Maximum Likelihood Estimation

EPSY 905: Fundamentals of Multivariate Modeling

Online Lecture #6

In This Lecture…

- The basics of maximum likelihood estimation
  - The engine that drives most modern statistical methods

- Additional information from maximum likelihood estimator (MLEs)
  - Likelihood ratio tests
  - Wald tests
  - Information criteria

- MLEs for GLMs
  - An introduction to the NLME (non-linear mixed effects) and LME (linear mixed effects) packages in R
  - We'll also use the lavaan package in R (ML for Path Analysis)

Today's Example Data #1

- Imagine an employer is looking to hire employees for a job where IQ is important
  - We will only use 5 observations so as to show the math behind the estimation calculations
- The employer collects two variables:
  - IQ scores
  - Job performance
- Descriptive Statistics:

Variable	Mean	SD
IQ	114.4	2.30
Performance	12.8	2.28

Covariance Matrix		
IQ	5.3	5.1
Performance	5.1	5.2

Observation	IQ	Performance
1	112	10
2	113	12
3	115	14
4	118	16
5	114	12

How Estimation Works (More or Less)

- Most estimation routines do one of three things:

1. **<u>Minimize Something:</u>** 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. **<u>Maximize Something:</u>** 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. **<u>Use Simulation to Sample from Something:</u>** 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.

AN INTRODUCTION TO MAXIMUM LIKELIHOOD ESTIMATION

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

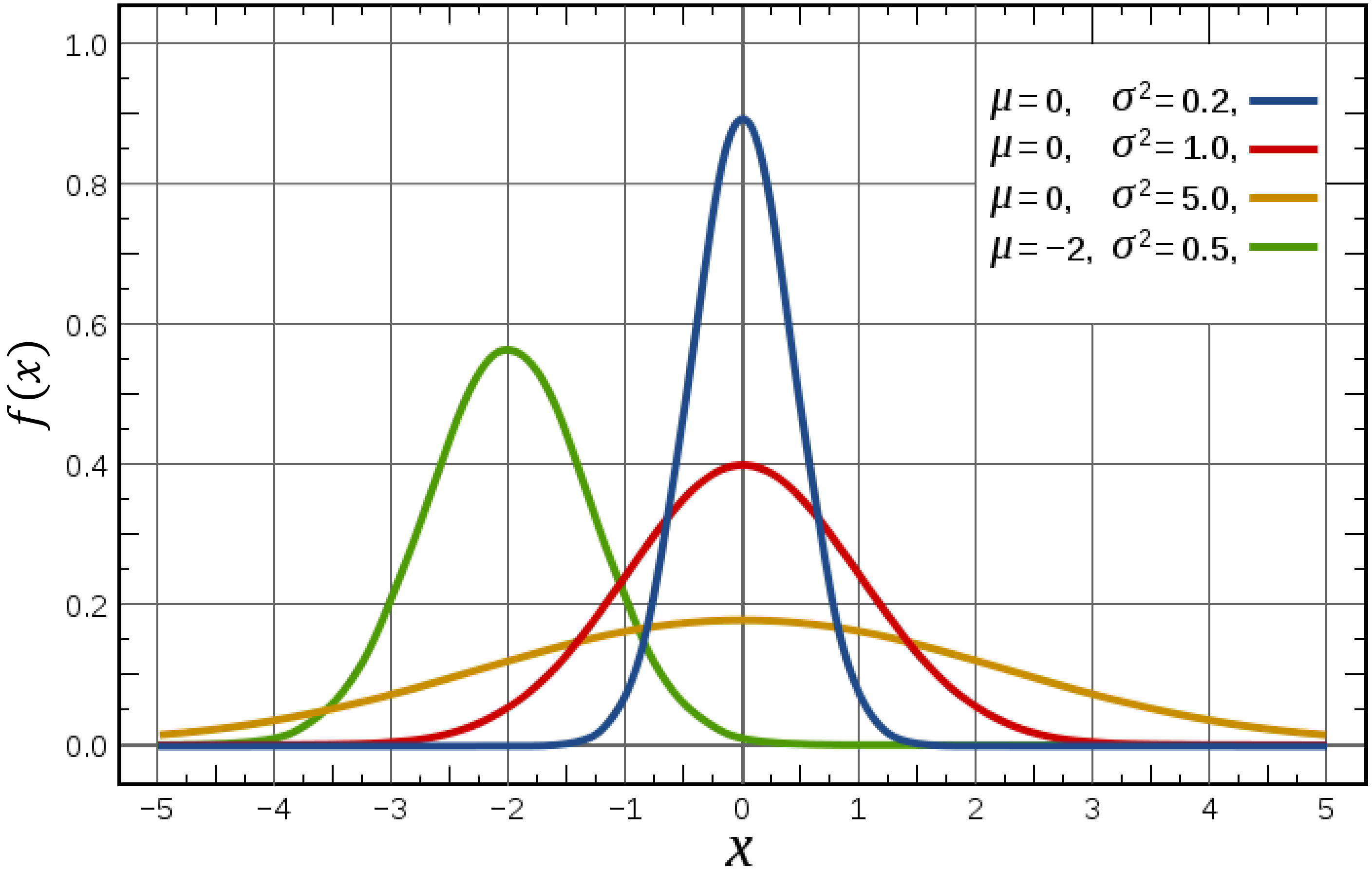
- For a single random variable x, the univariate normal distribution is

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

  - 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



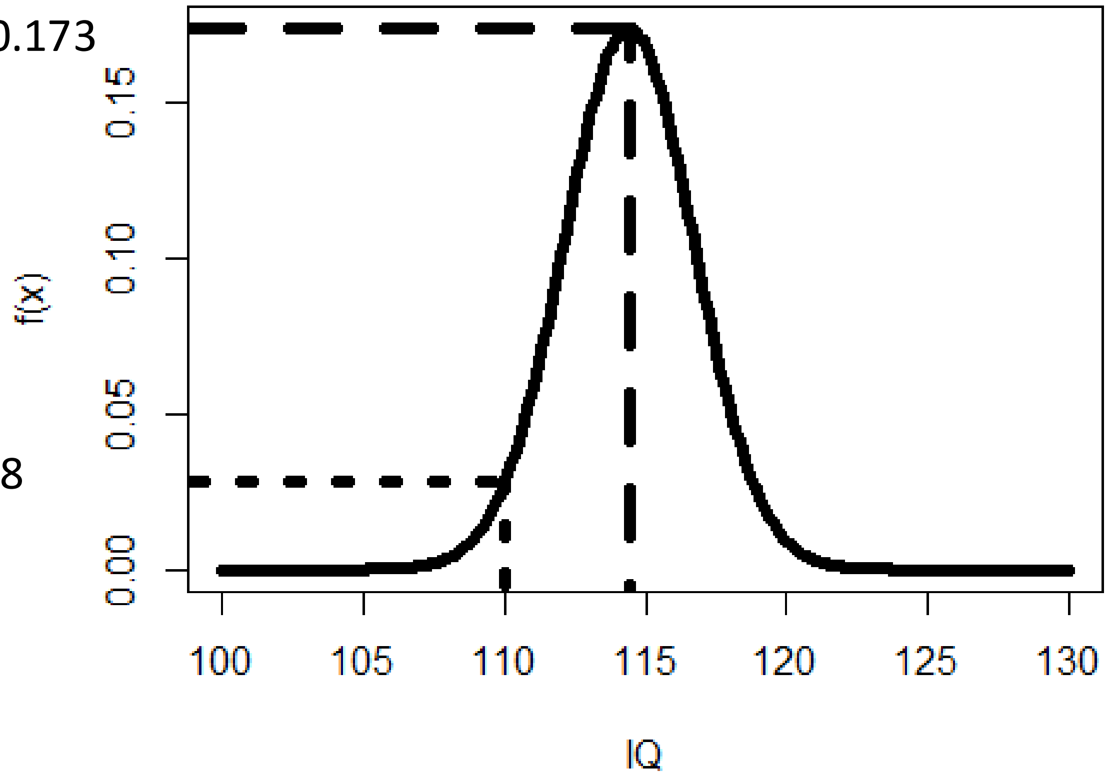
For any value of x , μ_x, and σ_x^2, $f(x)$ gives the height of the curve (relative frequency)

Example Distribution Values

- Let's examine the distribution values for the IQ variable
  - We assume that we **know** $\mu_x = 114.4$ and $\sigma_x^2 = 5.29$ ($\sigma_x = 2.30$)
    - In reality we do not know what these values happen to be

For $x = 114.4$, $f(114.4) = 0.173$

For $x = 110$, $f(110) = 0.028$



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 the first value of IQ ($x = 112$)
  - That IQ comes from a normal distribution
  - That the variance of x is known to be 5.29 ($\sigma_x^2 = 5.29$)
    - 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 ($x = 112; \sigma_x^2 = 5.29$), we can create the likelihood function for the mean:

$$L(\mu_x | x = 112, \sigma_x^2 = 5.29) = \frac{1}{\sqrt{2\pi * 5.29}} \exp\left(-\frac{(112 - \mu_x)^2}{2 * 5.29}\right)$$

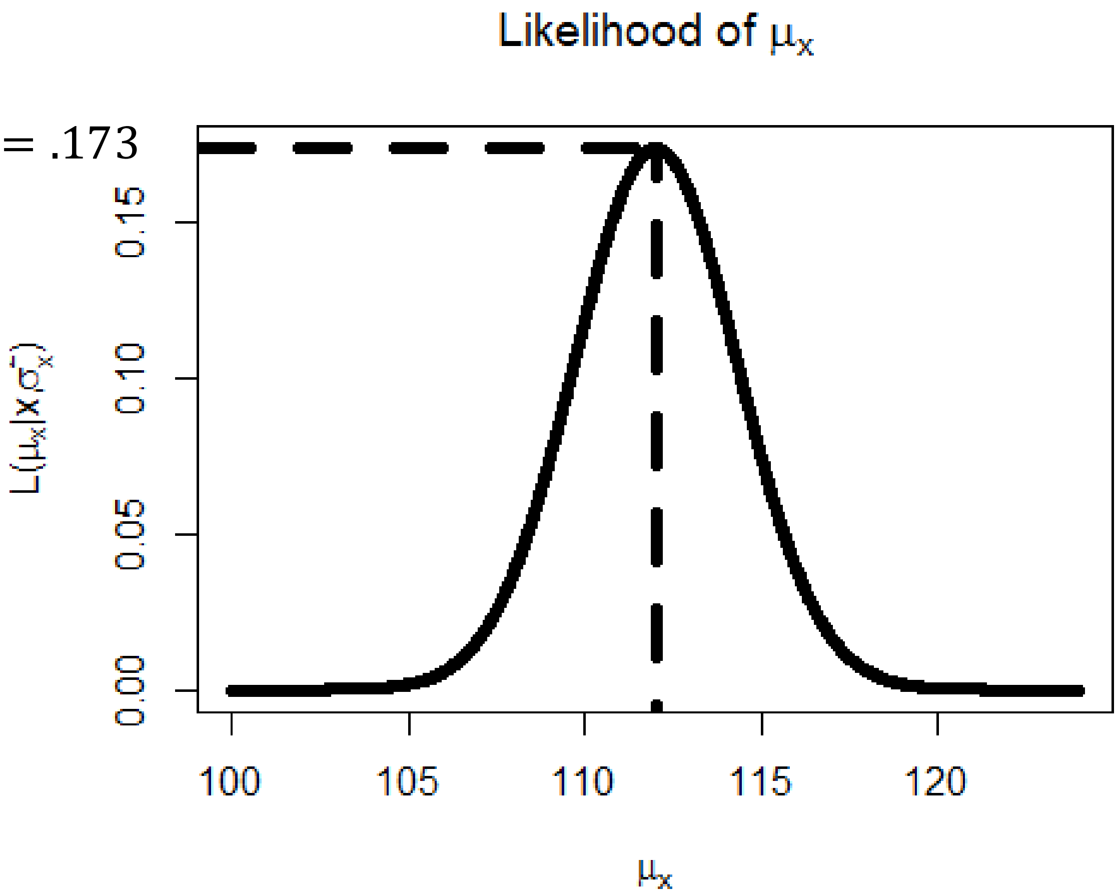
- For every value of μ_x *could be*, the likelihood function now returns a number that is called **the likelihood**
  - The actual value of the likelihood is not relevant (yet)

- 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 $L(112|x, \sigma_x^2) = .173$

For $\hat{\mu}_x = 112$, $L(112|x, \sigma_x^2) = .173$



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

$$
\begin{aligned}
L(\mu_x, \sigma_x^2 | x_1, \dots, x_N) &= L(\mu_x, \sigma_x^2 | x_1) \times L(\mu_x, \sigma_x^2 | x_2) \times \cdots \times L(\mu_x, \sigma_x^2 | x_N) \\
&= \prod_{p=1}^{N} f(x_p) = \prod_{p=1}^{N} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x_p - \mu_x)^2}{2\sigma_x^2}\right) = \\
&(2\pi\sigma_x^2)^{-\frac{N}{2}} \exp\left(-\sum_{p=1}^{N} \frac{(x_p - \mu_x)^2}{2\sigma_x^2}\right)
\end{aligned}
$$

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

The Sample Likelihood Function

- From the previous slide:

$$L(x_1, \dots, x_N | \mu_x, \sigma_x^2) = L = (2\pi\sigma_x^2)^{-\frac{N}{2}} \exp\left(-\sum_{p=1}^{N} \frac{(x_p - \mu_x)^2}{2\sigma_x^2} \right)$$

- For this function, there is one mean (μ_x), one variance (σ_x^2), and all of the data $(x_1, \dots, x_N)$
- If we observe the data but **do not know** the mean and/or variance, 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 possible values of μ_x ***and*** σ_x^2
  - ***Goal of Maximum Likelihood is to find* values of $\boldsymbol{\mu_x}$ *and* $\boldsymbol{\sigma_x^2}$ *that maximize this function***

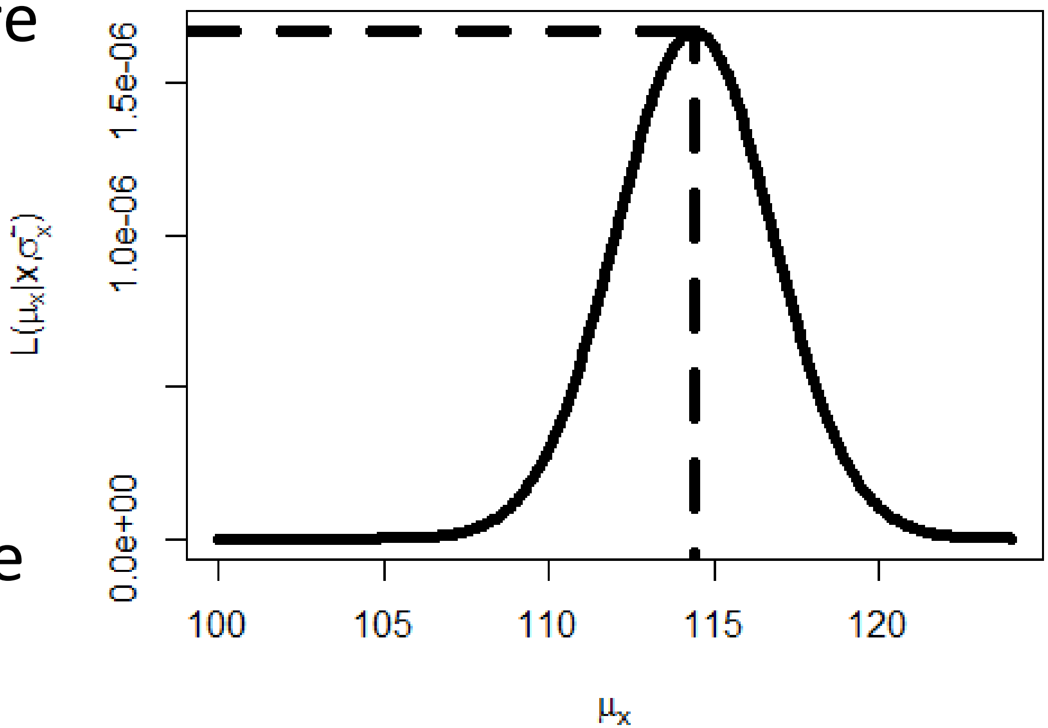
Likelihood Function for All Five Observations

- Imagine we know that $\sigma_x^2 = 5.29$ but we do not know μ_x

- The likelihood function will give us the likelihood of a range of values of μ_x:

- The value of μ_x where L is the maximum is the MLE for μ_x:
- $\hat{\mu}_x = 114.4$
- $L = 1.67e - 06$

- Note: likelihood value abbreviated as L



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

$$\log L = \log L(\mu_x, \sigma_x^2 | x_1, \ldots, x_N)$$
$$= \log\big(L(\mu_x, \sigma_x^2 | x_1) \times L(\mu_x, \sigma_x^2 | x_2) \times \cdots \times L(\mu_x, \sigma_x^2 | x_N)\big)$$
$$= \sum_{p=1}^{N} \log L\big(\mu_x, \sigma_x^2 \big| x_p\big) = \log\left[(2\pi\sigma_x^2)^{-\frac{N}{2}} \exp\left(-\sum_{p=1}^{N} \frac{\left(x_p - \mu_x\right)^2}{2\sigma_x^2}\right)\right] =$$
$$-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma_x^2) - \sum_{p=1}^{N} \frac{\left(x_p - \mu_x\right)^2}{2\sigma_x^2}$$

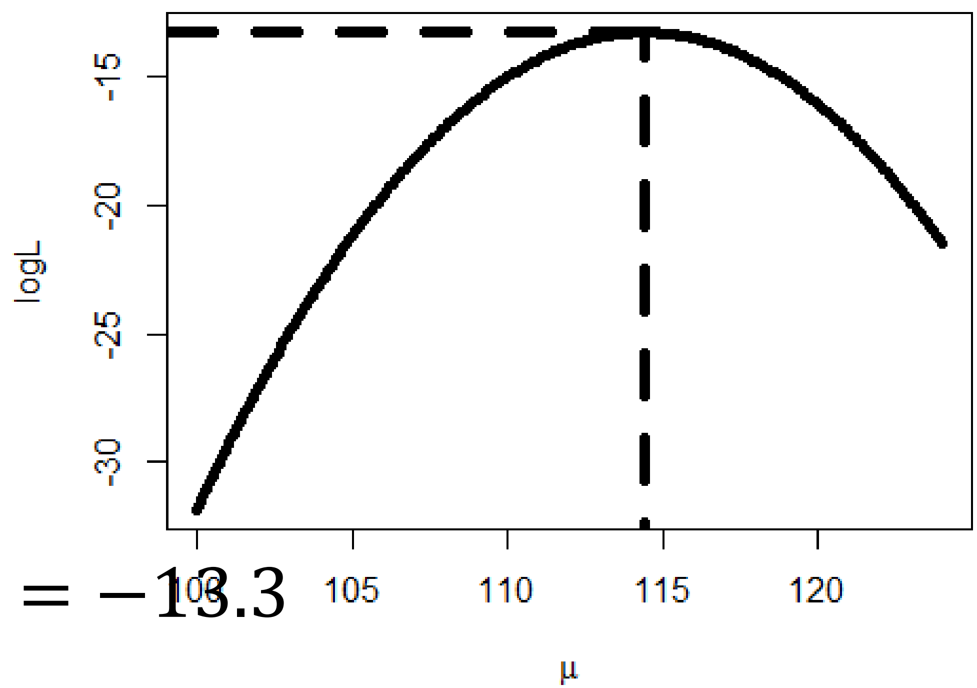
- The log-likelihood and the likelihood have a maximum at the same location of μ_x and σ_x^2

Log-Likelihood Function In Use

- Imagine we know that $\sigma_x^2 = 5.29$ but we do not know μ_x

- The log-likelihood function will give us the likelihood of a range of possible values of μ_x

- The value of μ_x where $\log L$ is the maximum is the MLE for μ_x:



- $\hat{\mu}_x = 114.4$

- $\log L = \log 1.67e - 06 = -13.3$

But…What About the Variance?

- Up to this point, we have assumed the sample variance was known
  - Not likely to happen in practice

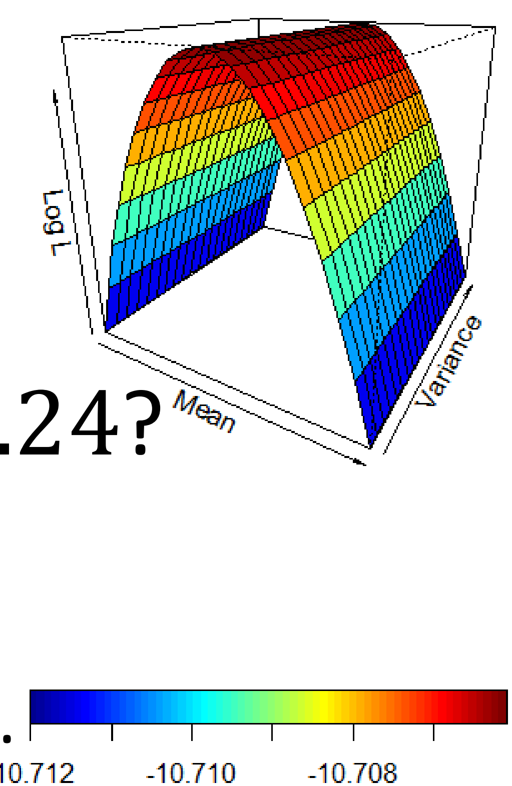
- We can jointly estimate the mean and the variance using the same log likelihood (or likelihood) function
  - The variance is now a parameter in the model
  - The likelihood function now will be with respect to two dimensions
    - Each unknown parameter is a dimension

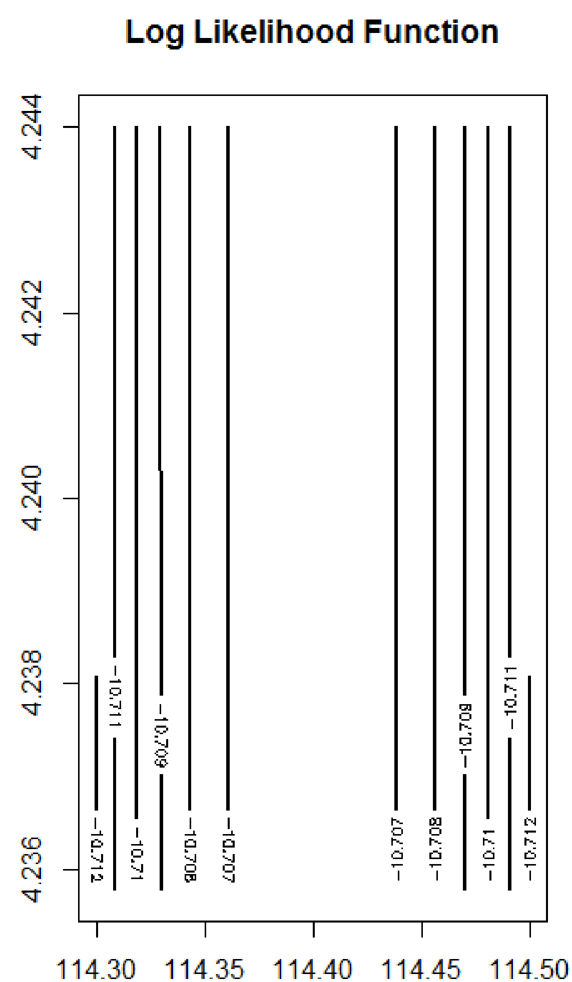
$$\log L = \log L(\mu_x, \sigma_x^2 | x_1, \dots, x_N)$$

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

The Log Likelihood Function for Two Parameters

- The point where log *L* is the maximum is the MLE for μ_x and σ_x^2
- $\log L = -10.7$
- $\hat{\mu} = 114.4$
- $\sigma_x^2 = 4.24$
- Wait…$\sigma_x^2 = 4.24$?
  - ➢ It was 5.29 on slide 3
  - ➢ Why? Think $\frac{1}{N}$…





Maximizing the Log Likelihood Function

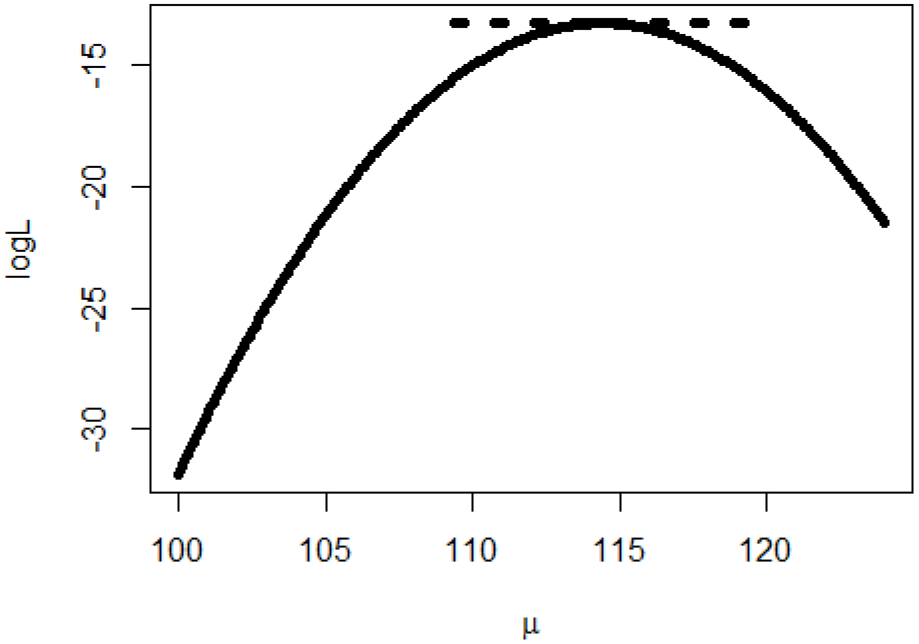
- The process of finding the values of μ_x *and* σ_x^2 that maximize the likelihood function is complicated
  - What was shown was a grid search: trial-and-error process

- 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)

Using Calculus: The First Derivative

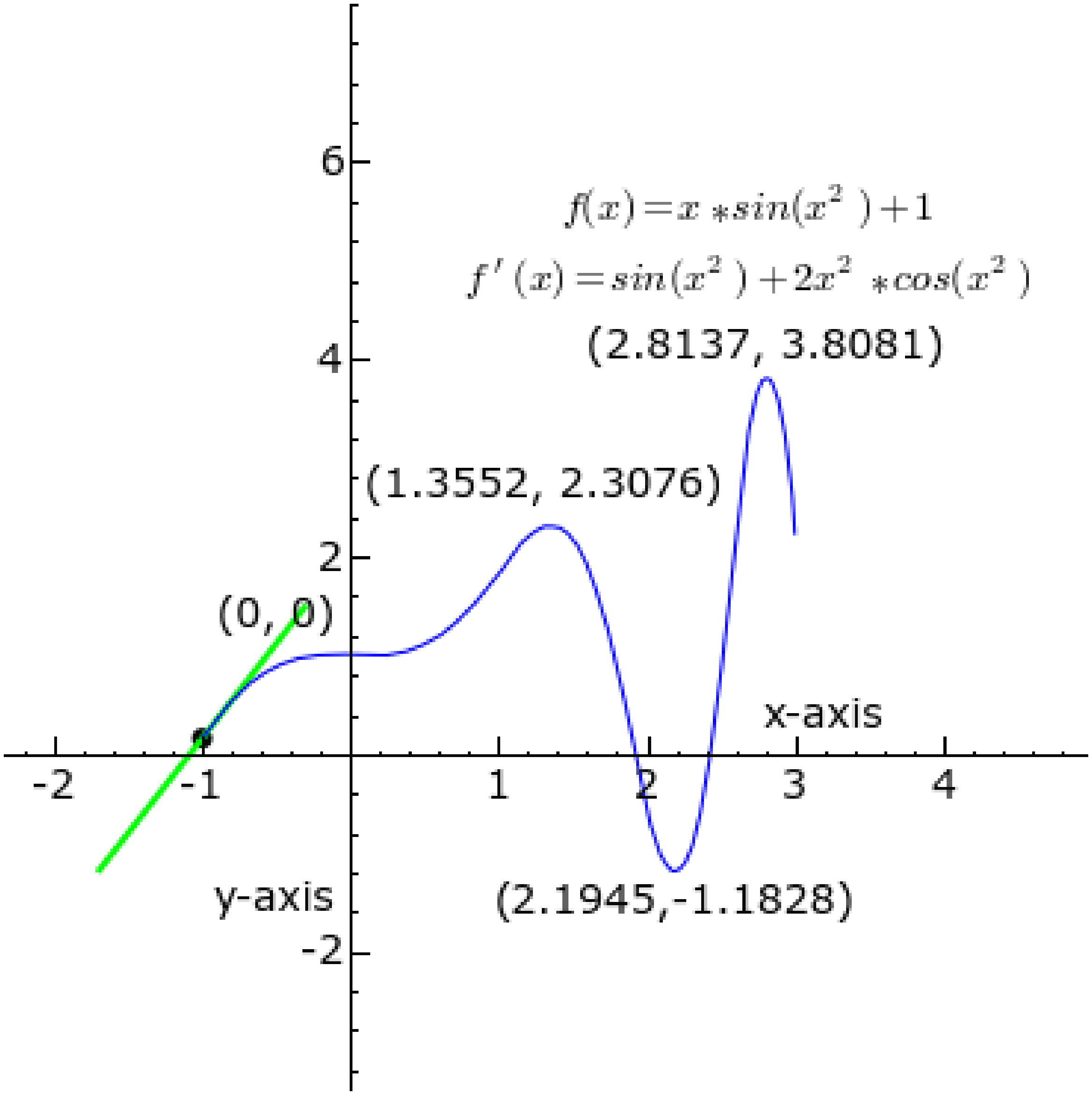
- The calculus method to find the maximum of a function makes use of the
  first derivative
  - Slope of line that is tangent to a point on the curve

- When the first derivative is zero (slope is flat), the maximum of the function
  is found
  - Could also be at a minimum – but our functions will
    be inverted Us (convex)



First Derivative = Tangent Line

From:
Wikipedia



The First Derivative for the Sample Mean

- Using calculus, we can find the first derivative for the mean from our normal distribution example (the slope of the tangent line for any value of μ_x):

$$\frac{\partial \log L}{\partial \mu_x} = \frac{1}{\sigma_x^2}\left(-N\mu_x + \sum_{p=1}^{N} x_p\right)$$

- To find where the maximum is, we set this equal to zero and solve for μ_x (giving us an ML estimate $\hat{\mu}_x$):

$$\frac{1}{\sigma_x^2}\left(-N\mu_x + \sum_{p=1}^{N} x_p\right) = 0 \rightarrow \hat{\mu}_x = \frac{1}{N}\sum_{p=1}^{N} x_p$$

The First Derivative for the Sample Variance

- Using calculus, we can find the first derivative for the variance (slope of the tangent line for any value of σ_x^2):

$$\frac{\partial \log L}{\partial \sigma_x^2} = -\frac{\mathrm{N}}{2\sigma_x^2} + \sum_{p=1}^{N} \frac{\left(x_p - \mu_x\right)^2}{2\sigma_x^4}$$

- To find where the maximum is, we set this equal to zero and solve for σ_x^2 (giving us an ML estimate $\hat{\sigma}_x^2$):

$$-\frac{\mathrm{N}}{2\sigma_x^2} + \sum_{p=1}^{N} \frac{\left(x_p - \mu_x\right)^2}{2\sigma_x^4} = 0 \rightarrow \hat{\sigma}_x^2 = \frac{1}{\mathrm{N}} \sum_{p=1}^{N} \left(x_p - \mu_x\right)^2$$

  - Where the $\frac{1}{N}$ version of the variance/standard deviation comes from

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

- Variance of the sample mean:

$$\frac{\partial^2 \log L}{\partial \mu_x^2} = \frac{-N}{\sigma_x^2} \rightarrow Var(\hat{\mu}_x) = \frac{\sigma_x^2}{N}$$

ML ESTIMATION OF GLMS: THE NLME/LME4 PACKAGES IN R

Maximum Likelihood Estimation for GLMs in R: NLME and LME4

- Maximum likelihood estimation of GLMs can be performed in ;a NLME and LME4 packages in R
  - Also: SAS PROC MIXED; XTMIXED in Stata
  - lavaan is be used for ML estimation in a structural equation modeling format

- These packages will grow in value to you as time goes on: most advanced analyses can be run with these programs:
  - Multilevel models
  - Repeated measures
  - Some factor analysis models

- The **MIXED** part of Non-Linear/Linear Mixed Effects refers to the type of model it can estimate: **General Linear Mixed Models**
  - Mixed models *extend* the GLM to be able to model dependency between observations (either within a person or within a group, or both)

Likelihood Functions in NLME and LME4

- Both packages use a common (but very general) log-likelihood function based on the GLM: the conditional distribution of Y given **X**

$$f\left(Y_p \middle| X_p, Z_p\right) \sim N\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p, \sigma_e^2\right)$$

  - Y is normally distributed conditional on the values of the predictors

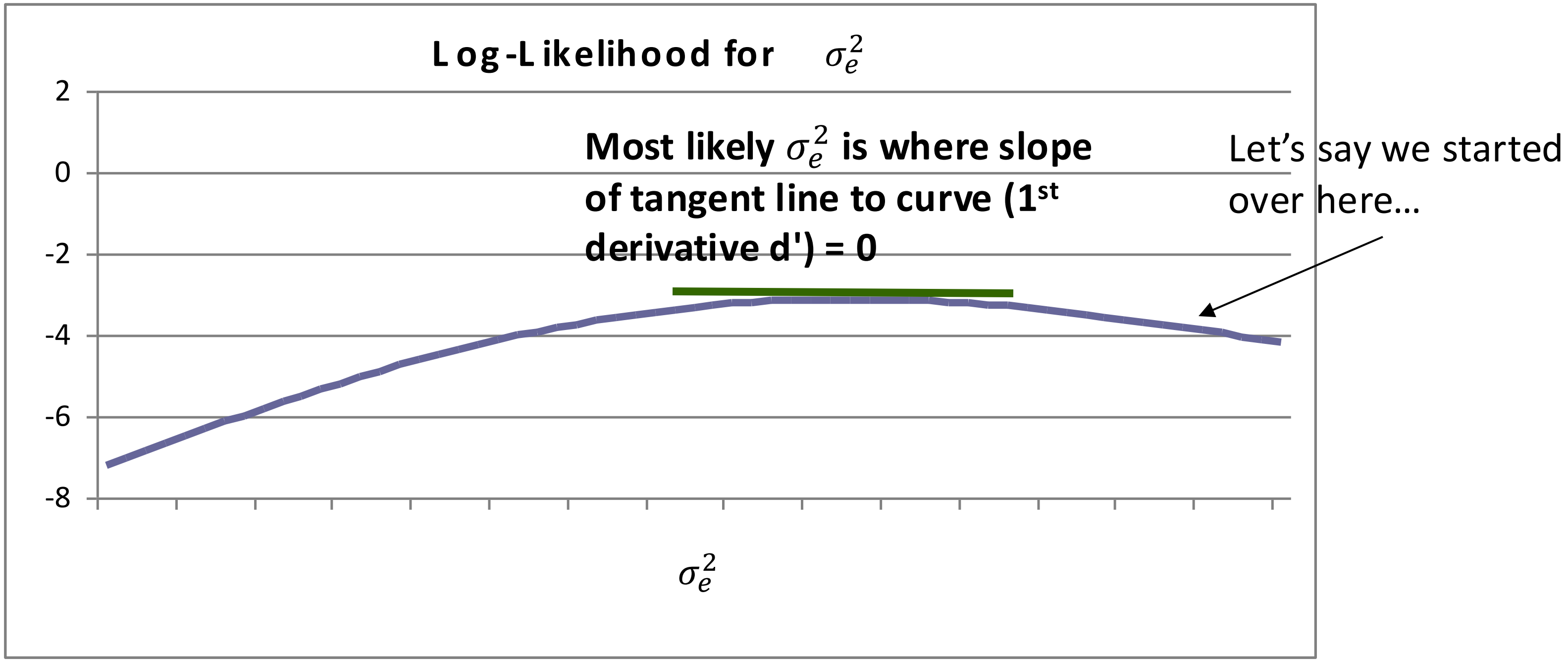
- The log likelihood for Y is then

$$\log L = \log L(\sigma_e^2 | x_1, \ldots, x_N) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma_e^2) - \sum_{p=1}^{N} \frac{\left(Y_p - \hat{Y}_p\right)^2}{2\sigma_e^2}$$

- Furthermore, there is a **closed form** (a set of equations) for the fixed effects (and thus $\hat{Y}_p$) for any possible value of σ_e^2
  - The programs seek to find σ_e^2 at the maximum of the log likelihood function – and after that finds everything else from equations
  - Begins with a naïve guess…then uses Newton-Raphson to find maximum

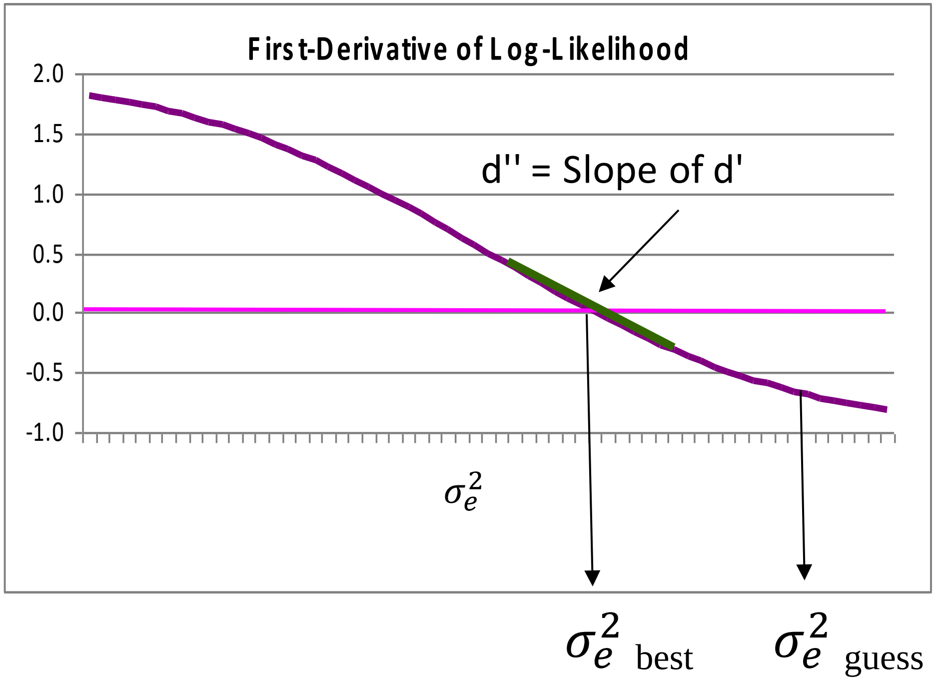
σ_e^2 Estimation via Newton Raphson

- We could calculate the likelihood over wide range of σ_e^2 for each person and plot those log likelihood values to see where the peak is…
  - But we have lives to lead, so we can solve it mathematically instead by finding where the slope of the likelihood function (the 1st derivative, d') = 0 (its peak)
- Step 1: Start with a guess of σ_e^2, **calculate 1st derivative d'** of the log likelihood with respect to σ_e^2 at that point
  - Are we there (d' = 0) yet? Positive d' = too low, negative d' = too high



σ_e^2 Estimation via Newton Raphson

- Step 2: **Calculate the 2nd derivative** (slope of slope, d'') at that point
  - Tells us **how far off we are**, and is used to figure out how much to adjust by
  - d'' will always be negative as approach top, but d' can be positive or negative
- Calculate new guess of σ_e^2 : σ_e^2 new = σ_e^2 old – (d'/d'')
  - If (d'/d'') < 0 → σ_e^2 increases
    If (d'/d'') > 0 → σ_e^2 decreases
    If (d'/d'') = 0 then you are done
- **2nd derivative d'' also tells you how *good* of a peak you have**
  - Need to know where your best σ_e^2 is (at d'=0), as well as how precise it is (from d'')
  - If the function is flat, d'' will be smallish
  - **Want large d'' because 1/SQRT(d'') = σ_e^2's SE**



Trying It Out: Using NLME with Our Example Data

- For now, we will know NLME to be largely like LM
  - Even the glht function from MULTCOMP works the same

- The first model will be the empty model where IQ is the DV
  - Linking NLME's gls function to our previous set of slides
  - After that, we will replicate a previous analysis : predicting Performance from IQ
  - What we are estimating is $\sigma_x^2 = \sigma_e^2$ (the variance of IQ, used in the likelihood function) and $\beta_0^{IQ} = \mu_x$ (the mean IQ, found from equations)

- The NLME function we will use is called gls
  - The empty model is:
  - The only difference from the lm function is the inclusion of the option method="ML"

```
> model01 = gls(iq~1,data=data01,method="ML")
> summary(model01)
Generalized least squares fit by maximum likelihood
  Model: iq ~ 1
  Data: data01
      AIC      BIC    logLik
  25.4122 24.63108 -10.7061

Coefficients:
            Value Std.Error  t-value p-value
(Intercept) 114.4  1.029563 111.1151       0
```

The Basics of PROC MIXED Output

- Here are some of the names of the object returned by the gls function:

```
> names(model01)
 [1] "modelStruct"  "dims"         "contrasts"    "coefficients" "varBeta"      "sigma"        "apVar"        "logLik"       "numIter"      "groups"       "call"
[12] "method"       "fitted"       "residuals"    "parAssign"    "na.action"
```

- Dimensions: see Subjects and Max Obs Per Subject

```
> model01$dims
$N
[1] 5

$p
[1] 1

$REML
[1] 0
```

- Note: if no results – no convergence – bad news
  - If you do not have the MLE all the good things about the MLE don't apply
    to your results

Further Unpacking Output

- The estimated σ_e is shown in the summary() function

```
Residual standard error: 2.059126
Degrees of freedom: 5 total; 4 residual
```

  - Note: R found the same estimate of σ_e^2 as we did – just reported as the un-squared version
  - Also: the SE of σ_e^2 is the SD of a variance – not displayed in this package but does happen in others

- The Information Criteria section shows statistics that can be used for model comparisons

```
     AIC      BIC    logLik
 25.4122 24.63108 -10.7061
```

Finally…the Fixed Effects

- The coefficients (also referred to as fixed effects) are where the estimated regression slopes are listed – here $\beta_0^{IQ} = \mu_x$

```
Coefficients:
               Value Std.Error  t-value p-value
(Intercept) 114.4   1.029563 111.1151       0
```

  - This also is the value we estimated in our example from before

- Not listed: traditional ANOVA table with Sums of Squares, Mean Squares, and F statistics
  - The Mean Square Error is no longer the estimate of σ_e^2: this comes directly from the model estimation algorithm itself
  - The traditional R^2 change test also changes under ML estimation

```
> anova(model01)
Denom. DF: 4
            numDF  F-value p-value
(Intercept)     1 9877.253  <.0001
```

USEFUL PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATES

Useful Properties of MLEs

- Next, we demonstrate three useful properties of MLEs (not just for GLMs)
  - Likelihood ratio (aka Deviance) tests
  - Wald tests
  - Information criteria

- To do so, we will consider our example where we wish to predict job performance from IQ (but will now center IQ at its mean of 114.4)

- We will estimate two models, both used to demonstrate how ML estimation differs slightly from LS estimation for GLMs
  - Empty model predicting just performance: $Y_p = \beta_0 + e_p$
  - Model where mean centered IQ predicts performance:

$$Y_p = \beta_0 + \beta_1(IQ - 114.4) + e_p$$

R gls Syntax

- Syntax for the empty model predicting performance:

```
#emtpy model predicting performance
model02 = gls(perf~1,data=data01,method="ML")
summary(model02)
```

- Syntax for the conditional model where mean centered IQ predicts performance:

```
#centering IQ at mean of 114.4
data01$iq114 = data01$iq-114.4

#Regression with ML:
model03a = gls(perf~iq114,data=data01,method="ML")
summary(model03a)
```

- Questions in comparing between the two models:
  - How do we test the hypothesis that IQ predicts performance?
    - Likelihood ratio tests (can be multiple parameter/degree-of-freedom)
    - Wald tests (usually for one parameter)
  - If IQ does significantly predict performance, what percentage of variance in performance does it account for?
    - Relative change in σ_e^2 from empty model to conditional model

Likelihood Ratio (Deviance) Tests

- The likelihood value from MLEs can help to statistically test competing models assuming the models are nested

- 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 null hypothesis that IQ did not predict performance:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

- The difference between the empty model and the conditional model is one parameter
  - Null model: one intercept β_0 and one residual variance σ_e^2 estimated = 2 parameters
  - Alternative model: one intercept β_0, one slope β_1, and one residual variance σ_e^2 estimated = 3 parameters

- Difference in parameters: 3-2 = 1 (will be degrees of freedom)

LRT/Deviance Test Procedure

- Step #1: estimate null model (get -2*log likelihood)

```
> summary(model02)
Generalized least squares fit by maximum likelihood
  Model: perf ~ 1
  Data: data01
       AIC      BIC    logLik
  25.31696 24.53584 -10.65848
```

- Step #2: estimate alternative model (get -2*log likelihood)

```
> summary(model03a)
Generalized least squares fit by maximum likelihood
  Model: perf ~ iq114
  Data: data01
       AIC      BIC    logLik
  12.92641 11.75472 -3.463204
```

- Step #3: compute test statistic

$$D = -2 \times (\log L_{H0} - \log L_{HA}) = -2 \times (-10.658 - 3.463) = 14.4$$

- Step #4: calculate p-value from Chi-Square Distribution with 1 DF
  - I used the pchisq() function (with the upper tail)
  - p-value = 0.000148

```
> lrt02v03
[1] 14.39055
> pchisq(lrt02v03,df=1,lower.tail=FALSE)
[1] 0.0001485457
```

- Inference: the regression slope for IQ was significantly different from zero -- we prefer our alternative model to the null model

- Interpretation: IQ significantly predicts performance

Likelihood Ratio Tests in R

- R makes this process much easier by embedding likelihood ratio tests in the ANOVA() function for nested models:

```
> anova(model02,model03a)
         Model df      AIC      BIC     logLik   Test  L.Ratio p-value
model02      1  2 25.31696 24.53584 -10.658480                        
model03a     2  3 12.92641 11.75472  -3.463204 1 vs 2 14.39055   1e-04
```

Wald Tests (Usually 1 DF Tests in Software)

- 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 (Z)
  - Exception: bounded parameters can have issues (variances)

- We can further add that variances are estimated, switching this standard normal distribution to a t distribution (R does this for us for some packages)
  - Note: some don't like calling this a "true" Wald test

Wald Test Example

- We could have used a Wald test to compare between the empty and conditional model, or:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

- R provides this for us in the from the summary() function:

```
> summary(model03a)
Generalized least squares fit by maximum likelihood
  Model: perf ~ iq114
  Data: data01
       AIC      BIC    logLik
  12.92641 11.75472 -3.463204

Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 12.800000 0.2792623 45.83505  0.0000
iq114        0.962264 0.1356218  7.09521  0.0058
```

  - Note: these estimates are identical to the glht estimates from last week

- Here, the slope estimate has a t-test statistic value of 7.095 (p = .0058), meaning we would reject our null hypothesis

- Typically, Wald tests are used for one additional parameter
  - Here, one slope

Model Comparison with R^2

- To compute an R^2, we use the ML estimates of σ_e^2:
  - Empty model: $\sigma_e^2 = 4.160\ (2.631)$
  - Conditional model: $\sigma_e^2 = 0.234\ (0.148)$

- The R^2 for variance in performance accounted for by IQ is:

$$R^2 = \frac{4.160 - 0.234}{4.160} = .944$$

  - Hall of fame worthy

```
> (model02$sigma^2-model03a$sigma^2)/model02$sigma^2
[1] 0.9062651
>
```

Information Criteria

- Information criteria are statistics that help determine the relative fit of a model for non-nested models
  - Comparison is fit-versus-parsimony
- R reports a set of criteria (from conditional model)

```
> anova(model02,model03a)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model02      1  2 25.31696 24.53584 -10.658480
model03a     2  3 12.92641 11.75472  -3.463204 1 vs 2 14.39055   1e-04
```

  - Each uses -2*log-likelihood as a base
    - Choice of statistic is **very** arbitrary and depends on field
- Best model is one with *smallest* value
- Note: don’t use information criteria for nested models
  - LRT/Deviance tests are more powerful

How ML and LS Estimation of GLMs Differ

- You may have recognized that the ML and the LS estimates of the fixed effects were identical
  - And for these models, they will be

- Where they differ is in their estimate of the residual variance σ_e^2:
  - From Least Squares (MSE): $\sigma_e^2 = 0.390$ (no SE)
  - From ML (model parameter): $\sigma_e^2 = 0.234$ (no SE in R)

- The ML version uses a **biased estimate** of σ_e^2 (it is too small)

- Because σ_e^2 plays a role in all SEs, the Wald tests differed from LS and ML

- Troubled by this? Don't be: a fix will come in a few weeks…
  - HINT: use method="REML" rather than method="ML" in gls()

WRAPPING UP

Wrapping Up

- This lecture was our first pass at maximum likelihood estimation
- The topics discussed today apply to all statistical models, not just GLMs
- Maximum likelihood estimation of GLMs helps when the basic assumptions are obviously violated
  - Independence of observations
  - Homogeneous σ_e^2
  - Conditional normality of Y (normality of error terms)
- Maximum likelihood is a primary method for handling missing data that are assumed missing at random
  - But…must be full information version…for which we may rely on structural equation modeling methods