Professor Ross L. Matsueda

LECTURE 13: TEST STATISTICS AND MULTIPLE GROUP MODELS

This lecture discusses two miscellaneous topics in covariance structure analysis: alternative test statistics and multiple-sample models. It begins by discussing other test statistics used in covariance structure analysis, beginning with Wald and Lagrangian Multiplier tests, and then considering alternative measures of fit designed in part to offset the dependence of sample size on goodness-of-fit statistic. It then concludes with a brief discussion of multiple group models, which allow one to examine a covariance structure model in different populations, and test hypotheses of interactions or invariance of parameters across populations.

I. ALTERNATIVE TEST STATISTICS.

A. WALD AND LAGRANGIAN MULTIPLIER TESTS.

These two tests are closely affiliated to the likelihood ratio test; however, they provide specific tests that can be constructed with a single computer run. Stated informally, for a given model, the Wald test provides a test of whether a free parameter(s) can be constrained; the Lagrangian multiplier test, in contrast, provides a test of whether a constrained parameter(s) can be freed. Recall that the likelihood ratio test can test either hypothesis, but requires two runs, estimating the more- and less-restricted models, and computing the difference in log-likelihoods. Consider Model H as a restrictive model, with free parameters θ_r and Model A as a less-restrictive alternative, with free parameters θ. Furthermore, consider the log-likelihood of Model H as $F_H(\theta)_{ML}$ and the log-likelihood of Model A as $F_A(\theta)_{ML}$. Then the likelihood ratio test statistic is:

$$\nu = -2 \ln \lambda = n - 1 \, [F_H(\theta)_{ML} - F_A(\theta)_{ML}] \quad \text{and} \quad \operatorname{plim}_{n\to\infty} \nu \sim \chi^2 \quad \text{with k degrees of freedom}$$

The Lagrangian Multiplier (LM) test subjects the restrictions in Model H to a test against Model A, but only estimates Model H. The test is based on the first-order partial derivatives of Model H, $\partial F_H(\theta)_{ML}/\partial\theta$. Depending on the model, the vector $(\theta)_{ML}$ contains both fixed and free parameters. Recall that the first-order partial derivatives give the slope of the likelihood function at current values of parameters; thus it tells us how much the function would change if we changed the value of the parameter by one unit. For the free parameters, the partial derivatives are equal to zero, since the maximum likelihood solution was obtained by setting these partial derivatives equal to zero (and solving for parameters). The partial derivatives of constrained parameters will be close to zero if the model is correct and these constraints hold in the population. But they won't be exactly zero because of sampling variability. The larger the first derivatives depart from zero, the larger is the estimated departure from the constraint of the model, and the less likely the departure is due solely to sampling variability (and therefore the more likely the constraint is invalid). To give a formal test of whether the constraint is true in the population, we need to consider sampling variability in the degree to which the first derivative departs from zero (or the nonzero value to which it was constrained). To do this, we use the estimated covariance matrix of Model H, which recall is:

$$\text{ACOV}(\theta_{ML}) = R(\theta)^{-1} = -\,E\{\partial^2 \ln \mathcal{L}[\Sigma_H(\theta_{ML})]/\partial\theta\partial\theta'\}^{-1} = 2/(N-1)\,E\{\partial^2\,[F_H(\theta)_{ML}]/\partial\theta\partial\theta'\}^{-1}$$

Then, the LM statistic is:

$$LM = (N-1)/2\,[\partial F_H(\theta)_{ML}/\partial\theta]'\,E[\partial^2\,F_H(\theta)_{ML})/\partial\theta\partial\theta']^{-1}\,[\partial F_H(\theta)_{ML}/\partial\theta]$$

This statistic is sometimes expressed as:

$$LM = [s(\theta_H)]' \, R(\theta)^{-1} \, s(\theta_H)$$

where $s(\theta_H) = \partial \ln \mathcal{L}[\Sigma_H(\theta_{ML})/\partial\theta]$ is called the score vector or vector of efficient scores; therefore, the test is sometimes called the "score test" or "efficient score test." We can visualize this by noting that LM takes the first-order partial derivative, squares it, and then divides by the appropriate elements of the asymptotic covariance matrix of the estimates, which weights each derivative (the departure from constraints) by the inverse of a measure of sampling variability $ACOV(\theta_{ML})$. Further insights into the LM statistic can be gained by noting that when Model H and Model A differ by one parameter, LM is a χ^2 variate with one degree of freedom, which is equivalent to a z^2 variate:

$$LM = MI = s(\theta_H)_i^2 / AVAR(\theta_{ML})_i$$

where $s(\theta_H)_i^2$ here is a scalar (the partial derivative of $\partial \ln \mathcal{L}[\Sigma_H(\theta_{ML})/\partial\theta_i]$ with respect to a single parameter θ_i and $AVAR(\theta_{ML})_i$ is (a scalar) the asymptotic variance of θ_i. These one-degree-of-freedom LM tests are called "modification indices" (MI) in the LISREL 8. They tell us the expected drop in χ^2 would be if we relaxed one constraint. LISREL 8 also prints out "expected change" values for each fixed parameter. We can define this approximately. Let θ_0 be the value of a given restricted parameter and let θ_i be the value of the restricted parameter if it were to be estimated. Then, Saris, Satorra, and Sörbom (1987) argue that the expected change in the restricted parameter is approximately:

$\theta_i - \theta_0 = MI/\partial \ln[\mathcal{L}\Sigma_H(\theta_{ML})]/\partial\theta_i$ and since $\partial \ln \mathcal{L}[\Sigma_H(\theta_{ML})]/\partial\theta_i = -(n-1)\partial F_H(\theta)_{ML}/\partial\theta_i$

$\theta_i - \theta_0 = MI/-(n-1)FD(\theta_0)$ where $FD(\theta_0) = \partial F_H(\theta)_{ML}/\partial\theta_0$ = LISREL 8 first-order derivatives

Also, note that the asymptotic standard error of the "expected change," $\theta_i - \theta_0$, is the square root of the ratio of the expected change squared divided by the modification index:

$$[AVAR(\theta_{ML})_i]^{1/2} = [(\theta_i - \theta_0)^2/MI]^{1/2}$$

We could compute multivariate modification indices (Lagrangian Multiplier tests) by taking the asymptotic covariance matrix of the estimator, and pre- and post-multiply it by a vector of first-order derivatives, after setting to zero those derivatives in which we are not interested. This can be done using LISREL 8 by saving the asymptotic covariance matrix using EC = *filename* on the OU line, obtaining first derivatives using FD in the OU line, and then using a matrix program to multiply the quadratic form. An easier route would be to use EQS, which allows one to ask for multivariate LM statistics.

The Wald test (W) subjects a model's free parameters to hypothesis testing. Thus, it answers the conventional hypothesis testing question of whether an estimated parameter is significantly different from the null hypothesis. It requires estimation of the less-restrictive alternative model (Model A) only. Suppose we wanted to test the following hypothesis:

H_0: $r(\theta) = c$ or equivalently, $r(\theta) - c = 0$

H_A: $r(\theta) \neq c$ or equivalently, $r(\theta) - c \neq 0$

where r and c are vectors of constants, and θ is a vector of parameters. Thus, if the restrictions (null hypothesis) are true, then θ should satisfy them within sampling error; if it is false, then $r(\theta) - c$ would be further from zero than we would expect by sampling variability alone. The test for this hypothesis uses the Wald statistic:

$$W = [r(\theta) - c]' \, \{ACOV[r(\theta) - c]\}^{-1} \, [r(\theta) - c]$$

Like the Lagrangian Multiplier statistic, the Wald statistic takes the departure of each parameter estimate from the null hypothesis and weights them by the inverse of the amount of sampling variability in the estimate. Given that ACOV[r(θ) - c] = C ACOV(θ) C', where C = [∂ r(θ)/∂θ'] we can write this as:

$$W = [r(\theta) - c]' \{[\partial r(\theta)/\partial\theta]' \, ACOV(\theta) \, [\partial r(\theta)/\partial\theta]\}^{-1} [r(\theta) - c]$$

A special case would arise for a general linear restriction in which the null hypothesis would be r(θ) - c = gθ - c = 0. Note also that in the one-degree-of-freedom case, W becomes the z-statistic squared. If the null hypothesis is H_0: θ - θ_0 = 0 versus the alternative H_A: θ - $\theta_0 \neq 0$, then

$$W = (\theta - \theta_0) \, ACOV(\theta - \theta_0)^{-1} (\theta - \theta_0)$$

$$= (\theta - \theta_0)^2 \, AVAR(\theta)$$

$$= Z^2$$

When the null hypothesis is zero, the Wald statistic is equal to the squared parameter estimate divided by the asymptotic variance of that estimate. Thus, what LISREL prints out as the asymptotic "t-value" is actually a special case of the Wald statistic.

The Wald, Lagrangian Multiplier, and likelihood-ratio test statistics all have the same asymptotic distribution. Thus, as the sample size approaches infinity, the three test statistics converge on the same value. The finite distributions of the test statistics, however, are not identical. Thus, one test may yield more statistical power than others. The finite-sample properties of the tests will also vary depending on the specific kind of model estimated. In general, then, the test statistics will not yield the identical results in finite samples. The choice of statistic is usually predicated on practical grounds -- that is, whichever test is easiest to compute is typically used. For example, the likelihood ratio test statistic is readily available in statistical packages, but requires estimating both restrictive and alternative models (except when the alternative model is just-identified). The Lagrangian Multiplier test requires estimating only the more restrictive model; the Wald test requires estimating only the nonrestrictive (alternative) model. If each test is available, then one might use the LM test when the baseline model (the one being reported) is the restrictive model, and the Wald test when the baseline model is the nonrestrictive model. The LISREL program provides the likelihood ratio test and one-degree-of-freedom Lagrangian Multiplier and Wald tests. The EQS program also allows users to specify multivariate Lagrangian Multiplier and Wald tests.

B. ALTERNATIVE FIT INDEXES.

Because the likelihood-ratio (and Wald and Lagrangian Multiplier) tests are sensitive to the sample size, several researchers have developed alternative fit indexes and test statistics. The problem is that for a given population model, the overall goodness-of-fit χ^2 test will be more likely to reject the model in large samples than in small samples. This is because the test will have more power to detect departures from overidentifying restrictions. In the extreme case, with a huge sample size, the test will reject models that have substantively trivial departures from restrictions. Using classical principles of statistical inference, the formally correct way to address this problem is to compute the power of the test and protect against a desired level of both type I and type II errors. But for large models with many overidentifying restrictions, this becomes impractical. Unfortunately, most fit indices do not directly resolve the problem because they factor out N (the sample size) but fail to control for the amount of sampling variability (which is determined in part by the sample size). That is, in small samples, given a true model, we would still expect larger discrepancies from restrictions [S - $\Sigma(\theta_H)$] because we have greater sampling variability. Conversely, in large samples, we would expect smaller discrepancies from restrictions because of smaller sampling variability. Simply offsetting n in the test statistic does not take this into consideration.

Here are some of the more popular indices.

1. χ^2/df

Some researchers argue for the use of χ^2/df, since one would expect larger χ^2 values for larger degrees of freedom. While this is true, the degrees of freedom are taken into consideration in the critical values of the χ^2_α, and it doesn't take sample size into account.

2. Standardized χ^2

Since the expected value of the χ^2 variate is the number of degrees of freedom, some have proposed χ^2/df as an index that would be independent of sample size. Others suggest standardizing by subtracting the expected value and dividing by the standard deviation of the χ^2, which is $(2df)^{1/2}$:

$$\text{Standardized } \chi^2 = (\chi^2 - df)/(2df)^{1/2}$$

Unfortunately, this statistic is just as dependent on sample size (and sampling variability) as the χ^2. And it's statistical distribution is not known.

3. Adjusted Goodness-of-Fit Index (AGFI).

This is Jöreskog and Sörbom's index, available in the LISREL 8 program. The GFI indexes the relative amount of S that is predicted by $\Sigma\theta$:

$$GFI = 1 - tr\{[\Sigma_H(\theta_{ML})^{-1} S - I]^2\}/tr\{[\Sigma_H(\theta_{ML})^{-1} S]^2\}$$

To adjust for the number of degrees of freedom, they define an Adjusted Goodness-of-Fit Index:

$$AGFI = 1 - [(p + q)(p + q + 1)/2df][1 - GFI]$$

While this is a useful relative fit index, it also is dependent on sampling variability (and thus sample size) and has an unknown distribution as a test statistic.

4. Bentler and Bonett's Incremental Fit Index.

Bentler and Bonett (1980) borrow from Tucker and Lewis (1973) and present an "incremental fit index." Here's the logic. Normally, in using the overall goodness-of-fit statistic, we are comparing our hypothesized model to a just-identified alternative. In large samples, we will likely reject our hypothesized model because our model pales in comparison to a perfectly fitting-model. But, even though our model is not significantly better than a "perfect" model, it may still be accounting for some nontrivial proportion of the observed covariance matrix S. Therefore, perhaps we should compare it to a very imperfect model -- call it the "null model." The "null model" should be the most restrictive, theoretically defensible model. They follow Tucker and Lewis (1973) and present as a generic null model one that leaves the observed variances free, but constrains all observed covariances to be zero. (Thus, all variables have variances estimated as sample variances, but zero intercorrelations.) If one's hypothesized model fits better than this null model, then perhaps it is capturing something important and shouldn't be rejected. Formally, their normed fit index is:

$$\Delta_{HA} = (F_H - F_A)/F_B \quad = \quad (\chi^2_H - \chi^2_A)/\chi^2_B$$

where F_H, F_A, and F_B refer to minimized values of the fitting function for the hypothesized model, the alternative model, and the baseline (or null) model, which are each equivalent to the overall goodness-of-fit χ^2 variate. Because $0 \le (F_H - F_A) \le F_B$, Δ_{HA} is bounded by zero and one. This index assesses the usual discrepancy between hypothesized and alternative models $\chi^2_H - \chi^2_A$ as a proportion of χ^2_B (the test of the discrepancy between **S** and $\Sigma_B(\theta)$. Stated crudely, if your favorite model doesn't fit relative to a *perfect* model, perhaps it'll fit compared to a

terrible model, i.e., the null model. Note that if the alternative model is the usual just-identified (perfectly fitting) model, then Δ_{HA} reduces to the ratio of the χ^2s for hypothesized versus null models. This index can also be used to test nested models by specifying Model H and Model A to correspond to specific nested hypotheses. The index can be useful in assessing fit, but has the drawback of an unknown statistical distribution. Furthermore, as Bentler and Bonett point out, it tests a different hypothesis than the usual χ^2 test, and is therefore, often not a substitute for the usual test.

5. Hoelter's Critical N (CN).

Hoelter (1983) presents an index called Critical N (CN) that is independent of sample size. This index tells us for a given discrepancy between $\Sigma_H(\theta)$ and $\Sigma_A(\theta)$, and a given level of protection against type I error, α, the sample size below which the null hypothesis could not be rejected. That is, if the sample were any smaller, one could not reject the null hypothesis. For the likelihood ratio χ^2 with r degrees of freedom (and α significance level),

$$CN = \chi^2_\alpha(r)/(\upsilon/N)$$

When $r < 100$, the critical value of the χ^2 distribution, χ^2_α, can be obtained from statistical tables of the χ^2 distribution. For larger values of r, the numerator of CN can be approximated by $(1/2)[Z_\alpha + (2r - 1)^{1/2}]^2$, where Z_α is the critical value for the standard normal distribution (see Matsueda and Bielby 1986). The index can tell us when we may have such a large sample (too much statistical power) that we are rejecting useful models. If we know that a particular set of models retained in samples of 200 or more observations have small discrepancies between $\Sigma_H(\theta)$ and $\Sigma_A(\theta)$, then we may safely conclude that so long as $CN < 200$, departures from the null hypothesis detected with large samples are substantively trivial. It can also tell us when we have such a small sample (too little statistical power) that we are retaining poor models. If we retain a null hypothesis and CN is much larger than 200, then there are likely to be substantively large but statistically nonsignificant discrepancies between $\Sigma_H(\theta)$ and $\Sigma_A(\theta)$.

6. BIC Statistic.

A very different test statistic departs from classical principles of statistical inference and can be justified with a Bayesian interpretation. The statistic, called a Bayesian Information Criterion (BIC) by Raftery is defined as:

$$BIC = \chi^2_H - df \ln n$$

This statistic has Bayesian interpretation given that

$$\operatorname{plim}_{n\to\infty} BIC \sim -2 \ln B \qquad \text{where B is P(Model H)/P(Model A)}$$

and P(Model H) is the probability that Model H is true given the data; P(Model A) is the probability that Model A is true given the data. These probabilities are treated as conditional on the researcher's degree of belief, but Raftery notes that prior beliefs exert little effect when the sample is extremely large. Furthermore, the Bayesian justification is not essential to use BIC as a model selection procedure: Asymptotically (as the sample size goes to infinity), the BIC statistic selects correct models with a high probability. A negative value of the BIC statistic indicates the model fits the data better than the alternative just identified model. The BIC can also be used to compare nested models, with the smaller BIC statistic indicating a better-fitting model. This use of the BIC is very useful. As a general fit measure in practice, this statistic appears useful in extremely large samples.

7. Indexes Based on the Noncentrality Parameter of υ.

Browne and Cudeck (1993) show that $n - 1\ [F_H(\theta)]$ is an unbiased estimator of the fit between S and $\Sigma_H(\theta)$ -- that is, between the sample covariance matrix and the covariance matrix implied by the model's estimates -- but is a *biased* estimator of the population discrepancy between Σ and $\Sigma_H(\theta)$ -- that is, the population covariance matrix and the

covariance matrix implied by the best-fitting population model. The latter is not a zero matrix because most social science models will not fit the population perfectly; thus, the discrepancy is due to misspecification in the population. Since in testing models, we're really interested in testing whether our population model fits the population covariance matrix, we may want an unbiased estimator of the discrepancy between Σ and $\Sigma_H(\theta)$ (see Matsueda and Bielby 1986). If we define $F_0 = F[\Sigma, \Sigma_H(\theta)]$ as the discrepancy between the population moment matrix and the population moment matrix fitted by the model, then under the alternative hypothesis, $\Sigma \neq \Sigma_H(\theta)$, $\upsilon = n - 1\ [F_H(\theta)]$ has a noncentral χ^2 distribution with noncentrality parameter $\tau = (n - 1)\ F_0$. Therefore, an approximation to the expected value of υ is $E(\upsilon) = F_0 + r/(n-1)$, where r is the number of degrees of freedom and n - 1 is the sample size. It follows that a less biased estimator of $F_0 = \Sigma$ and $\Sigma_H(\theta)$ is

$$\hat{F}_0 = \upsilon - r/(n - 1) = n - 1\ [F_H(\theta)] - r/N$$

This statistic, however, could be negative, so McDonald (1989) suggests simply bounding it at zero:

$$\hat{F}_0 = \text{Max}\ \{\upsilon - r/(n - 1), 0\}$$

That is, take as $\hat{F}_0$ the larger of the two values, either 0 or $\upsilon - r/(n - 1)$. But $\hat{F}_0$ is a random variable estimated from sample to sample, and therefore has sampling variability. Therefore, it may be useful to include not only the point estimate but also an interval estimate and construct 90% confidence intervals. A 90% confidence interval is given by

$$[(n-1)^{-1}\ \lambda_L;\ (n-1)^{-1}\ \lambda_U]$$

where λ_L is the solution for λ of the nonlinear equation for $G(\upsilon \mid \lambda, r) = .95$ if $G(\upsilon \mid \lambda, r) \geq .95$ and $\lambda_L = 0$ otherwise, and $\upsilon = n - 1\ [F_H(\theta)]$, and r = degrees of freedom. Similarly, λ_U is the solution for λ for $G(\upsilon \mid \lambda, r) = .05$ if $G(\upsilon \mid \lambda, r) \geq .05$ and $\lambda_L = 0$ otherwise.

The above statistic decreases when parameters are added to the model, which would induce researchers to add parameters to improve fit, departing from the principle of parsimony. To counteract this tendency, one can correct for degrees of freedom:

$$\epsilon_a = (F_0 / r) = \text{RMSEA (Root Mean Square Error of Approximation)}$$

To estimate this, one uses the sample estimators above:

$$\hat{\epsilon}_a = (\hat{F}_0 / r) = \text{Max}\ \{(\upsilon/r) - 1/(n - 1), 0\}$$

The 90% confidence interval is given by

$$(\hat{\epsilon}_{aL}\ ;\ \hat{\epsilon}_{aU}) = [(\lambda_L/(n - 1)r)^{1/2};\ \lambda_U/(n - 1)r)^{1/2}]$$

Interestingly, RMSEA is closely related to the simple index, χ^2/df:

$$\chi^2/df = n\text{-}1\ \hat{\epsilon}_a + 1 \quad \text{whenever } \hat{\epsilon}_a > 1$$

The difference is that RMSEA excludes the sample size, n, and therefore reduces the effect of n. But it doesn't eliminate sample size as a factor in fit, since n affects sampling variability in general.

In general, the larger the estimated RMSEA, the worse the fit of the model; similarly the larger the estimated noncentrality parameter, the worse the fit of the model. Browne and Cudeck (1993) suggest, based on their own experience, that RMSEA of .05 or less indicates a close fit of the model; RMSEA of .08 indicates a reasonable model; and RMSEA greater than .1 indicates a poor-fitting model. Using a 90% confidence interval, if the interval

covers zero (at the lower interval), one would not reject the null hypothesis at the .05 level. The interval at the upper end serves as a reminder that the model cannot be regarded as correct. One can also perform a test of "close fit" as opposed to exact fit in the population. If the null hypothesis of an exact fit is H_0: $\epsilon_a = 0$, which may be an implausible hypothesis in social science applications, one may want to replace this point hypothesis with an interval hypothesis of a "close fit": H_0: $\epsilon_a \leq 0.05$. This hypothesis will not be rejected at the $p > .05$ level if the lower limit of the 90% confidence interval of $\hat{\epsilon}_a$ is less than .05. One can also set up a test of close fit using the same test statistic but with an exceedance probability (p-value of close fit) given by:

$$P = 1 - G(\upsilon | \lambda^*, r)$$

where $\upsilon = n - 1[F_H(\theta)]$, $\lambda^* = r(n - 1)(.05)^2$. The null hypothesis is rejected in favor of the alternative hypothesis, H_0: $\epsilon_a > 0.05$, if P is less than a prespecified level of significance (e.g., $P < .05$).

8. Indexes Based on Information Theory and Cross-Validation

In covariance structure analysis -- as more generally in statistics and econometrics -- a number of selection criteria have been proposed to select the "best" among a set of nested models. Here are three fit statistics that are given in LISREL 8: (1) Akaike's Information Criterion (AIC); (2) Bozdogan's (1987) Consistent Akaike Information Criterion; and (3) Browne and Cudeck's (1989) single sample Expected Cross-Validation Index (ECVI).

Akaike's Information Criterion selects among a set of nested models that differ by number of parameters, and thus, numbers of degrees of freedom. Such models can be indexed by their degrees of freedom, r, where $r = 1, 2, \ldots K$). Then, for a model with r degrees of freedom, AIC is defined as:

$$AIC(r) = -2 \ln \lambda + 2r = n - 1 [F_H(\theta_{ML})] + 2r$$

The first term of AIC is simply the log-likelihood of the hypothesized model, which gives the likelihood ratio test statistic for that model, testing the overidentifying restrictions of the model using classical statistical inference. Thus, it is a measure of the badness of fit when the model is estimated by maximum likelihood. The second term measures the complexity or the penalty due to increased unreliability or compensation for bias in the first term, which depends on the number of parameters used to fit the data. The more parameters one uses, the greater the likelihood that the model will fit, so this needs to be included as a penalty. When comparing several models estimated by maximum likelihood, one chooses the model that has a minimum value of AIC. The best model will be the one that gives the highest information gain while minimizing complexity (and maximizing parsimony). This procedure is called the *minimum AIC procedure* and the chosen model is called the *minimum AIC estimate.*

One problem with AIC noted in the literature is that the multiplier $2r$ is chosen arbitrarily -- the penalty could just as well be $3r$ or $4r$, etc. Bozdogan (1987) notes that this multiplier should be chosen to so that AIC gives a consistent estimator of the true model -- that is as $n \rightarrow \infty$, the procedure should select the correct r. Since the AIC contains the noncentral χ^2, then degrees of freedom should be an increasing function of sample size, n. Bozdogan uses $r[(\ln n) + 1]$ as a multiplier, which gives a Consistent Akaike Information Criterion (CAIC):

$$CAIC(r) = -2 \ln \lambda + r[(\ln n) + 1] = n - 1 [F_H(\theta_{ML})] + r[(\ln n) + 1]$$

By selecting nested models that minimize CAIC, the model chosen will be the correct model asymptotically, as $n \rightarrow \infty$. However, consistency is a very *weak* property here. Note that $r[(\ln n) + 1]$ results in increasing the penalty of adding additional parameters, so that CAIC will select more restrictive models than AIC. Both AIC and CAIC can be applied to the hypothesized model as well as alternative baseline models, such as a model of independence (which assumes that all observed variables are orthogonal to each other and a saturated model (which is a just-identified alternative model). Under some substantive situations, these comparison models may be useful. Note that the saturated model will not fit perfectly using AIC and CAIC (unlike the saturated χ^2, which is zero with zero degrees of freedom), since these indices penalize for adding parameters.

Browne and Cudeck (1989) introduced a single sample Expected Cross-Validation Index (ECVI). This index is rooted in a cross-validation test of a model using two independent samples. Rather than computing cross-validation indexes for both calibration and validation samples, it proposes to compute the expected value of the calibration sample alone. The ECVI is defined as:

$$ECVI = n - 1\ [F_H(\theta_{ML})]/[(n - 1) + 2t/(n - 1)]$$

This gives an estimate of the discrepancy between the fitted covariance matrix to sample data and the expected value of a covariance matrix fitted to another sample of the same size. EVCI is linearly related to AIC and therefore is a consistent estimator. Moreover, when testing a series of nested models, it will give the same ordering of fit as AIC. One can also obtain a confidence interval on the estimated EVCI:

$$(c_L; c_U) = [(\lambda_L + k(k + 1)t/(n - 1); \lambda_U + k(k + 1)t/(n - 1)]$$

The confidence intervals can be used to determine whether the lower confidence interval is less than .05.

II. MULTIPLE GROUP MODELS.

Covariance structure analysis allows one to model covariance structures in different populations or groups. This provides a rigorous way of examining whether a given model is invariant across populations -- i.e., whether the parameters of a model interact with groups. For example, one may be interested in whether a status attainment model is invariant across black and white races. One could examine first whether the measurement parameters are invariant across race, and then examine whether returns to education varied across race. Thus, one can disentangle racial differences in measurement parameters with racial differences in substantive parameters. Specific hypotheses about invariance of parameters across groups can be tested using the likelihood ratio method (and when available the Wald and LM tests). The multi-group option requires that (1) one has drawn independent random samples from each population; (2) the model's structure is the same across groups (although this can be modified slightly by "tricking" the program with "pseudo variables"); and (3) the sample sizes of each group is sufficiently large to capitalize on asymptotic properties.

Consider a set of G populations, which are any clearly defined and mutually-exclusive populations such as social classes, races, genders, experimental groups, etc. Then the covariance structure model specifies the following parameter matrices for group g, where g = 1, 2,..., G:

$$\Lambda_y^{(g)}, \Lambda_x^{(g)}, \mathbf{B}^{(g)}, \Gamma^{(g)}, \Phi^{(g)}, \Psi^{(g)}, \Theta_\epsilon^{(g)}, \Theta_\delta^{(g)}$$

The general fitting function for groups pooled would be a weighted function of the fitting function for each group:

$$F = \sum_{g=1}^{G} (N_g/N)\ F_g\ [\mathbf{S}^{(g)}, \Sigma(\theta)^{(g)}, \mathbf{W}^{(g)}]$$

where N_g is the sample size of the gth group, N is the pooled sample size such that $N = N_1 + N_2 + ... + N_G$, $\mathbf{S}^{(g)}$ is the sample covariance of the gth group, $\Sigma(\theta)^{(g)}$ is the covariance matrix implied by the population parameters of the model, $\mathbf{W}^{(g)}$ is the weight matrix for WLS, and $F_g\ [\mathbf{S}^{(g)}, \Sigma(\theta)^{(g)}]$ is the fit function of the gth group. For ML, ULS, and GLS, the fitting function is:

$$F = \sum_{g=1}^{G} (N_g/N)\ F_g\ [\mathbf{S}^{(g)}, \Sigma(\theta)^{(g)}]$$

The maximum likelihood fitting function for each group, F_{gML}, used in LISREL 8 is:

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

Thus, maximum likelihood estimation will minimize each of these fitting functions (by minimizing the discrepancies between $S^{(g)}$ and $\Sigma(\theta)^{(g)}$, and minimize a weighted combination of the g functions. Note than when g = 1, this reduces to the fitting function we discussed under earlier when introducing ML estimation.
The program will provide a likelihood ratio test statistic, which gives the minimum value of the pooled fitting function, with degrees of freedom equal to the sum of degrees of freedom from each group:

$df = \frac{1}{2} G(p + q)(p + q + 1) - t$ where t is the total number of parameters estimated in all groups

LISREL 8 gives a single overall goodness-of-fit χ^2 statistic for all groups combined, but separate GFI, AGFI, and RMR for each group. The χ^2 will allow us to test hypotheses within groups, as well as between groups.

A. SETTING UP IN LISREL 8.

To estimate a model in multiple samples using LISREL 8, you basically follow the usual strategy with the following modifications:

1. On the data (DA) line for group 1, define the number of groups NG = *no. groups*. For example:

   DA NG=2 NI=4 NO=200

2. On the model (MO) line for group 1, define the number of variables and the *form* of the parameter matrices (ZE, ID, DI, SY, or FU). These must not change across groups. (However, the specification of fixed or free matrices *can* vary across groups.

3. Following the OU line for group 1, begin the control cards for group 2 (beginning with the title), which are followed by group 3, etc. This includes lines indicating where to read the data matrix for each group.

4. Within each group, parameter matrices are referred to in the usual way. LISREL will keep track of the group in which you are specifying parameters. However, at some point you will probably want to specify constraints across groups. To do this, specify for a given parameter matrix the group number first, followed by the elements of the matrix. For example,

   In group 1: FR BE 2 1 BE 3 1

   In group 2: FR BE 3 1
   EQ BE 1 2 1 BE 2 1

   In group 3: FR BE 3 1
   EQ BE 1 2 1 BE 2 1

   The above specifies that BE 2 1 and BE 3 1 are free parameters in the first group. In the second group, BE 2 1 is constrained to be equal to BE 2 1 of group 1, whereas BE 3 1 is unconstrained. The third group specifies the same. If one wanted to test whether BE 2 1 was invariant across the three groups, one could follow the above model with a model that relaxes the equality constraints.

   In group 1: FR BE 2 1 BE 3 1
   In group 2: FR BE 2 1 BE 3 1
   In group 3: FR BE 2 1 BE 3 1

   If this is the only difference between the two models, the difference in χ^2s of the two models will give a two-degree-of-freedom test of invariant β_{21}s.

5. There are some shortcuts in specifying multiple-sample models. These appear in the MO line in groups 2,...,G:

a. SP: If a given matrix has the identical pattern of fixed and free parameters (which is likely) of the previous group, one can specify SP to signify *same pattern.*

b. SS: If a given matrix has the same start values as the matrix in a previous group, one can specify SS, for *same start values.*

c. PS: If a given matrix has both the same pattern and start values, one can specify PS.

d. IN: If a given matrix is invariant across all groups (same pattern and values of all fixed and free parameters), one can specify IN for *invariant.*

A word of caution in examining the standardized solution in multiple groups. Typically one would not want to compare standardized coefficients across samples, since they take the parameters assumed to be invariant across samples (unstandardized slopes) and standardize them using sample-specific quantities (standard deviations of variables), which we assume are not invariant across samples. Because of this LISREL 8 will give a standardized solution based on a common scale *for all groups.* It does this by taking the covariance matrix of latent variables implied by the model of each group, computing a weighted average of these matrices (weighted by the group's sample size) and then transforming the weighted average covariance matrix into a correlation matrix. This solution, obtained with SS on the OU card, is called "SOLUTION STANDARDIZED TO A COMMON METRIC." The program will also give a standardized solution within each group, called "WITHIN GROUP STANDARDIZED SOLUTION." These are standardized in latent variables only. To obtain the same pair of standardized solutions (common metric and within group), one can specify SC on the OU card.

B. MODEL-FITTING AND HYPOTHESIS TESTING.

The specific tests of invariance carried out in multi-sample models will vary depending on the substantive context and research questions. The following, however, is likely to be a useful generic strategy for typical covariance structure models with multiple indicators of latent variables, and multiple equations of latent variables.

1. $H_{\text{Model H}}$: $\mathbf{S}^{(1)} = \Sigma(\theta)_H^{(1)}$; $\mathbf{S}^{(2)} = \Sigma(\theta)_H^{(2)}$; ... $\mathbf{S}^{(g)} = \Sigma(\theta)_H^{(g)}$

2. $H_{\Lambda x}$: $\Lambda_x^{(1)} = \Lambda_x^{(2)} = \ldots = \Lambda_x^{(G)}$ $\quad$ $H_{\Lambda y}$: $\Lambda_y^{(1)} = \Lambda_y^{(2)} = \ldots = \Lambda_y^{(G)}$

3. $H_{\theta\delta}$: $\theta_\delta^{(1)} = \theta_\delta^{(2)} = \ldots = \theta_\delta^{(G)}$ $\quad$ $H_{\theta\epsilon}$: $\theta_\epsilon^{(1)} = \theta_\epsilon^{(2)} = \ldots = \theta_\epsilon^{(G)}$

4. H_Γ: $\Gamma^{(1)} = \Gamma^{(2)} = \ldots = \Gamma^{(G)}$ $\quad$ H_B: $\mathbf{B}^{(1)} = \mathbf{B}^{(2)} = \ldots = \mathbf{B}^{(G)}$

5. H_Ψ: $\Psi^{(1)} = \Psi^{(2)} = \ldots = \Psi^{(G)}$

6. H_Φ: $\Phi^{(1)} = \Phi^{(2)} = \ldots = \Phi^{(G)}$

The first test tests whether the model fits reasonably well in each sample; thus it tests whether the same structural model fits in each sample. If this causes one to reject the models in one or more sample, then one may want to stop and conclude that different structures underlie the data in different populations.

The second test is important because it tells us whether the latent variables are measured on the same metric in each population. If this test fails to reject the null hypothesis of invariant lambdas, one can assume that the latent variables have identical metrics across populations. If the test rejects the null hypothesis, then one cannot be assured that latent variables have the same metrics, which could compromise one's ability to compare unstandardized coefficients across populations. For example, if occupational prestige scales have a different metric for blacks than nonblacks, estimating a model of returns to occupational status could end up being misleading. What can one do? First, examine the magnitude of departures form lambda invariance across groups. It could be

that the differences are substantively negligible, but a large sample leads to ample statistical power to detect minute differences across populations. Second, if in the unlikely event, one knows that one indicator has the identical lambda (unstandardized measurement slope λ_{ij}) across populations, normalize on this lambda, and proceed. Third, lacking such knowledge, one might vary the reference indicator and examine whether the relative effects among latent variables varies (the βs and γs). For any given latent construct, because we have to normalize on a reference indicator, only the *ratio* of lambdas is identified. Therefore, varying the reference indicator for a given construct (in each population) could vary the ratio of coefficients (βs and γs) across populations. For example, if the black-white difference in returns to prestige varies depending on which reference indicator we normalize on, one would not be able to conclude that returns really vary by race, since the result could be an artifact of the reference indicator chosen, because metrics really vary across race. However, if this doesn't alter substantive conclusions drawn, one would be more confident that comparisons are meaningful. Fourth, some (such as Jöreskog and Sörbom) recommend simply assuming away the problem by forcing Λ_x and Λ_y to be invariant. If you follow this strategy, proceed at your own risk!

The third test assesses whether the measurement error variances are invariant across populations. For example, one can test whether blacks respond with greater random measurement error than do whites. Note that this may be an important test for some models, since random measurement error in independent variables attenuates estimated regression coefficients. Failure to correct for measurement error when it is larger in some populations will result in greater attenuation in those populations, and therefore, biased cross-population comparisons.

The fourth test is typically the most critical, since it tests for invariance in structural parameters relating latent variables. Often one would set up specific a priori hypotheses about subsets of coefficients corresponding to specific substantive or theoretical propositions. For example, in a two-gender human capital model, one might be interested in the one-degree of freedom test of whether returns to education is invariant across gender.

The fifth and sixth tests are less-likely to be of substantive interest. The fifth assesses whether the variances of structural disturbances of equations for latent variables are invariant across populations. The sixth assesses whether the elements of the covariance matrix of exogenous variables is invariant across populations.

In setting up multiple-sample models, one typically begins by estimating the models separately in each sample, and then proceeds to test for invariance according to a selected hierarchy of hypotheses. However, depending on the substantive context, one might want to constrain all coefficients to be equal across groups, and then test selected hypotheses by relaxing specific parameter restrictions. Either way, the likelihood ratio method can be used to test restrictions, requiring that one estimate both the less-restrictive and more-restricted models for each hypothesis, and test the hypothesis with the difference in χ^2s. In the latter strategy, the EQS program provides both univariate and multivariate LM tests, which allows one to test constraints across populations while estimating only the more restrictive model. The univariate LM test, which is has a z^2 distribution, tests whether relaxing single equality constraints on parameters across populations will significantly improve the fit of the model. The multivariate LM test (distributed as χ^2 with df = number of parameters tested) assesses whether relaxing multiple equality constraints across populations will significantly improve the fit of the model. LISREL 8, on the other hand, does not provide modification indices (univariate LM statistics) to test restrictions across populations.