$F_A(\theta_{ML}) = \ln|S| + tr[S\,S^{-1}] - \ln|S| - (p+q)$$

$$= \ln|S| - \ln|S| + tr(\mathbf{I}) - (p+q) \qquad \text{where } \mathbf{I} = (p+q)x(p+q) \text{ identity matrix}$$

$$= 0 \qquad \text{since } tr(\mathbf{I}) = p+q$$

Therefore, the test of the overall goodness-of-fit of the model is

$$\nu = n - 1\ [F_H(\theta_{ML}) - F_A(\theta_{ML})] = n - 1\ [F_H(\theta_{ML}) - 0]$$

$$\nu = n - 1\ \{\ln|\Sigma_H(\theta_{ML})| + tr[S\,\Sigma_H(\theta_{ML})^{-1}] - \ln|S| - (p+q)\}$$

with degrees of freedom (r) equal to the number of moments minus parameters (t) estimated by the model[(k)(k + 1)/2] - t, where k = p + q, the number of observable variables. LISREL 7 and EQS print out this statistic for every model. Again, remember it is giving a simultaneous test of all of the overidentifying restrictions in the model. Note

that if a model (the moment matrix implied by the estimates) fits the sample data perfectly, $\Sigma_H(\theta_{ML}) = \mathbf{S}$ and $F_H(\theta_{ML}) = 0$ (which is very very unlikely!), then:

$$\nu = n - 1 \{\ln|\mathbf{S}| + \text{tr}[\mathbf{S}(\mathbf{S})^{-1}] - \ln|\mathbf{S}| - (p + q)\}$$

$$= n - 1 [\ln|\mathbf{S}| - \ln|\mathbf{S}| + \text{tr}(\mathbf{I}) - (p + q)] \qquad \text{where } \mathbf{I} = (p+q)\text{x}(p+q) \text{ identity matrix}$$

$$= 0 \qquad \text{since } \text{tr}(\mathbf{I}) = p + q$$

Thus, for a model whose overidentifying restrictions are perfectly satisfied in the sample, the χ^2 is zero. Of course, this is very unlikely to occur, since *even if the model holds exactly in the population*, one would not expect it to reproduce the sample covariance matrix perfectly, because of sampling error. In fact, the expected value of ν is $E(\chi^2) = [(k)(k + 1)/2] - t$, the number of degrees of freedom in the model.

Aside: Note that for nested models differing by one parameter, the one degree of freedom χ^2 for nested models is equal to the squared z-statistic (t-value) differentiating the two models (and found in the less-constrained alternative model), where the z statistic is $\theta_{ML}/[\text{VAR}(\theta_{ML})]^{1/2}$.