Maximum Likelihood and Robust Maximum Likelihood

Latent Trait Measurement and
Structural Equation Models
Lecture #2 – January 16, 2013

Today's Class

- A review of maximum likelihood estimation
  - How it works
  - Properties of MLEs

- Robust maximum likelihood for MVN outcomes
  - Augmenting likelihood functions for data that aren't quite MVN

Today's Example Data

- We use the data given in Enders (2010)
  - Imagine an employer is looking to hire employees for a job where IQ is important
  - Two variables:
    - IQ scores
    - Job performance (which potentially is missing)

- Enders reports three forms of the data:
  - Complete
  - Performance with Missing Completely At Random (MCAR) missingness
    - MCAR = missing process does not depend on observed or unobserved data
  - Performance with Missing At Random missingness
    - MAR = missing process depends on observed data only

- For our purposes, we will focus only on the complete data today
  - Note: ML (and Robust ML from today) can accommodate missing data assumed to be Missing At Random

IQ	Performance: Complete	Performance: MCAR	Performance: MAR
78	9	-	-
84	13	13	-
84	10	-	-
85	8	8	-
87	7	7	-
91	7	7	7
92	9	9	9
94	9	9	9
94	11	11	11
96	7	-	7
99	7	7	7
105	10	10	10
105	11	11	11
106	15	15	15
108	10	10	10
112	10	-	10
113	12	12	12
115	14	14	14
118	16	16	16
134	12	-	12

Descriptive Statistics

Variable	Mean	SD
IQ	100	14.13
Perf-C	10.35	2.68
Perf-MCAR	10.60	2.92
Perf-MAR	10.67	2.79

Covariance Matrix (denom = N)

Complete Data		
IQ	189.6	19.5
Performance	19.5	6.8

MCAR Data (Pairwise Deletion)		
IQ	115.6	19.4
Performance	19.4	8.0

MAR Data (Pairwise Deletion		
IQ	130.2	19.5
Performance	19.5	7.3

AN INTRODUCTION TO MAXIMUM LIKELIHOOD ESTIMATION

Why Estimation is Important

- In "applied" statistics courses estimation is not discussed frequently
  - Can be very technical…very intimidating

- Estimation is of critical importance
  - Quality and validity of estimates (and of inferences made from them) depends on how they were obtained

- Consider an absurd example:
  - I say the mean for IQ should be 20 – just from what I feel
  - Do you believe me? Do you feel like reporting this result?
    - Estimators need a basis in reality (in statistical theory)

How Estimation Works (More or Less)

- Most estimation routines do one of three things:

1. **Minimize Something:** Typically found with names that have "least" in the title. Forms of least squares include "Generalized", "Ordinary", "Weighted", "Diagonally Weighted", "WLSMV", and "Iteratively Reweighted." Typically the estimator of last resort…

2. **Maximize Something:** Typically found with names that have "maximum" in the title. Forms include "Maximum likelihood", "ML", "Residual Maximum Likelihood" (REML), "Robust ML". Typically the gold standard of estimators.

3. **Use Simulation to Sample from Something:** more recent advances in simulation use resampling techniques. Names include "Bayesian Markov Chain Monte Carlo", "Gibbs Sampling", "Metropolis Hastings", "Metropolis Algorithm", and "Monte Carlo". Used for complex models where ML is not available or for methods where prior values are needed.

Properties of Maximum Likelihood Estimators

- Provided several assumptions ("regularity conditions") are met, maximum likelihood estimators have good statistical properties:

1. Asymptotic Consistency: as the sample size increases, the estimator converges in probability to its true value

2. Asymptotic Normality: as the sample size increases, the distribution of the estimator is normal (with variance given by "information" matrix)

3. Efficiency: No other estimator will have a smaller standard error

- Because they have such nice and well understood properties, MLEs are commonly used in statistical estimation

Maximum Likelihood: Estimates Based on Statistical Distributions

- Maximum likelihood estimates come from statistical distributions – assumed distributions of data
  - We will begin today with the univariate normal distribution but quickly move to other distributions (see this Friday's class)

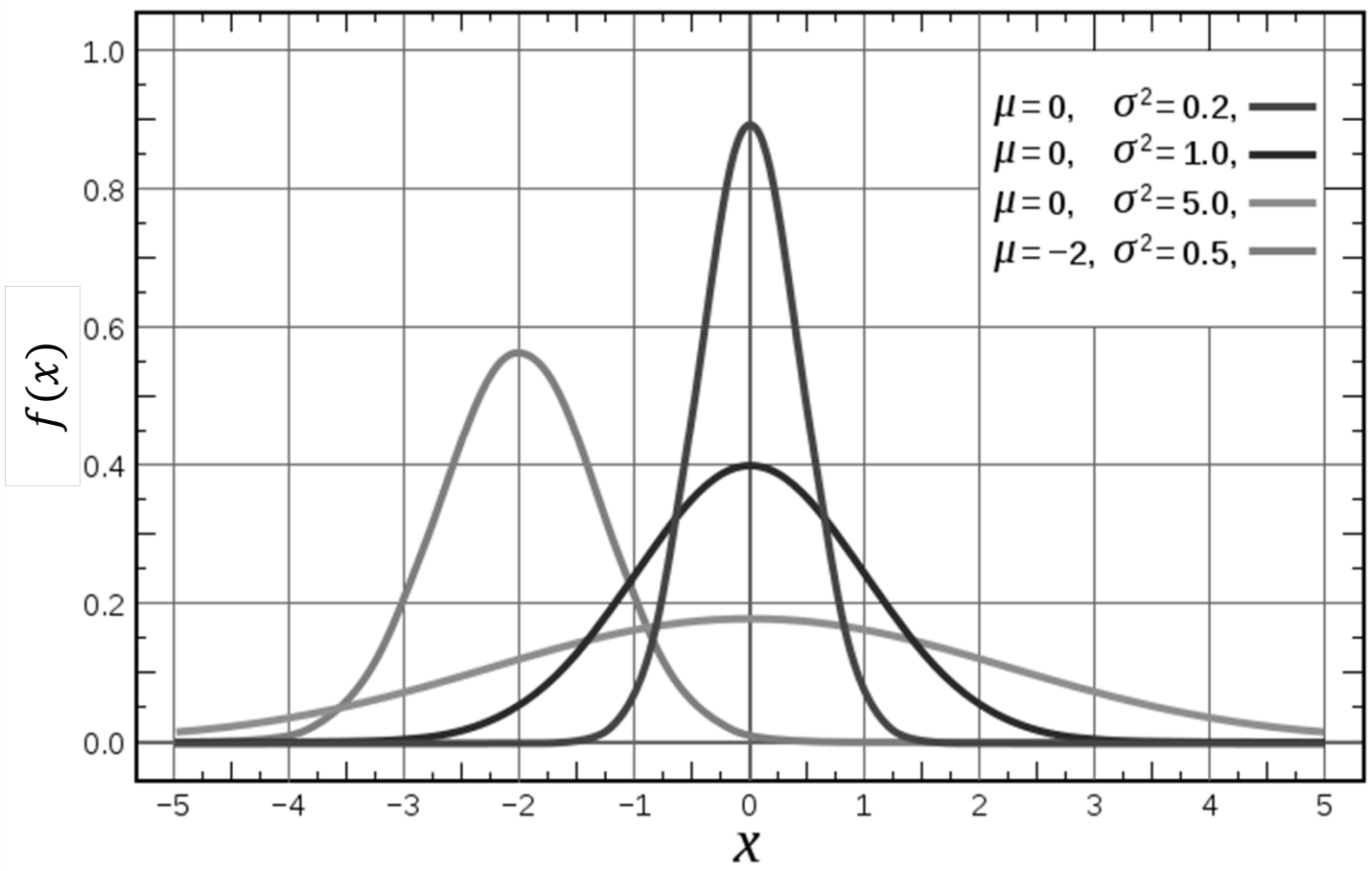
- For a single random variable x, the univariate normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$$

  - Provides the height of the curve for a value of x, μ_x, and σ_x^2

- Last week we pretended we knew μ_x and σ_x^2
  - Today we will only know x (and maybe σ_x^2)

Univariate Normal Distribution



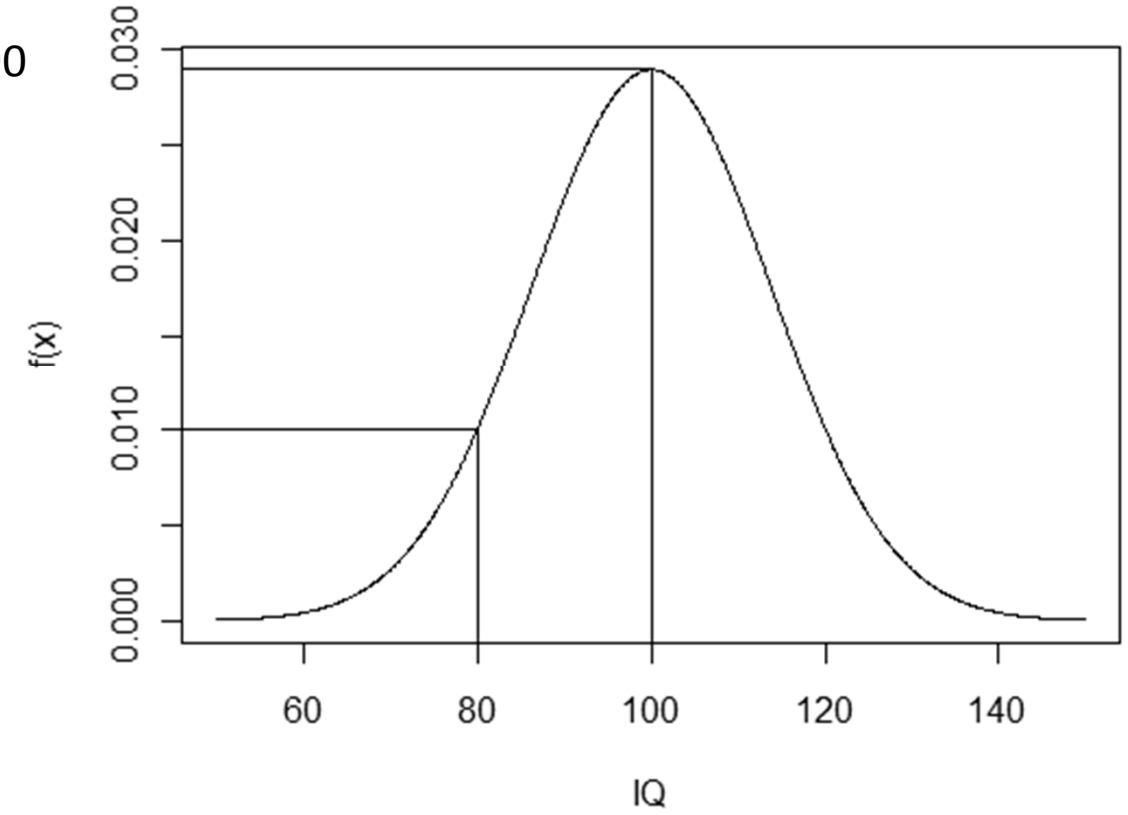
For any value of x , μ_x, and σ_x^2, $f(x)$ gives the height of the curve (relative frequency)

Example Distribution Values

- Let's examine the distribution values for the IQ variable
  - We assume that we **know** $\mu_x = 100$ and $\sigma_x^2 = 189.6$
    - Later we will not know what these values happen to be

For $x = 100$, $f(100) = 0.0290$

For $x = 80$, $f(80) = 0.0101$



Constructing a Likelihood Function

- Maximum likelihood estimation begins by building a **likelihood function**
  - A likelihood function provides a value of a likelihood (think height of a curve) for a set of statistical parameters

- Likelihood functions start with probability density functions (PDFs)
  - Density functions are provided for each observation individually (marginal)

- The likelihood function for the entire sample is the function that gets used in the estimation process
  - The sample likelihood can be thought of as a joint distribution of all the observations, simultaneously
  - In univariate statistics, observations are considered independent, so the joint likelihood for the sample is constructed through a product

- To demonstrate, let's consider the likelihood function for one observation

A One-Observation Likelihood Function

- Let's assume the following:
  - We have observed IQ (for the person where $x = 112$)
  - That IQ comes from a normal distribution
  - That the variance of x is known to be 189.6 ($\sigma_x^2 = 189.6$)
    - This is to simplify the likelihood function so that we only don't know one value
    - More on this later…empirical under-identification

- For this one observation, the likelihood function takes its assumed distribution and uses its PDF:

$$f(x, \mu_x, \sigma_x^2) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$$

- The PDF above now is expressed in terms of the three unknowns that go into it: x, μ_x, σ_x^2

A One-Observation Likelihood Function

- Because we know two of these terms ($x = 112; \sigma_x^2 = 189.6$), we can create the likelihood function for the mean:

$$L(\mu_x | x = 112, \sigma_x^2 = 189.6) = \frac{1}{\sqrt{2\pi * 189.6}} \exp\left(-\frac{(112 - \mu_x)^2}{2 * 189.6}\right)$$

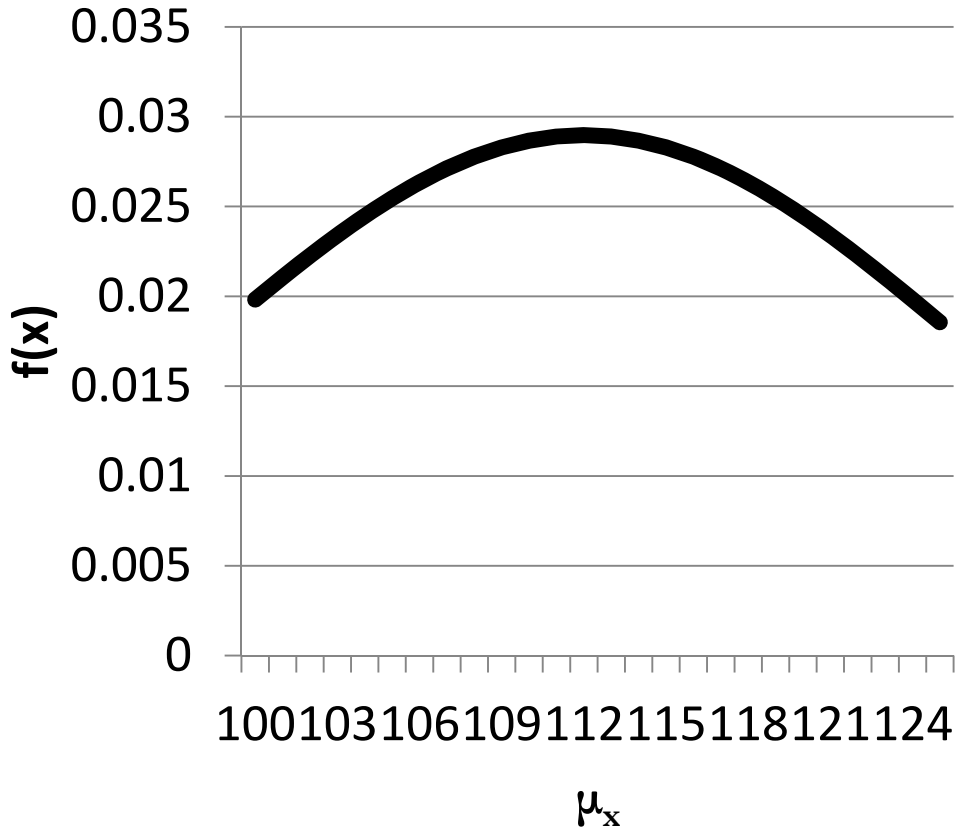
- For every value of μ_x *could be*, the likelihood function now returns a number that is called **the likelihood**
  - The actual value of the likelihood is not relevant (yet)

- The value of μ_x with the highest likelihood is called the **maximum likelihood estimate (MLE)**
  - For this one observation, what do you think the MLE would be?
  - This is asking: what is the most likely mean that produced these data?

The MLE is…

- The value of μ_x that maximizes $L(\mu_x|x, \sigma_x^2)$ is $\hat{\mu}_x = 112$
  - The value of the likelihood function at that point is $L(112|x, \sigma_x^2) = .029$

For $\hat{\mu}_x = 112$, $L(112|x, \sigma_x^2) = .029$



From One Observation…To The Sample

- The likelihood function shown previously was for one observation, but we will be working with a sample
  - Assuming the sample observations are independent and identically distributed, we can form the joint distribution of the sample
  - For normal distributions, this means the observations have the same mean and variance

Multiplication comes from independence assumption:
Here, $L(\mu_x, \sigma_x^2 | x_p)$ is the univariate normal PDF for x_p, μ_x, and σ_x^2

$$L(\mu_x, \sigma_x^2 | x_1, \dots, x_N) = L(\mu_x, \sigma_x^2 | x_1) \times L(\mu_x, \sigma_x^2 | x_2) \times \cdots \times L(\mu_x, \sigma_x^2 | x_N)$$

$$= \prod_{p=1}^{N} f(x_p) = \prod_{p=1}^{N} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x_p - \mu_x)^2}{2\sigma_x^2} \right) =$$

$$(2\pi\sigma_x^2)^{-\frac{N}{2}} \exp\left(-\sum_{p=1}^{N} \frac{(x_p - \mu_x)^2}{2\sigma_x^2} \right)$$

Maximizing the Log Likelihood Function

- The process of finding the values of μ_x *and* σ_x^2 that maximize the likelihood function is complicated
  - What was shown was a grid search: trial-and-error process

- For relatively simple functions, we can use calculus to find the maximum of a function mathematically
  - Problem: not all functions can give closed-form solutions (i.e., one solvable equation) for location of the maximum
  - Solution: use efficient methods of searching for parameter (i.e., Newton-Raphson)

Standard Errors: Using the Second Derivative

- Although the estimated values of the sample mean and variance are needed, we also need the standard errors

- For MLEs, the standard errors come from the **information matrix**, which is found from the square root of -1 times the inverse matrix of second derivatives (only one value for one parameter)
  - Second derivative gives curvature of log-likelihood function

MAXIMUM LIKELIHOOD WITH THE MULTIVARIATE NORMAL DISTRIBUTION

ML with the Multivariate Normal Distribution

- The example from the first part of class focused on a single variable from a univariate normal distribution
  - We typically have multiple variables (p) from a multivariate normal distribution

$$f(\mathbf{x}_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)}{2}\right]$$

The Multivariate Normal Distribution

$$f(\mathbf{x}_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)}{2}\right]$$

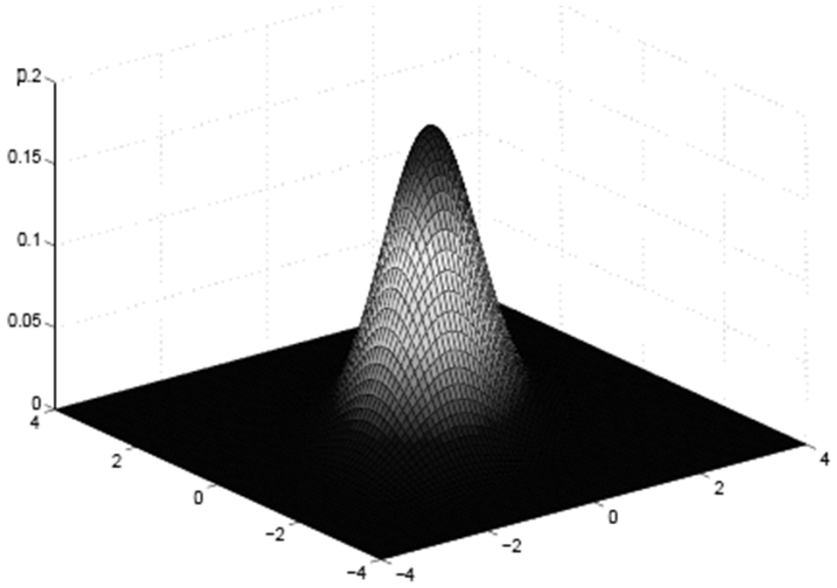
- The mean vector is $\boldsymbol{\mu} = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \\ \vdots \\ \mu_{x_p} \end{bmatrix}$

- The covariance matrix is $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \cdots & \sigma_{x_1x_p} \\ \sigma_{x_1x_2} & \sigma_{x_2}^2 & \cdots & \sigma_{x_2x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_1x_p} & \sigma_{x_2x_p} & \cdots & \sigma_{x_p}^2 \end{bmatrix}$

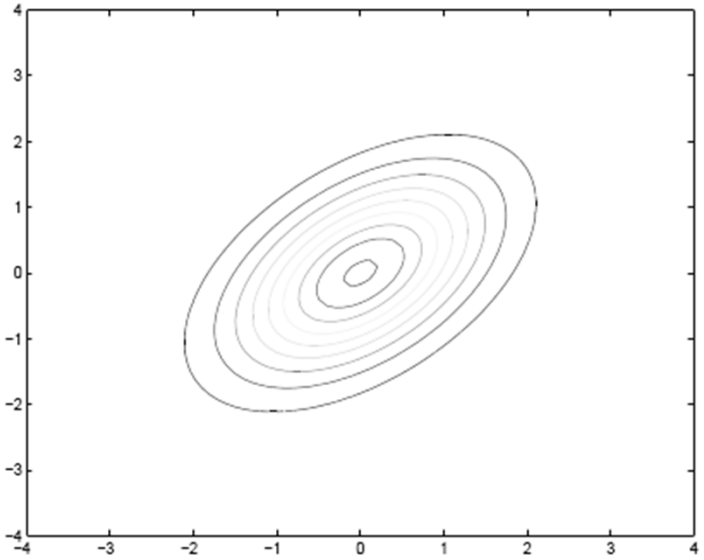
  - The covariance matrix must be non-singular (invertible)

Multivariate Normal Plot

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{x_1} & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma^2_{x_2} \end{bmatrix} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$$



Density Surface (3D)



Density Surface (2D):
Contour Plot

Example Distribution Values

- Let's examine the distribution values for the both variables
  - We assume that we **know** $\boldsymbol{\mu} = \begin{bmatrix} 100 \\ 10.35 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 189.6 & 19.5 \\ 19.5 & 6.8 \end{bmatrix}$
    - We will not know what these values happen to be in practice

- The MVN distribution function gives the height of the curve for values of both variables: IQ and Performance
  - $f(\boldsymbol{x}_i = [100 \quad 10.35]) = 0.0052$
    - This is an observation exactly at the mean vector – highest likelihood
  - $f(\boldsymbol{x}_i = [130 \quad 13]) = 0.0004$
    - This observation is distant from the mean vector – lower likelihood

From One Observation…To The Sample

- The distribution function shown on the last slide was for one observation, but we will be working with a sample
  - Assuming the sample are independent and identically distributed, we can form the joint distribution of the sample

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = f(\boldsymbol{x}_1) \times f(\boldsymbol{x}_2) \times \cdots \times f(\boldsymbol{x}_N) = \prod_{i=1}^{N} f(\boldsymbol{x}_i) =$$

$$\prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)}{2} \right] =$$

$$(2\pi)^{-\frac{Np}{2}} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\left[\sum_{i=1}^{N} -\frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)}{2} \right]$$

The Sample MVN Likelihood Function

- From the previous slide:

$$L(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) =$$

$$L = (2\pi)^{-\frac{Np}{2}} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\left[\sum_{i=1}^{N} -\frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)}{2}\right]$$

- For this function, there is one mean vector ($\boldsymbol{\mu}$), one covariance matrix ($\boldsymbol{\Sigma}$), and all of the data ($\mathbf{X}$)

- If we observe the data but **do not know** the mean vector and/or covariance matrix, then we call this the sample likelihood function

- Rather than provide the height of the curve of any value of x, it provides the ***likelihood*** for any values of***$\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$***
  - ***Goal of Maximum Likelihood is to find* values of $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *that maximize this function***

The Log-Likelihood Function

- The likelihood function is more commonly re-expressed as the log-likelihood: $\log L = \ln(L)$
  - The natural log of L

$$\log L = \log\left[= (2\pi)^{-\frac{Np}{2}} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\left[\sum_{i=1}^{N} -\frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)}{2}\right]\right] =$$

$$-\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log(|\boldsymbol{\Sigma}|) - \sum_{i=1}^{N} \frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i^T - \boldsymbol{\mu}\right)}{2}$$

- The log-likelihood and the likelihood have a maximum at the same location of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Log-Likelihood Function In Use

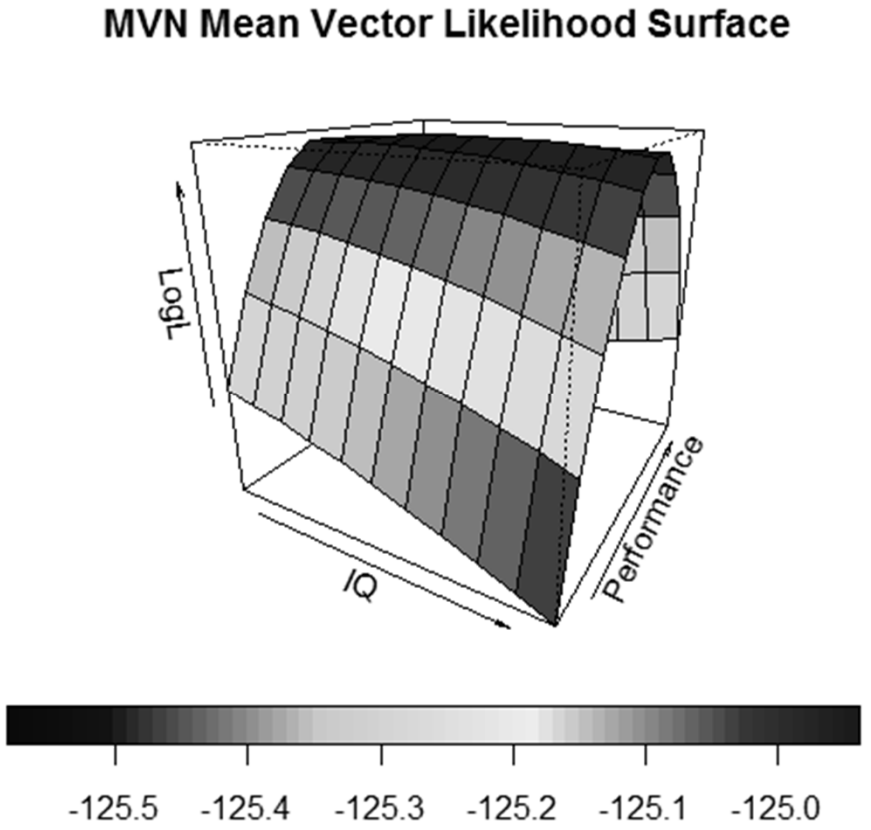
- Imagine we know that $\boldsymbol{\Sigma} = \begin{bmatrix} 189.6 & 19.5 \\ 19.5 & 6.8 \end{bmatrix}$ but not $\boldsymbol{\mu}$

- The log-likelihood function will give us the likelihood of a range of values of $\boldsymbol{\mu}$

- The value of $\boldsymbol{\mu}$ where $\log L$ is the maximum is the MLE for $\boldsymbol{\mu}$:

- $\widehat{\boldsymbol{\mu}} = \begin{bmatrix} 100 \\ 10.35 \end{bmatrix}$

- $\log L = \log 5.494e - 55$
  $= -124.9385$



Finding MLEs In Practice

- Most likelihood functions do not have closed form estimates
  - Iterative algorithms must be used to find estimates

- Iterative algorithms begin at a location of the log-likelihood surface and then work to find the peak
  - Each iteration brings estimates closer to the maximum
  - Change in log-likelihood from one iteration to the next should be small

- If models have latent (random) components, then these components are "marginalized" – removed from equation
  - Called Marginal Maximum Likelihood

- Once the algorithm finds the peak, then the estimates used to find the peak are called the MLEs
  - And the information matrix is obtained providing standard errors for each

MAXIMUM LIKELIHOOD BY MVN: USING MPLUS FOR ESTIMATION

Using MVN Likelihoods in MPlus

- In Mplus, the default assumption for variables is a linear (mixed) models procedure that uses (full information) ML with the multivariate normal distribution
  - Full Information = All Data Contribute

- You can use Mplus to do analyses for all sorts of linear models including:
  - MANOVA
  - Repeated Measures ANOVA
  - Multilevel models/Hierarchical Linear Models
  - (Some) Factor Models

- The MVN is what we will use for the first part of this class
  - Later, we will work with distributions for categorical outcomes

An Unconditional Model in Mplus

- An unconditional model (where no predictors are used) will give us ML estimates of the mean vector and covariance matrix when using Mplus

```
VARIABLE:
NAMES = ID IQ perfC perfMCAR perfMAR;
IDVARIABLE = ID;
USEVARIABLE = IQ perfC;
MISSING = .;

MODEL:
IQ WITH perfC;
```

- By default Mplus:
  - …Enters all named variables into an analysis
    - The USEVARIABLE command limits the number of variables
  - …Specifies a multivariate normal distribution for all variables
    - If your data are categorical you must change this (see Week 14)
  - …Specifies all variables to be uncorrelated
    - Use the WITH command to estimate the covariance between variables

Mplus Output: Model Information

```
SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                          20

Number of dependent variables                    2
Number of independent variables                  0
Number of continuous latent variables            0

Observed dependent variables

  Continuous
   IQ          PERFC

Variables with special functions

  ID variable           ID

Estimator                                       ML
Information matrix                        OBSERVED
Maximum number of iterations                  1000
Convergence criterion                    0.500D-04
Maximum number of steepest descent iterations   20
Maximum number of iterations for H1           2000
Convergence criterion for H1             0.100D-03
```

Check # of Subjects

Check # of Dependent Variables

Check Estimator (ML should be gold standard)

Mplus Output: Iteration History

- Mplus uses an iterative algorithm to find MLEs:
  - You can see the history using the OUTPUT: TECH5 command

- Important message (if not shown, don't trust output –not at peak of log-likelihood function):

- Iteration summary:

```
THE MODEL ESTIMATION TERMINATED NORMALLY
```

```
TECHNICAL 5/6 OUTPUT


TECHNICAL OUTPUT FROM EM ALGORITHM ITERATIONS FOR THE H1 MODEL

   ITER        FUNCTION     ABS CHANGE   REL CHANGE

      1     0.91684618D+01
      2     0.88180975D+01   -0.3503643   -0.0382141
      3     0.88180975D+01   -0.3503643   -0.0382141
```

Mplus Output: Covariance Parameters

- Covariance Parameter Estimates w/SEs
  - More on the Est./S.E. and Two-Tailed P-Value shortly (hint: Wald Test)

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
IQ WITH				
PERFC	19.500	9.151	2.131	0.033
Means				
IQ	100.000	3.079	32.478	0.000
PERFC	10.350	0.584	17.714	0.000
Variances				
IQ	189.600	59.957	3.162	0.002
PERFC	6.827	2.159	3.162	0.002

- Mean Vector: $\hat{\boldsymbol{\mu}} = \begin{bmatrix} 100.00 \\ 10.35 \end{bmatrix}$
- Covariance Matrix: $\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 189.600 & 19.500 \\ 19.500 & 6.827 \end{bmatrix}$

This is the "saturated" model: no more parameters are possible. Consider this to be the target for the best model

USEFUL PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATES

Likelihood Ratio (Deviance) Tests

- The likelihood value from MLEs can help to statistically test competing models
  - Assuming none of the parameters are at their boundary
    - Boundary issues happen when testing some covariance parameters as a variance cannot be less than zero

- Likelihood ratio tests take the ratio of the likelihood for two models and use it as a test statistic

- Using log-likelihoods, the ratio becomes a difference
  - The test is sometimes called a deviance test

$$D = \Delta - 2\log L = -2 \times (\log L_{H0} - \log L_{HA})$$

  - D is tested against a Chi-Square distribution with degrees of freedom equal to the difference in number of parameters

Deviance Test Example

- Imagine we wanted to test the hypothesis that the unstructured covariance matrix in our empty model was different from what we would have if the data were from independent observations
- Null Model: $\mathbf{R} = \sigma_e^2\mathbf{I} = 98.21\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} = \begin{bmatrix}98.21 & 0\\ 0 & 98.21\end{bmatrix}$
- Alternative Model: $\mathbf{R} = \mathbf{\Sigma} = \begin{bmatrix}189.6 & 19.5\\ 19.5 & 6.8\end{bmatrix}$
- The difference between the two models is two parameters
  - Null model: one variance estimated = 1 parameter
  - Alternative model: two variances and one covariance estimated = 2 parameters

Deviance Test Procedure

- Step #1: estimate null model (get log likelihood)

```
MODEL:
IQ WITH perfC @0;
IQ (1);
perfC (1);
```

```
MODEL FIT INFORMATION

Number of Free Parameters                       3

Loglikelihood

          H0 Value                       -148.500
          H1 Value                       -124.939
```

- Step #2: estimate alternative model (get log likelihood)
  - Note: Mplus does this automatically for the unstructured covariance matrix (we do this here to show the process)

```
MODEL:
IQ WITH perfC;
```

```
MODEL FIT INFORMATION

Number of Free Parameters                       5

Loglikelihood

          H0 Value                       -124.939
          H1 Value                       -124.939
```

Deviance Test Procedure

- Step #3: compute test statistic

$D = -2 \times (\log L_{H0} - \log L_{H1}) = -2 * (-148.500 - -124.939) = 47.122$

  - Note, this is actually output in Mplus:
  - From Null Model

Chi-Square Test of Model Fit

Value	47.124
Degrees of Freedom	2
P-Value	0.0000

- Step #4: calculate p-value from Chi-Square Distribution with 2 degrees of freedom (I used =chidist() from Excel)
  - p-value < 0.0001

- Inference: the two parameters were significantly different from zero -- we prefer our alternative model to the null model

- Interpretation: the unstructured covariance matrix fits better than the independence model

Residual Covariance Matrix

- When an reduced form model (not saturated /unstructured) is estimated a way of determining how "close" the reduced model fits the full model is to look at the residual covariances
  - Residual = Full Model Covariance – Reduced Model Covariance
  - Obtained in Mplus by adding the sord "RESIDUAL" under the OUTPUT section:
- Reduced Model: $\widehat{\boldsymbol{\Sigma}}_R = \sigma_e^2 \mathbf{I} = 98.21 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 98.21 & 0 \\ 0 & 98.21 \end{bmatrix}$
- Full Model: $\widehat{\boldsymbol{\Sigma}}_F = \begin{bmatrix} 189.6 & 19.5 \\ 19.5 & 6.8 \end{bmatrix}$
- Residual Covariance = $\widehat{\boldsymbol{\Sigma}}_F - \widehat{\boldsymbol{\Sigma}}_R = \begin{bmatrix} 91.386 & 19.5 \\ 19.5 & -91.386 \end{bmatrix}$

```
OUTPUT:
TECH5 RESIDUAL;
```

Wald Tests

- For each parameter θ, we can form the Wald statistic:

$$\omega = \frac{\hat{\theta}_{MLE} - \theta_0}{SE(\hat{\theta}_{MLE})}$$

  - (typically $\theta_0 = 0$)

- As N gets large (goes to infinity), the Wald statistic converges to a standard normal distribution $\omega \sim N(0,1)$
  - Gives us a hypothesis test of $H_0: \theta = 0$

- If we divide each parameter by its standard error, we can compute the two-tailed p-value from the standard normal distribution
  - Exception: bounded parameters can have issues (variances)

Wald Test Example

- Although the Wald tests for the variance parameters shouldn't be used, Mplus computes them:

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
IQ WITH				
PERFC	19.500	9.151	2.131	0.033
Means				
IQ	100.000	3.079	32.478	0.000
PERFC	10.350	0.584	17.714	0.000
Variances				
IQ	189.600	59.957	3.162	0.002
PERFC	6.827	2.159	3.162	0.002

- Similarly, we could test whether the mean job performance was equal to zero using

$$\omega = \frac{10.35}{0.5843} = 17.7;\ p < 0.0001$$

Information Criteria

- Information criteria are statistics that help determine the relative fit of a model
  - Comparison is fit-versus-parsimony
  - Often used to compare non-nested models
- Mplus reports a set of criteria (from unstructured model)

```
Information Criteria

          Akaike (AIC)                       259.877
          Bayesian (BIC)                     264.856
          Sample-Size Adjusted BIC           249.442
            (n* = (n + 2) / 24)
```

  - Each uses -2*log-likelihood as a base
    - Choice of statistic is **very** arbitrary and depends on field (I use BIC)
- Best model is one with smallest value
  - Information criteria from independence model (unstructured wins):

```
Information Criteria

          Akaike (AIC)                       303.001
          Bayesian (BIC)                     305.988
          Sample-Size Adjusted BIC           296.740
            (n* = (n + 2) / 24)
```

ROBUST ML IN CFA/SEM

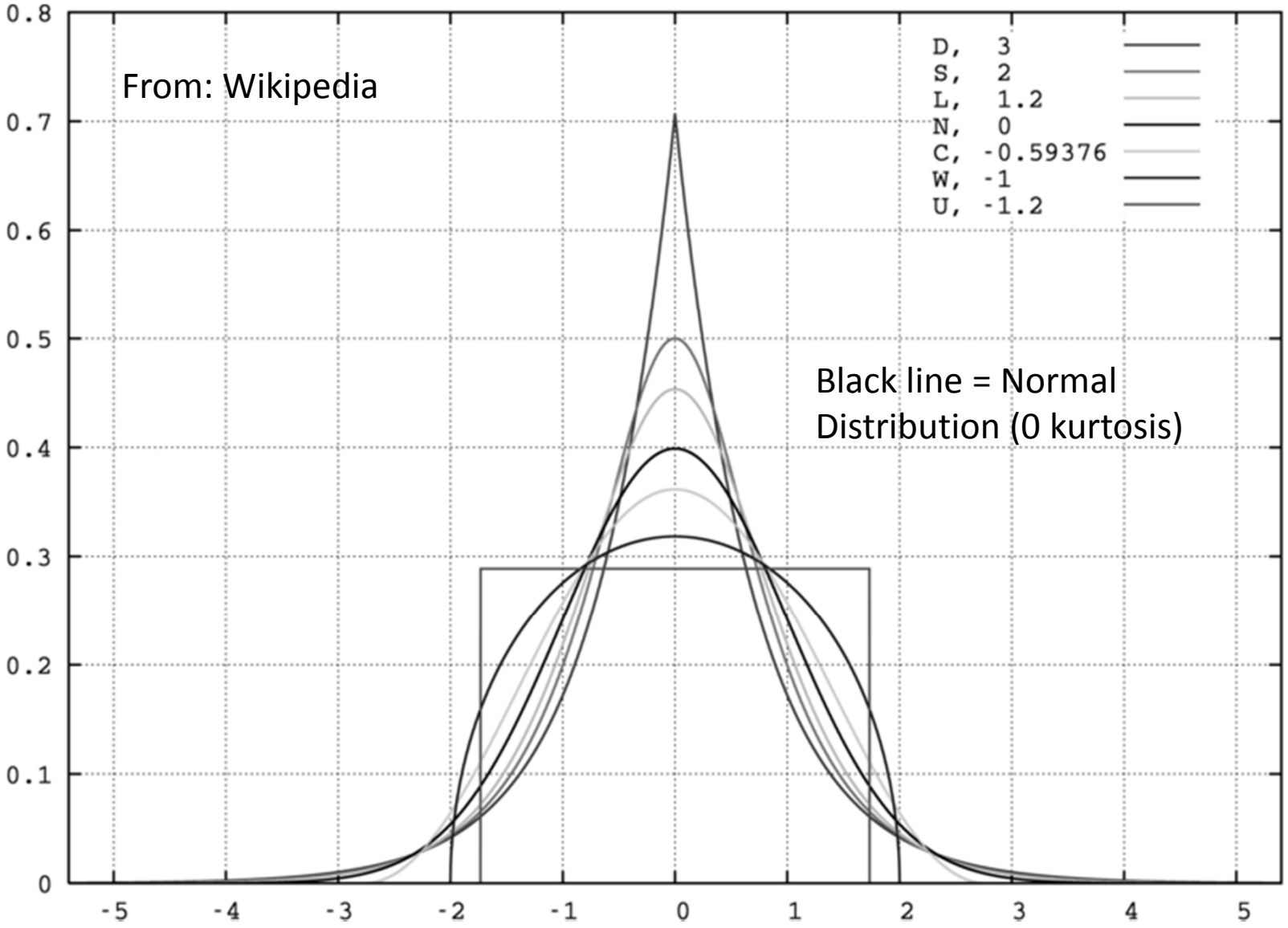
Robust Estimation: The Basics

- Robust estimation in ML still assumes the data follow a multivariate normal distribution
  - But that the data have more or less kurtosis than would otherwise be common in a normal distribution

- Kurtosis: measure of the shape of the distribution
  - From Greek word for bulging
  - Can be estimated for data (either marginally for each item or jointly across all items)

- The degree of kurtosis in a data set is related to how incorrect the log-likelihood value will be
  - Leptokurtic data (too-fat tails): χ^2 inflated, SEs too small
  - Platykurtic data (too-thin tails): χ^2 depressed, SEs too large

Visualizing Kurtosis



Robust ML for Non-Normality in Mplus: MLR

- Robust ML can be specified very easily in Mplus:
  - Add ANALYSIS: ESTIMATOR = MLR; to your code

- The model parameter **estimates** will all be identical to those found under regular maximum likelihood
  - And…if data are MVN – then no adjustment is made (so we can use MLR for everything!)

- MLR adjusts:
  - Model χ^2 (and associated fit statistics that use it: RMSEA, CFI, TLI) – closely related to Yuan-Bentler T_2 (permits MCAR or MAR missing data)

  - Model **standard errors**: uses Huber-White "sandwich" estimator to adjust standard errors
    - Sandwich estimator found using information matrix of the partial first derivatives to correct information matrix from the partial second derivatives

Adjusted Model Fit Statistics

- Under MLR, model fit statistics are adjusted based on an estimated scaling factor:
  - Scaling factor = 1.000
    - Perfectly MVN data
  - Scaling factor > 1.000
    - Leptokurtosis (too-fat tails; fixes too big χ^2)
  - Scaling factor < 1.000
    - Platykurtosis (too-thin tails; fixes too small χ^2)

- The scaling factor will now show up in all likelihood ratio tests (deviance tests)
  - So you must add it to your calculations

Adjusted Standard Errors

- The standard errors for all parameter estimates will be different under MLR
  - Remember, these are used in Wald tests

- If the data show leptokurtosis (too-fat tails):
  - Increases information matrix
  - Fixes too small SEs

- If the data show platykurtosis (too-thin tails):
  - Lowers values in information matrix
  - Fixes too big SEs

Data Analysis Example with MLR

- To demonstrate, we will revisit our analysis of the example data for today's class using MLR

- So far, we have estimated two models:
  - Saturated model
  - Independence model

- The results of the two analyses (ML v. MLR) will be compared

- Because MLR is something does not affect our results if we have MVN data, we should have been using MLR all along

Mplus Syntax

- Saturated model syntax:

```
TITLE:
Example of Multivariate Normal Distribution
Using Robust Maximum Likelihood
Unstructured Covariance Matrix

DATA:
FILE = jobperf.csv;

VARIABLE:
NAMES = ID IQ perfC perfMCAR perfMAR;
IDVARIABLE = ID;
USEVARIABLE = IQ perfC;
MISSING = .;

ANALYSIS:
ESTIMATOR = MLR;

MODEL:

!Listing the variable by itself estimates the variance under MVN
IQ; perfC;

!The WITH statement provides an estimate of the covariance
IQ WITH perfC;

OUTPUT:
TECH5 STANDARDIZED RESIDUAL SAMPSTAT;
```

Mplus Output: Log-likelihoods Under ML and MLR

Under ML

```
MODEL FIT INFORMATION

Number of Free Parameters                       5

Loglikelihood

          H0 Value                       -124.939
          H1 Value                       -124.939
```

Under MLR

```
MODEL FIT INFORMATION

Number of Free Parameters                       5

Loglikelihood

          H0 Value                       -124.939
          H0 Scaling Correction Factor      0.9095
            for MLR
          H1 Value                       -124.939
          H1 Scaling Correction Factor      0.9095
            for MLR
```

- The actual log-likelihoods are the same
  - But, under MLR, the log-likelihood gets re-scaled

- The scaling factor for the saturated model is 0.9095 – indicates slightly platykurtic data

Adding Scaling Factors to the Analysis

- The MLR-estimated scaling factors are used to rescale the log-likelihoods under LR test model comparisons
  - Extra calculations are needed

- The rescaled LR test is given by:

$$LR_{RS} = \frac{-2\left(\log L_{restricted} - \log L_{full}\right)}{c_{LR}}$$

- The denominator is found by the scaling factors (c) and number of parameters (q) in each model:

$$c_{LR} = \left|\frac{(q_{restricted})(c_{restricted}) - \left(q_{full}\right)\left(c_{full}\right)}{\left(q_{restricted} - q_{full}\right)}\right|$$

- Sometimes c_{LR} can be negative - so take the absolute value

Model Comparison: Independence v. Saturated Model

Independence Model

```
MODEL FIT INFORMATION

Number of Free Parameters                       3

Loglikelihood

          H0 Value                       -148.500
          H0 Scaling Correction Factor      1.2223
            for MLR
          H1 Value                       -124.939
          H1 Scaling Correction Factor      0.9095
            for MLR
```

Saturated Model

```
MODEL FIT INFORMATION

Number of Free Parameters                       5

Loglikelihood

          H0 Value                       -124.939
          H0 Scaling Correction Factor      0.9095
            for MLR
          H1 Value                       -124.939
          H1 Scaling Correction Factor      0.9095
            for MLR
```

- To compare the independence model against the saturated model we must first calculate the scaling factor
  - $q_{full} = 5$ – number of parameters in saturated model
  - $c_{full} = 0.9095$ – scaling factor from saturated model
  - $q_{reduced} = 3$ – number of parameters in one-factor model
  - $c_{reduced} = 1.2223$ – scaling factor from one-factor model

- The scaling factor for the LR test is then:

$$c_{LR} = \frac{3 * 1.2223 - 5 * 0.9095}{3 - 5} = \frac{-0.8806}{-2} = .4403$$

Model Comparison #1: Independence v. Saturated Model

- The next step is to calculate the re-scaled likelihood ratio test using the original log-likelihoods and the scaling factor:

$$LR_{RS} = \frac{-2\left(\log L_{restricted} - \log L_{full}\right)}{c_{LR}}$$

$$= \frac{-2(-148.500 - -124.939)}{.4403} = 107.0225$$

- Finally, we use the rescaled LR test as we would in any other LR test- compare it to a χ^2 with df = difference in number of parameters (here 2)
  - I use "=chidist(107.0225,1)" in Excel
  - Our test has p-value < .001 – so the one-factor model is not preferred to the H1 saturated model

Mplus Output: Chi-Square Test of Model Fit

- Our model comparison is what Mplus calculates and reports under the section "Chi-Square Test of Model Fit"

```
Chi-Square Test of Model Fit

          Value                             107.007*
          Degrees of Freedom                      2
          P-Value                           0.0000
          Scaling Correction Factor         0.4404
            for MLR

*   The chi-square value for MLM, MLMV, MLR, ULSMV, WLSM and WLSMV cannot be used
    for chi-square difference testing in the regular way.  MLM, MLR and WLSM
    chi-square difference testing is described on the Mplus website.  MLMV, WLSMV,
    and ULSMV difference testing is done using the DIFFTEST option.
```

- We will see that in the use of SEM, many caution against using this test…even so, the MLR version has better performance
  - Still not used as frequently as the other fit indices

Results Unchanged Under MLR

- Information criteria are unchanged under MLR:

Under ML

Information Criteria

Akaike (AIC)	259.877
Bayesian (BIC)	264.856
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	249.442

Under MLR

Information Criteria

Akaike (AIC)	259.877
Bayesian (BIC)	264.856
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	249.442

Standard Errors/Wald Tests Under MLR

Under ML

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
IQ WITH				
PERFC	19.500	9.151	2.131	0.033
Means				
IQ	100.000	3.079	32.478	0.000
PERFC	10.350	0.584	17.714	0.000
Variances				
IQ	189.600	59.957	3.162	0.002
PERFC	6.827	2.159	3.162	0.002

Under MLR

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
IQ WITH				
PERFC	19.500	6.578	2.964	0.003
Means				
IQ	100.000	3.079	32.478	0.000
PERFC	10.350	0.584	17.714	0.000
Variances				
IQ	189.600	56.722	3.343	0.001
PERFC	6.827	1.831	3.729	0.000

- The SEs of our model under MLR are smaller than the SEs under ML
  - As such, the values of the Wald tests are larger (SEs are the denominator)

MLR: The Take-Home Point

- If you feel you have continuous data that are (tenuously) normally distributed, use MLR
  - Any time you use SEM/CFA/Path Analysis as we have to this point
  - In general, likert-type items with 5 or more categories are treated this way
  - If data aren't/cannot be considered normal we should still use different distributional assumptions

- If data truly are MVN, then MLR doesn't adjust anything

- If data are not MVN (but are still continuous), then MLR adjusts the important inferential portions of the results

WRAPPING UP

Wrapping Up

- Today was a refresher course on ML and "Robust" ML estimation
  - These topics will come back to us each week – so lots of opportunity for practice

- These topics are important when using Mplus as there are quite a few different estimators in the package
  - ML is not always the default

- Homework #2: available on our website – due next Wednesday (January 23rd) at 11:59am