Maximum likelihood estimation

Peter Hoff
Duke STA 610

Estimation frameworks

Review of ML estimation

ML for HNM

Method of moments

```
aovfit<-anova(lm(y~as.factor(g)) )

MSG<-aovfit[1,3]
MSE<-aovfit[2,3]

t2<-(MSG-MSE)/n

s2<-MSE

t2 ; sqrt(t2)

##         1
## 0.3840768
##         1
## 0.6197393

s2 ; sqrt(s2)

## [1] 1.787206
## [1] 1.336864

mean(y)

## [1] 16.3064
```

Maximum likelihood estimation

```
lmer                    package:lme4                    R Documentation

Fit Linear Mixed-Effects Models

Description:

     Fit a linear mixed-effects model (LMM) to data, via REML or
     maximum likelihood.

Usage:

     lmer(formula, data = NULL, REML = TRUE, control = lmerControl(),
          start = NULL, verbose = 0L, subset, weights, na.action,
          offset, contrasts = NULL, devFunOnly = FALSE)
```

```r
library(lme4)

lmer(y~1+(1|g))

## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 + (1 | g)
## REML criterion at convergence: 177.9876
## Random effects:
##  Groups   Name        Std.Dev.
##  g        (Intercept) 0.6197
##  Residual             1.3369
## Number of obs: 50, groups:  g, 10
## Fixed Effects:
## (Intercept)
##       16.31
```

A more complicated example

```
nels[1:10,]

##     school enroll flp public urbanicity hwh   ses mscore
## 1     1011      5   3      1      urban   2 -0.23  52.11
## 2     1011      5   3      1      urban   0  0.69  57.65
## 3     1011      5   3      1      urban   4 -0.68  66.44
## 4     1011      5   3      1      urban   5 -0.89  44.68
## 5     1011      5   3      1      urban   3 -1.28  40.57
## 6     1011      5   3      1      urban   5 -0.93  35.04
## 7     1011      5   3      1      urban   1  0.36  50.71
## 8     1011      5   3      1      urban   4 -0.24  66.17
## 10    1011      5   3      1      urban   8 -1.07  46.17
## 11    1011      5   3      1      urban   2 -0.10  58.76
```

A more complicated example

$$y_{i,j} = (\beta_0 + \beta_{0,j}) + \beta_1 \times \mathsf{flp}_j + \beta_2 \times \mathsf{enroll}_j + (\beta_3 + \beta_{3,j}) \times \mathsf{ses}_{i,j} + \epsilon_{i,j}$$

```
fit<-lmer(mscore~flp+enroll+ses+(ses|school),data=nels,REML=FALSE)
```

```r
summary(fit)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: mscore ~ flp + enroll + ses + (ses | school)
##    Data: nels
##
##      AIC      BIC   logLik deviance df.resid
##  92397.7  92457.5 -46190.9  92381.7    12966
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9797 -0.6399  0.0180  0.6681  4.5053
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  school   (Intercept)  9.004   3.001
##           ses          1.600   1.265    0.05
##  Residual             67.260   8.201
## Number of obs: 12974, groups:  school, 684
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 55.429341   0.402910 137.573
## flp         -2.411521   0.185312 -13.013
## enroll       0.007095   0.082024   0.087
## ses          4.116881   0.125381  32.835
##
## Correlation of Fixed Effects:
##        (Intr) flp    enroll
## flp    -0.815
## enroll -0.300 -0.193
## ses    -0.202  0.212  0.007
```

Models and inference

A *statistical model* is a collection of probability distributions for observed data:

$$\mathcal{P} = \{p(y|\gamma), \gamma \in \Gamma\}$$

- y is the data;
- Γ is the set of parameter values;
- $p(y|\gamma)$ is a probability (density) for each $\gamma \in \Gamma$.

Models and inference

A *statistical model* is a collection of probability distributions for observed data:

$$\mathcal{P} = \{p(y|\gamma), \gamma \in \Gamma\}$$

- y is the data;
- Γ is the set of parameter values;
- $p(y|\gamma)$ is a probability (density) for each $\gamma \in \Gamma$.

Models and inference

A *statistical model* is a collection of probability distributions for observed data:

$$\mathcal{P} = \{p(y|\gamma), \gamma \in \Gamma\}$$

- y is the data;
- Γ is the set of parameter values;
- $p(y|\gamma)$ is a probability (density) for each $\gamma \in \Gamma$.

Models and inference

A *statistical model* is a collection of probability distributions for observed data:

$$\mathcal{P} = \{p(y|\gamma), \gamma \in \Gamma\}$$

- y is the data;
- Γ is the set of parameter values;
- $p(y|\gamma)$ is a probability (density) for each $\gamma \in \Gamma$.

Models and inference

A *statistical model* is a collection of probability distributions for observed data:

$$\mathcal{P} = \{p(y|\gamma), \gamma \in \Gamma\}$$

- y is the data;
- Γ is the set of parameter values;
- $p(y|\gamma)$ is a probability (density) for each $\gamma \in \Gamma$.

Models and inference

A *statistical model* is a collection of probability distributions for observed data:

$$\mathcal{P} = \{p(y|\gamma), \gamma \in \Gamma\}$$

- y is the data;
- Γ is the set of parameter values;
- $p(y|\gamma)$ is a probability (density) for each $\gamma \in \Gamma$.

Example: Normal model

For example, the normal model is

$$\{p(y|\theta,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\{-(y-\theta)^2/(2\sigma^2)\}, \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}.$$

- y is a single scalar data value;
- $\gamma = \{\theta, \sigma^2\}$ is the parameter (or are the parameters);
- $\Gamma = \mathbb{R} \times \mathbb{R}^+$ is the set of possible parameter values;
- $p(y|\theta,\sigma^2)$ is the normal probability density for each θ, σ^2.

Example: Normal model

For example, the normal model is

$$\{p(y|\theta,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\{-(y-\theta)^2/(2\sigma^2)\}, \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}.$$

- y is a single scalar data value;
- $\gamma = \{\theta, \sigma^2\}$ is the parameter (or are the parameters);
- $\Gamma = \mathbb{R} \times \mathbb{R}^+$ is the set of possible parameter values;
- $p(y|\theta,\sigma^2)$ is the normal probability density for each θ, σ^2.

Example: Normal model

For example, the normal model is

$$\{p(y|\theta,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\{-(y-\theta)^2/(2\sigma^2)\}, \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}.$$

- y is a single scalar data value;
- $\gamma = \{\theta, \sigma^2\}$ is the parameter (or are the parameters);
- $\Gamma = \mathbb{R} \times \mathbb{R}^+$ is the set of possible parameter values;
- $p(y|\theta, \sigma^2)$ is the normal probability density for each θ, σ^2.

Example: Normal model

For example, the normal model is

$$\{p(y|\theta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-(y-\theta)^2/(2\sigma^2)\}, \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}.$$

- y is a single scalar data value;
- $\gamma = \{\theta, \sigma^2\}$ is the parameter (or are the parameters);
- $\Gamma = \mathbb{R} \times \mathbb{R}^+$ is the set of possible parameter values;
- $p(y|\theta, \sigma^2)$ is the normal probability density for each θ, σ^2.

Example: Normal model

For example, the normal model is

$$\{p(y|\theta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-(y-\theta)^2/(2\sigma^2)\}, \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}.$$

- y is a single scalar data value;
- $\gamma = \{\theta, \sigma^2\}$ is the parameter (or are the parameters);
- $\Gamma = \mathbb{R} \times \mathbb{R}^+$ is the set of possible parameter values;
- $p(y|\theta, \sigma^2)$ is the normal probability density for each θ, σ^2.

Example: Normal model

For example, the normal model is

$$\{p(y|\theta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-(y-\theta)^2/(2\sigma^2)\}, \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}.$$

- y is a single scalar data value;
- $\gamma = \{\theta, \sigma^2\}$ is the parameter (or are the parameters);
- $\Gamma = \mathbb{R} \times \mathbb{R}^+$ is the set of possible parameter values;
- $p(y|\theta, \sigma^2)$ is the normal probability density for each θ, σ^2.

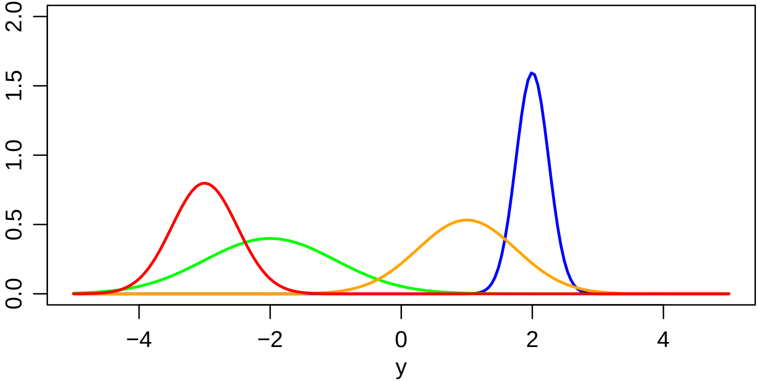
Example: Normal model

For example, the normal model is

$$\{p(y|\theta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-(y-\theta)^2/(2\sigma^2)\}, \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}.$$

- y is a single scalar data value;
- $\gamma = \{\theta, \sigma^2\}$ is the parameter (or are the parameters);
- $\Gamma = \mathbb{R} \times \mathbb{R}^+$ is the set of possible parameter values;
- $p(y|\theta, \sigma^2)$ is the normal probability density for each θ, σ^2.

Example: Normal model



Model-based inference

Model-based statistical inference involves

Estimation: Obtain a value $\hat{\gamma} \in \Gamma$ that "best" represents the population.

Inference: Evaluate the plausibility of other γ values.

Inference includes things like: confidence intervals, hypotheses tests.

Likelihood-based statistical inference:

- a type of model based inference;
- estimation and inference are based on the likelihood function.

Model-based inference

Model-based statistical inference involves

Estimation: Obtain a value $\hat{\gamma} \in \Gamma$ that "best" represents the population.

Inference: Evaluate the plausibility of other γ values.

Inference includes things like: confidence intervals, hypotheses tests.

Likelihood-based statistical inference:

- a type of model based inference;
- estimation and inference are based on the likelihood function.

Model-based inference

Model-based statistical inference involves

Estimation: Obtain a value $\hat{\gamma} \in \Gamma$ that "best" represents the population.

Inference: Evaluate the plausibility of other γ values.

Inference includes things like: confidence intervals, hypotheses tests.

Likelihood-based statistical inference:

- a type of model based inference;
- estimation and inference are based on the likelihood function.

Model-based inference

Model-based statistical inference involves

Estimation: Obtain a value $\hat{\gamma} \in \Gamma$ that "best" represents the population.

Inference: Evaluate the plausibility of other γ values.

Inference includes things like: confidence intervals, hypotheses tests.

Likelihood-based statistical inference:

- a type of model based inference;
- estimation and inference are based on the likelihood function.

Model-based inference

Model-based statistical inference involves

Estimation: Obtain a value $\hat{\gamma} \in \Gamma$ that "best" represents the population.

Inference: Evaluate the plausibility of other γ values.

Inference includes things like: confidence intervals, hypotheses tests.

Likelihood-based statistical inference:

- a type of model based inference;
- estimation and inference are based on the likelihood function.

Model-based inference

Model-based statistical inference involves

Estimation: Obtain a value $\hat{\gamma} \in \Gamma$ that “best” represents the population.

Inference: Evaluate the plausibility of other γ values.

Inference includes things like: confidence intervals, hypotheses tests.

Likelihood-based statistical inference:

- a type of model based inference;
- estimation and inference are based on the likelihood function.

Model-based inference

Model-based statistical inference involves

Estimation: Obtain a value $\hat{\gamma} \in \Gamma$ that "best" represents the population.

Inference: Evaluate the plausibility of other γ values.

Inference includes things like: confidence intervals, hypotheses tests.

Likelihood-based statistical inference:

- a type of model based inference;
- estimation and inference are based on the likelihood function.

Model-based inference

Model-based statistical inference involves

Estimation: Obtain a value $\hat{\gamma} \in \Gamma$ that "best" represents the population.

Inference: Evaluate the plausibility of other γ values.

Inference includes things like: confidence intervals, hypotheses tests.

Likelihood-based statistical inference:

- a type of model based inference;
- estimation and inference are based on the likelihood function.

Model-based inference

Model-based statistical inference involves

Estimation: Obtain a value $\hat{\gamma} \in \Gamma$ that "best" represents the population.

Inference: Evaluate the plausibility of other γ values.

Inference includes things like: confidence intervals, hypotheses tests.

Likelihood-based statistical inference:

- a type of model based inference;
- estimation and inference are based on the likelihood function.

Joint probability of the data

Independent events: Recall if A and B are independent events,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B).$$

Independent observations: If y_1 and y_2 are independent observations, then

$$\begin{aligned} p_{y_1 y_2}(y_1, y_2|\gamma) &= p(y_1|\gamma) \times p(y_2|\gamma) \\ &= \prod_{i=1}^{2} p(y_i|\gamma). \end{aligned}$$

Independent sample: If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are independent observations, then

$$\begin{aligned} p_{\boldsymbol{y}}(\boldsymbol{y}|\gamma) &= p(y_1|\gamma) \times \cdots \times p(y_n|\gamma) \\ &= \prod_{i=1}^{n} p(y_i|\gamma). \end{aligned}$$

$p_{\mathbf{y}}(\mathbf{y}|\gamma)$, as a function of $\mathbf{y}$, is the *joint probability (density)* of the data.

Joint probability of the data

Independent events: Recall if A and B are independent events,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B).$$

Independent observations: If y_1 and y_2 are independent observations, then

$$\begin{aligned} p_{y_1 y_2}(y_1, y_2|\gamma) &= p(y_1|\gamma) \times p(y_2|\gamma) \\ &= \prod_{i=1}^{2} p(y_i|\gamma). \end{aligned}$$

Independent sample: If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are independent observations, then

$$\begin{aligned} p_{\boldsymbol{y}}(\boldsymbol{y}|\gamma) &= p(y_1|\gamma) \times \cdots \times p(y_n|\gamma) \\ &= \prod_{i=1}^{n} p(y_i|\gamma). \end{aligned}$$

$p_{\mathbf{y}}(\mathbf{y}|\gamma)$, as a function of $\mathbf{y}$, is the *joint probability (density)* of the data.

Joint probability of the data

Independent events: Recall if A and B are independent events,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B).$$

Independent observations: If y_1 and y_2 are independent observations, then

$$\begin{aligned} p_{y_1 y_2}(y_1, y_2|\gamma) &= p(y_1|\gamma) \times p(y_2|\gamma) \\ &= \prod_{i=1}^{2} p(y_i|\gamma). \end{aligned}$$

Independent sample: If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are independent observations, then

$$\begin{aligned} p_{\boldsymbol{y}}(\boldsymbol{y}|\gamma) &= p(y_1|\gamma) \times \cdots \times p(y_n|\gamma) \\ &= \prod_{i=1}^{n} p(y_i|\gamma). \end{aligned}$$

$p_{\mathbf{y}}(\mathbf{y}|\gamma)$, as a function of $\mathbf{y}$, is the *joint probability (density)* of the data.

Joint probability of the data

Independent events: Recall if A and B are independent events,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B).$$

Independent observations: If y_1 and y_2 are independent observations, then

$$\begin{aligned} p_{y_1 y_2}(y_1, y_2|\gamma) &= p(y_1|\gamma) \times p(y_2|\gamma) \\ &= \prod_{i=1}^{2} p(y_i|\gamma). \end{aligned}$$

Independent sample: If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are independent observations, then

$$\begin{aligned} p_{\mathbf{y}}(\mathbf{y}|\gamma) &= p(y_1|\gamma) \times \cdots \times p(y_n|\gamma) \\ &= \prod_{i=1}^{n} p(y_i|\gamma). \end{aligned}$$

$p_{\mathbf{y}}(\mathbf{y}|\gamma)$, as a function of $\mathbf{y}$, is the *joint probability (density)* of the data.

Joint probability of the data

Independent events: Recall if A and B are independent events,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B).$$

Independent observations: If y_1 and y_2 are independent observations, then

$$\begin{aligned} p_{y_1 y_2}(y_1, y_2|\gamma) &= p(y_1|\gamma) \times p(y_2|\gamma) \\ &= \prod_{i=1}^{2} p(y_i|\gamma). \end{aligned}$$

Independent sample: If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are independent observations, then

$$\begin{aligned} p_{\boldsymbol{y}}(\boldsymbol{y}|\gamma) &= p(y_1|\gamma) \times \cdots \times p(y_n|\gamma) \\ &= \prod_{i=1}^{n} p(y_i|\gamma). \end{aligned}$$

$p_{\mathbf{y}}(\mathbf{y}|\gamma)$, as a function of $\mathbf{y}$, is the *joint probability (density)* of the data.

Joint probability of the data

Independent events: Recall if A and B are independent events,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B).$$

Independent observations: If y_1 and y_2 are independent observations, then

$$\begin{aligned} p_{y_1 y_2}(y_1, y_2|\gamma) &= p(y_1|\gamma) \times p(y_2|\gamma) \\ &= \prod_{i=1}^{2} p(y_i|\gamma). \end{aligned}$$

Independent sample: If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are independent observations, then

$$\begin{aligned} p_{\boldsymbol{y}}(\boldsymbol{y}|\gamma) &= p(y_1|\gamma) \times \cdots \times p(y_n|\gamma) \\ &= \prod_{i=1}^{n} p(y_i|\gamma). \end{aligned}$$

$p_{\mathbf{y}}(\mathbf{y}|\gamma)$, as a function of $\mathbf{y}$, is the *joint probability (density)* of the data.

Joint probability of the data

Independent events: Recall if A and B are independent events,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B).$$

Independent observations: If y_1 and y_2 are independent observations, then

$$\begin{aligned} p_{y_1 y_2}(y_1, y_2|\gamma) &= p(y_1|\gamma) \times p(y_2|\gamma) \\ &= \prod_{i=1}^{2} p(y_i|\gamma). \end{aligned}$$

Independent sample: If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are independent observations, then

$$\begin{aligned} p_{\boldsymbol{y}}(\boldsymbol{y}|\gamma) &= p(y_1|\gamma) \times \cdots \times p(y_n|\gamma) \\ &= \prod_{i=1}^{n} p(y_i|\gamma). \end{aligned}$$

$p_{\mathbf{y}}(\mathbf{y}|\gamma)$, as a function of $\mathbf{y}$, is the *joint probability (density)* of the data.

Joint probability of the data

Independent events: Recall if A and B are independent events,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B).$$

Independent observations: If y_1 and y_2 are independent observations, then

$$\begin{aligned} p_{y_1 y_2}(y_1, y_2|\gamma) &= p(y_1|\gamma) \times p(y_2|\gamma) \\ &= \prod_{i=1}^{2} p(y_i|\gamma). \end{aligned}$$

Independent sample: If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are independent observations, then

$$\begin{aligned} p_{\mathbf{y}}(\boldsymbol{y}|\gamma) &= p(y_1|\gamma) \times \cdots \times p(y_n|\gamma) \\ &= \prod_{i=1}^{n} p(y_i|\gamma). \end{aligned}$$

$p_{\mathbf{y}}(\mathbf{y}|\gamma)$, as a function of $\mathbf{y}$, is the *joint probability (density)* of the data.

Example: One sample normal model

$$y_1, \ldots, y_n \sim \text{ i.i.d. } N(\theta, \sigma^2)$$

For this model,

$$p(y_i|\theta, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-(y_i-\theta)^2/[2\sigma^2]}$$

$$p(y_1, \ldots, y_n|\theta, \sigma^2) = \prod_{i=1}^{n} p(y_i|\theta, \sigma^2)$$

$$= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp(-\sum (y_i - \theta)^2/[2\sigma^2])$$

Interpretation:

$p(\boldsymbol{y}|\theta)$ roughly quantifies how probable $\boldsymbol{y}$ is, for a particular (θ, σ^2).

Example: One sample normal model

$$y_1, \ldots, y_n \sim \text{ i.i.d. } N(\theta, \sigma^2)$$

For this model,

$$\begin{aligned} p(y_i|\theta,\sigma^2) &= (2\pi\sigma^2)^{-1/2} e^{-(y_i-\theta)^2/[2\sigma^2]} \\ p(y_1,\ldots,y_n|\theta,\sigma^2) &= \prod_{i=1}^{n} p(y_i|\theta,\sigma^2) \\ &= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp(-\sum(y_i-\theta)^2/[2\sigma^2]) \end{aligned}$$

Interpretation:
$p(\boldsymbol{y}|\theta)$ roughly quantifies how probable $\boldsymbol{y}$ is, for a particular (θ, σ^2).

Likelihood

The *likelihood* is the probability of the data as a function of the parameter:

$$L(\theta : \boldsymbol{y}) = p(\boldsymbol{y}|\theta)$$

The *maximum likelihood estimator* (*MLE*) is the value of θ that maximizes $L(\theta : \boldsymbol{y})$:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} L(\theta : \boldsymbol{y})$$

Likelihood

The *likelihood* is the probability of the data as a function of the parameter:

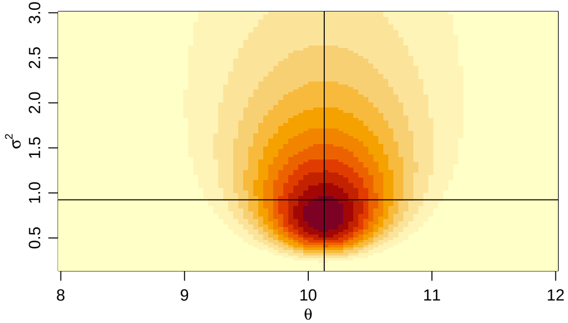
$$L(\theta : \boldsymbol{y}) = p(\boldsymbol{y}|\theta)$$

The *maximum likelihood estimator* (*MLE*) is the value of θ that maximizes $L(\theta : \boldsymbol{y})$:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} L(\theta : \boldsymbol{y})$$

Likelihood function

```
## some data
y

## [1]  9.373546 10.183643  9.164371 11.595281 10.329508

mean(y)

## [1] 10.12927

var(y)

## [1] 0.9235968
```



Log likelihoods

Likelihoods based on lots of data can give extreme numbers.

Alternatively, we can make inference with the *log-likelihood*:

If $\hat{\theta}$ maximizes $L(\theta : \boldsymbol{y})$ then it also maximizes $\log L(\theta : \boldsymbol{y}) = l(\theta : \boldsymbol{y})$.

$$\log p(\boldsymbol{y}|\theta, \sigma^2) = -\frac{1}{2}\left(n \log \sigma^2 + \sum_i (y_i - \theta)^2/\sigma^2\right) + c$$

Log likelihoods

Likelihoods based on lots of data can give extreme numbers.

Alternatively, we can make inference with the *log-likelihood*:

If $\hat{\theta}$ maximizes $L(\theta : \boldsymbol{y})$ then it also maximizes $\log L(\theta : \boldsymbol{y}) = l(\theta : \boldsymbol{y})$.

$$\log p(\boldsymbol{y}|\theta, \sigma^2) = -\frac{1}{2}\left(n \log \sigma^2 + \sum_i (y_i - \theta)^2/\sigma^2\right) + c$$

Log likelihoods

Likelihoods based on lots of data can give extreme numbers.

Alternatively, we can make inference with the *log-likelihood*:

If $\hat{\theta}$ maximizes $L(\theta : \boldsymbol{y})$ then it also maximizes $\log L(\theta : \boldsymbol{y}) = l(\theta : \boldsymbol{y})$.

$$\log p(\boldsymbol{y}|\theta, \sigma^2) = -\frac{1}{2}\left(n \log \sigma^2 + \sum_i (y_i - \theta)^2/\sigma^2\right) + c$$

Finding the MLE

Recall from calculus that the *tangent* or *derivative* of a function, at a local maximum, will be zero. This tells us how to find the MLE:

$$\hat{\gamma}_{MLE} \quad \text{satisfies} \quad \frac{d}{d\gamma} l(\gamma : \boldsymbol{y})|_{\gamma=\hat{\gamma}} = 0$$

Let's try this for the normal model. The derivative of the log-likelihood is

$$\frac{d}{d\gamma} l(\gamma : \boldsymbol{y}) = \begin{pmatrix} n(\bar{y} - \theta) \\ (-n/\sigma^2 + \sum_i (y_i - \theta)^2/\sigma^4)/2 \end{pmatrix}$$

The MLE of (θ, σ^2) is then

$$(\hat{\theta}, \hat{\sigma}^2) = \left(\bar{y}, \sum_i (y_i - \bar{y})^2/n \right).$$

So $\hat{\sigma}^2$ is biased for estimating σ^2.

Finding the MLE

Recall from calculus that the *tangent* or *derivative* of a function, at a local maximum, will be zero. This tells us how to find the MLE:

$$\hat{\gamma}_{MLE} \quad \text{satisfies} \quad \frac{d}{d\gamma} l(\gamma : \boldsymbol{y})|_{\gamma=\hat{\gamma}} = 0$$

Let's try this for the normal model. The derivative of the log-likelihood is

$$\frac{d}{d\gamma} l(\gamma : \boldsymbol{y}) = \begin{pmatrix} n(\bar{y} - \theta) \\ (-n/\sigma^2 + \sum_i (y_i - \theta)^2/\sigma^4)/2 \end{pmatrix}$$

The MLE of (θ, σ^2) is then

$$(\hat{\theta}, \hat{\sigma}^2) = \left(\bar{y}, \sum_i (y_i - \bar{y})^2/n\right).$$

So $\hat{\sigma}^2$ is biased for estimating σ^2.

Finding the MLE

Recall from calculus that the *tangent* or *derivative* of a function, at a local maximum, will be zero. This tells us how to find the MLE:

$$\hat{\gamma}_{MLE} \quad \text{satisfies} \quad \frac{d}{d\gamma} l(\gamma : \boldsymbol{y})|_{\gamma=\hat{\gamma}} = 0$$

Let's try this for the normal model. The derivative of the log-likelihood is

$$\frac{d}{d\gamma} l(\gamma : \boldsymbol{y}) = \begin{pmatrix} n(\bar{y} - \theta) \\ (-n/\sigma^2 + \sum_i (y_i - \theta)^2/\sigma^4)/2 \end{pmatrix}$$

The MLE of (θ, σ^2) is then

$$(\hat{\theta}, \hat{\sigma}^2) = \left(\bar{y}, \sum_i (y_i - \bar{y})^2 / n \right).$$

So $\hat{\sigma}^2$ is biased for estimating σ^2.

Information and precision

The precision of the MLE (how well it estimates the truth) depends on the *information* or second derivative of the log-likelihood.

Information: The *observed information* about γ is

$$I_n = -\frac{d^2}{d\gamma^2} l(\gamma : \boldsymbol{y})|_{\hat{\gamma}}$$

In many problems, the inverse of the information gives a variance estimate:

$$\mathrm{Var}[\hat{\gamma}] \approx I_n^{-1}$$
$$\mathrm{sd}(\hat{\gamma}) \approx 1/\sqrt{\mathrm{diag}(I_n)}$$

For the normal model,

$$I_n^{-1} = \begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -n/[2\hat{\sigma}^4]. \end{pmatrix}$$

So we have

$$\mathrm{Var}[\hat{\theta}] \approx \hat{\sigma}^2/n$$
$$\mathrm{Var}[\hat{\sigma}^2] \approx 2\hat{\sigma}^4/n$$

Information and precision

The precision of the MLE (how well it estimates the truth) depends on the *information* or second derivative of the log-likelihood.

Information: The *observed information* about γ is

$$I_n = -\frac{d^2}{d\gamma^2} l(\gamma : \boldsymbol{y})|_{\hat{\gamma}}$$

In many problems, the inverse of the information gives a variance estimate:

$$\mathrm{Var}[\hat{\gamma}] \approx I_n^{-1}$$

$$\mathrm{sd}(\hat{\gamma}) \approx 1/\sqrt{\mathrm{diag}(I_n)}$$

For the normal model,

$$I_n^{-1} = \begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -n/[2\hat{\sigma}^4]. \end{pmatrix}$$

So we have

$$\mathrm{Var}[\hat{\theta}] \approx \hat{\sigma}^2/n$$

$$\mathrm{Var}[\hat{\sigma}^2] \approx 2\hat{\sigma}^4/n$$

Information and precision

The precision of the MLE (how well it estimates the truth) depends on the *information* or second derivative of the log-likelihood.

Information: The *observed information* about γ is

$$I_n = -\frac{d^2}{d\gamma^2} l(\gamma : \boldsymbol{y})|_{\hat{\gamma}}$$

In many problems, the inverse of the information gives a variance estimate:

$$\mathrm{Var}[\hat{\gamma}] \approx I_n^{-1}$$
$$\mathrm{sd}(\hat{\gamma}) \approx 1/\sqrt{\mathrm{diag}(I_n)}$$

For the normal model,

$$I_n^{-1} = \begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -n/[2\hat{\sigma}^4]. \end{pmatrix}$$

So we have

$$\mathrm{Var}[\hat{\theta}] \approx \hat{\sigma}^2/n$$
$$\mathrm{Var}[\hat{\sigma}^2] \approx 2\hat{\sigma}^4/n$$

Information and precision

The precision of the MLE (how well it estimates the truth) depends on the *information* or second derivative of the log-likelihood.

Information: The *observed information* about γ is

$$I_n = -\frac{d^2}{d\gamma^2} l(\gamma : \boldsymbol{y})|_{\hat{\gamma}}$$

In many problems, the inverse of the information gives a variance estimate:

$$\mathrm{Var}[\hat{\gamma}] \approx I_n^{-1}$$
$$\mathrm{sd}(\hat{\gamma}) \approx 1/\sqrt{\mathrm{diag}(I_n)}$$

For the normal model,

$$I_n^{-1} = \begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -n/[2\hat{\sigma}^4]. \end{pmatrix}$$

So we have

$$\mathrm{Var}[\hat{\theta}] \approx \hat{\sigma}^2/n$$
$$\mathrm{Var}[\hat{\sigma}^2] \approx 2\hat{\sigma}^4/n$$

Information and precision

The precision of the MLE (how well it estimates the truth) depends on the *information* or second derivative of the log-likelihood.

Information: The *observed information* about γ is

$$I_n = -\frac{d^2}{d\gamma^2} l(\gamma : \boldsymbol{y})|_{\hat{\gamma}}$$

In many problems, the inverse of the information gives a variance estimate:

$$\begin{aligned} \mathrm{Var}[\hat{\gamma}] &\approx I_n^{-1} \\ \mathrm{sd}(\hat{\gamma}) &\approx 1/\sqrt{\mathrm{diag}(I_n)} \end{aligned}$$

For the normal model,

$$I_n^{-1} = \begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -n/[2\hat{\sigma}^4]. \end{pmatrix}$$

So we have

$$\begin{aligned} \mathrm{Var}[\hat{\theta}] &\approx \hat{\sigma}^2/n \\ \mathrm{Var}[\hat{\sigma}^2] &\approx 2\hat{\sigma}^4/n \end{aligned}$$

MLE for the hierarchical normal model

$$
\begin{aligned}
y_{i,j} &= \mu + a_j + \epsilon_{i,j} \\
\{\epsilon_{i,j}\} &\sim \text{iid } N(0, \sigma^2) \\
\{a_j\} &\sim \text{iid } N(0, \tau^2)
\end{aligned}
$$

Parameters to estimate:

- Fixed effects: μ
- Variance components: σ^2, τ^2
- Random effects: $a_1, \ldots, a_m$

Likelihood estimation focuses on estimation of $\theta = (\mu, \sigma^2, \tau^2)$

Alternative methods are required for estimation of $a_1, \ldots, a_m$.

MLE for the hierarchical normal model

$$
\begin{aligned}
y_{i,j} &= \mu + a_j + \epsilon_{i,j} \\
\{\epsilon_{i,j}\} &\sim \text{iid } N(0, \sigma^2) \\
\{a_j\} &\sim \text{iid } N(0, \tau^2)
\end{aligned}
$$

Parameters to estimate:

- Fixed effects: μ
- Variance components: σ^2, τ^2
- Random effects: $a_1, \ldots, a_m$

Likelihood estimation focuses on estimation of $\theta = (\mu, \sigma^2, \tau^2)$

Alternative methods are required for estimation of $a_1, \ldots, a_m$.

MLE for the hierarchical normal model

$$y_{i,j} = \mu + a_j + \epsilon_{i,j}$$
$$\{\epsilon_{i,j}\} \sim \text{iid } N(0, \sigma^2)$$
$$\{a_j\} \sim \text{iid } N(0, \tau^2)$$

Parameters to estimate:

- Fixed effects: μ
- Variance components: σ^2, τ^2
- Random effects: $a_1, \ldots, a_m$

Likelihood estimation focuses on estimation of $\theta = (\mu, \sigma^2, \tau^2)$

Alternative methods are required for estimation of $a_1, \ldots, a_m$.

MLE for the hierarchical normal model

$$
\begin{aligned}
y_{i,j} &= \mu + a_j + \epsilon_{i,j} \\
\{\epsilon_{i,j}\} &\sim \text{iid } N(0, \sigma^2) \\
\{a_j\} &\sim \text{iid } N(0, \tau^2)
\end{aligned}
$$

Parameters to estimate:

- Fixed effects: μ
- Variance components: σ^2, τ^2
- Random effects: $a_1, \ldots, a_m$

Likelihood estimation focuses on estimation of $\theta = (\mu, \sigma^2, \tau^2)$

Alternative methods are required for estimation of $a_1, \ldots, a_m$.

MLE for the hierarchical normal model

$$
\begin{aligned}
y_{i,j} &= \mu + a_j + \epsilon_{i,j} \\
\{\epsilon_{i,j}\} &\sim \text{iid } N(0, \sigma^2) \\
\{a_j\} &\sim \text{iid } N(0, \tau^2)
\end{aligned}
$$

Parameters to estimate:

- Fixed effects: μ
- Variance components: σ^2, τ^2
- Random effects: $a_1, \ldots, a_m$

Likelihood estimation focuses on estimation of $\theta = (\mu, \sigma^2, \tau^2)$

Alternative methods are required for estimation of $a_1, \ldots, a_m$.

MLE for the hierarchical normal model

$$y_{i,j} = \mu + a_j + \epsilon_{i,j}$$
$$\{\epsilon_{i,j}\} \sim \text{iid } N(0, \sigma^2)$$
$$\{a_j\} \sim \text{iid } N(0, \tau^2)$$

Parameters to estimate:

- Fixed effects: μ
- Variance components: σ^2, τ^2
- Random effects: $a_1, \ldots, a_m$

Likelihood estimation focuses on estimation of $\theta = (\mu, \sigma^2, \tau^2)$

Alternative methods are required for estimation of $a_1, \ldots, a_m$.

MLE for the hierarchical normal model

$$y_{i,j} = \mu + a_j + \epsilon_{i,j}$$
$$\{\epsilon_{i,j}\} \sim \text{iid } N(0, \sigma^2)$$
$$\{a_j\} \sim \text{iid } N(0, \tau^2)$$

Parameters to estimate:

- Fixed effects: μ
- Variance components: σ^2, τ^2
- Random effects: $a_1, \ldots, a_m$

Likelihood estimation focuses on estimation of $\theta = (\mu, \sigma^2, \tau^2)$

Alternative methods are required for estimation of $a_1, \ldots, a_m$.

HNM likelihood

Data:

$$\begin{aligned}\boldsymbol{y} &= (y_{1,1}, \ldots, y_{n_j,1}, \ldots, y_{1,m}, \ldots, y_{n_m,m}) \\ &= (\{y_{1,1}, \ldots, y_{n_j,1}\}, \ldots, \{y_{1,m}, \ldots, y_{n_m,m}\}) \\ &= (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)\end{aligned}$$

Likelihood:

$$l(\mu, \sigma^2, \tau^2 : \boldsymbol{y}) = p(\boldsymbol{y}|\mu, \tau^2, \sigma^2)$$

Recall: Under the HNM,

- observations within groups are correlated;
- observations across groups are independent.

$$\begin{aligned}l(\mu, \sigma^2, \tau^2 : \boldsymbol{y}) = p(\boldsymbol{y}|\mu, \tau^2, \sigma^2) &= p(\boldsymbol{y}_1|\mu, \tau^2, \sigma^2) \times \cdots \times p(\boldsymbol{y}_m|\mu, \tau^2, \sigma^2) \\ &= \prod_{j=1}^{m} p(\boldsymbol{y}_j|\mu, \tau^2, \sigma^2)\end{aligned}$$

HNM likelihood

Data:

$$
\begin{aligned}
\boldsymbol{y} &= (y_{1,1}, \ldots, y_{n_j,1}, \ldots, y_{1,m}, \ldots, y_{n_m,m}) \\
&= (\{y_{1,1}, \ldots, y_{n_j,1}\}, \ldots, \{y_{1,m}, \ldots, y_{n_m,m}\}) \\
&= (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)
\end{aligned}
$$

Likelihood:

$$
l(\mu, \sigma^2, \tau^2 : \boldsymbol{y}) = p(\boldsymbol{y}|\mu, \tau^2, \sigma^2)
$$

Recall: Under the HNM,

- observations within groups are correlated;
- observations across groups are independent.

$$
\begin{aligned}
l(\mu, \sigma^2, \tau^2 : \boldsymbol{y}) = p(\boldsymbol{y}|\mu, \tau^2, \sigma^2) &= p(\boldsymbol{y}_1|\mu, \tau^2, \sigma^2) \times \cdots \times p(\boldsymbol{y}_m|\mu, \tau^2, \sigma^2) \\
&= \prod_{j=1}^{m} p(\boldsymbol{y}_j|\mu, \tau^2, \sigma^2)
\end{aligned}
$$

HNM likelihood

Data:

$$\begin{aligned}\boldsymbol{y} &= (y_{1,1}, \ldots, y_{n_j,1}, \ldots, y_{1,m}, \ldots, y_{n_m,m}) \\ &= (\{y_{1,1}, \ldots, y_{n_j,1}\}, \ldots, \{y_{1,m}, \ldots, y_{n_m,m}\}) \\ &= (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)\end{aligned}$$

Likelihood:

$$l(\mu, \sigma^2, \tau^2 : \boldsymbol{y}) = p(\boldsymbol{y}|\mu, \tau^2, \sigma^2)$$

Recall: Under the HNM,

- observations within groups are correlated;
- observations across groups are independent.

$$\begin{aligned}l(\mu, \sigma^2, \tau^2 : \boldsymbol{y}) = p(\boldsymbol{y}|\mu, \tau^2, \sigma^2) &= p(\boldsymbol{y}_1|\mu, \tau^2, \sigma^2) \times \cdots \times p(\boldsymbol{y}_m|\mu, \tau^2, \sigma^2) \\ &= \prod_{j=1}^{m} p(\boldsymbol{y}_j|\mu, \tau^2, \sigma^2)\end{aligned}$$

HNM likelihood

Data:

$$
\begin{aligned}
\boldsymbol{y} &= (y_{1,1}, \ldots, y_{n_j,1}, \ldots, y_{1,m}, \ldots, y_{n_m,m}) \\
&= (\{y_{1,1}, \ldots, y_{n_j,1}\}, \ldots, \{y_{1,m}, \ldots, y_{n_m,m}\}) \\
&= (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)
\end{aligned}
$$

Likelihood:

$$
l(\mu, \sigma^2, \tau^2 : \boldsymbol{y}) = p(\boldsymbol{y}|\mu, \tau^2, \sigma^2)
$$

Recall: Under the HNM,

- observations within groups are correlated;
- observations across groups are independent.

$$
\begin{aligned}
l(\mu, \sigma^2, \tau^2 : \boldsymbol{y}) = p(\boldsymbol{y}|\mu, \tau^2, \sigma^2) &= p(\boldsymbol{y}_1|\mu, \tau^2, \sigma^2) \times \cdots \times p(\boldsymbol{y}_m|\mu, \tau^2, \sigma^2) \\
&= \prod_{j=1}^{m} p(\boldsymbol{y}_j|\mu, \tau^2, \sigma^2)
\end{aligned}
$$

Likelihood contribution from a single group

$$
\begin{aligned}
y_{i,j} &= \mu + a_j + \epsilon_{i,j} \\
\epsilon_{1,j}, \ldots, \epsilon_{n_j,j} &\sim \text{iid } N(0, \sigma^2) \\
a_j &\sim N(0, \tau^2)
\end{aligned}
$$

As we've discussed, the $y_{i,j}$'s are normal with

- $\mathrm{E}[y_{i,j}|\mu] = \mu$
- $\mathrm{Var}[y_{i,j}|\mu] = \sigma^2 + \tau^2$
- $\mathrm{Cov}[y_{i_1,j}, y_{i_2,j}|\mu] = \tau^2$

In vector form, we can express this as follows:

$$
\mathrm{E}[\boldsymbol{y}_j|\mu] = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \mu \mathbf{1} \quad \mathrm{Cov}[\mathbf{y}_j|\mu] = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 \end{pmatrix}
$$

Likelihood contribution from a single group

$$y_{i,j} = \mu + a_j + \epsilon_{i,j}$$
$$\epsilon_{1,j}, \ldots, \epsilon_{n_j,j} \sim \text{iid } N(0, \sigma^2)$$
$$a_j \sim N(0, \tau^2)$$

As we've discussed, the $y_{i,j}$'s are normal with

- $\mathrm{E}[y_{i,j}|\mu] = \mu$
- $\mathrm{Var}[y_{i,j}|\mu] = \sigma^2 + \tau^2$
- $\mathrm{Cov}[y_{i_1,j}, y_{i_2,j}|\mu] = \tau^2$

In vector form, we can express this as follows:

$$\mathrm{E}[\mathbf{y}_j|\mu] = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \mu\mathbf{1} \quad \mathrm{Cov}[\mathbf{y}_j|\mu] = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 \end{pmatrix}$$

Likelihood contribution from a single group

$$y_{i,j} = \mu + a_j + \epsilon_{i,j}$$
$$\epsilon_{1,j}, \ldots, \epsilon_{n_j,j} \sim \text{iid } N(0, \sigma^2)$$
$$a_j \sim N(0, \tau^2)$$

As we've discussed, the $y_{i,j}$'s are normal with

- $\mathrm{E}[y_{i,j}|\mu] = \mu$
- $\mathrm{Var}[y_{i,j}|\mu] = \sigma^2 + \tau^2$
- $\mathrm{Cov}[y_{i_1,j}, y_{i_2,j}|\mu] = \tau^2$

In vector form, we can express this as follows:

$$\mathrm{E}[\mathbf{y}_j|\mu] = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \mu\mathbf{1} \quad \mathrm{Cov}[\mathbf{y}_j|\mu] = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 \end{pmatrix}$$

Likelihood contribution from a single group

$$y_{i,j} = \mu + a_j + \epsilon_{i,j}$$
$$\epsilon_{1,j}, \ldots, \epsilon_{n_j,j} \sim \text{iid } N(0, \sigma^2)$$
$$a_j \sim N(0, \tau^2)$$

As we've discussed, the $y_{i,j}$'s are normal with

- $\mathrm{E}[y_{i,j}|\mu] = \mu$
- $\mathrm{Var}[y_{i,j}|\mu] = \sigma^2 + \tau^2$
- $\mathrm{Cov}[y_{i_1,j}, y_{i_2,j}|\mu] = \tau^2$

In vector form, we can express this as follows:

$$\mathrm{E}[\mathbf{y}_j|\mu] = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \mu\mathbf{1} \quad \mathrm{Cov}[\mathbf{y}_j|\mu] = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 \end{pmatrix}$$

Likelihood contribution from a single group

$$y_{i,j} = \mu + a_j + \epsilon_{i,j}$$
$$\epsilon_{1,j}, \ldots, \epsilon_{n_j,j} \sim \text{iid } N(0, \sigma^2)$$
$$a_j \sim N(0, \tau^2)$$

As we've discussed, the $y_{i,j}$'s are normal with

- $\mathrm{E}[y_{i,j}|\mu] = \mu$
- $\mathrm{Var}[y_{i,j}|\mu] = \sigma^2 + \tau^2$
- $\mathrm{Cov}[y_{i_1,j}, y_{i_2,j}|\mu] = \tau^2$

In vector form, we can express this as follows:

$$\mathrm{E}[\mathbf{y}_j|\mu] = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \mu\mathbf{1} \quad \mathrm{Cov}[\mathbf{y}_j|\mu] = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 \end{pmatrix}$$

Likelihood contribution from a single group

$$y_{i,j} = \mu + a_j + \epsilon_{i,j}$$
$$\epsilon_{1,j}, \ldots, \epsilon_{n_j,j} \sim \text{iid } N(0, \sigma^2)$$
$$a_j \sim N(0, \tau^2)$$

As we've discussed, the $y_{i,j}$'s are normal with

- $\text{E}[y_{i,j}|\mu] = \mu$
- $\text{Var}[y_{i,j}|\mu] = \sigma^2 + \tau^2$
- $\text{Cov}[y_{i_1,j}, y_{i_2,j}|\mu] = \tau^2$

In vector form, we can express this as follows:

$$\text{E}[\boldsymbol{y}_j|\mu] = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \mu\mathbf{1} \quad \text{Cov}[\mathbf{y}_j|\mu] = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 \end{pmatrix}$$

Multivariate normal distribution

This means that $\boldsymbol{y}_j$ has a *multivariate normal distribution*.

The density of a general multivariate normal$(\boldsymbol{\theta}, \Sigma)$ distribution is

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-(\boldsymbol{y}-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\theta})/2\}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & & \vdots \\ \sigma_{1,p} & \sigma_{2,p} & \cdots & \sigma_p^2 \end{pmatrix}.$$

```
ldmvnorm<-function(y, theta, Sig)
{
  -.5*(
 length(y)*log(2*pi) +
 log(det(Sig)) +
  t(y-theta)%*%solve(Sig)%*%(y-theta)
      )
}
```

Multivariate normal distribution

This means that $\boldsymbol{y}_j$ has a *multivariate normal distribution*.

The density of a general multivariate normal$(\boldsymbol{\theta}, \Sigma)$ distribution is

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-(\boldsymbol{y}-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\theta})/2\}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & & \vdots \\ \sigma_{1,p} & \sigma_{2,p} & \cdots & \sigma_p^2 \end{pmatrix}.$$

```
ldmvnorm<-function(y, theta, Sig)
{
  -.5*(
 length(y)*log(2*pi) +
 log(det(Sig)) +
  t(y-theta)%*%solve(Sig)%*%(y-theta)
      )
}
```

Multivariate normal distribution

This means that $\boldsymbol{y}_j$ has a *multivariate normal distribution*.

The density of a general multivariate normal$(\boldsymbol{\theta}, \Sigma)$ distribution is

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-(\boldsymbol{y}-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\theta})/2\}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & & \vdots \\ \sigma_{1,p} & \sigma_{2,p} & \cdots & \sigma_p^2 \end{pmatrix}.$$

```
ldmvnorm<-function(y, theta, Sig)
{
  -.5*(
 length(y)*log(2*pi) +
 log(det(Sig)) +
  t(y-theta)%*%solve(Sig)%*%(y-theta)
      )
}
```

Multivariate normal distribution

This means that $\boldsymbol{y}_j$ has a *multivariate normal distribution*.

The density of a general multivariate normal$(\boldsymbol{\theta}, \Sigma)$ distribution is

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\{-(\boldsymbol{y} - \boldsymbol{\theta})^T \Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\theta})/2\}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & & \vdots \\ \sigma_{1,p} & \sigma_{2,p} & \cdots & \sigma_p^2 \end{pmatrix}.$$

```
ldmvnorm<-function(y, theta, Sig)
{
  -.5*(
 length(y)*log(2*pi) +
 log(det(Sig)) +
  t(y-theta)%*%solve(Sig)%*%(y-theta)
       )
}
```

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$
\begin{aligned}
L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
&= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\
&= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2),
\end{aligned}
$$

where $\log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- θ with $\mu 1$
- Σ with the covariance matrix from the previous slide.

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$
\begin{aligned}
L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
&= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\
&= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2),
\end{aligned}
$$

where $\log p(\boldsymbol{y}_j|\mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- θ with $\mu 1$
- Σ with the covariance matrix from the previous slide.

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$
\begin{aligned}
L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
&= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\
&= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2),
\end{aligned}
$$

where $\log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- θ with $\mu 1$
- Σ with the covariance matrix from the previous slide.

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$
\begin{aligned}
L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
&= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\
&= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2),
\end{aligned}
$$

where $\log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- θ with $\mu 1$
- Σ with the covariance matrix from the previous slide.

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$\begin{aligned} L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ &= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\ &= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2), \end{aligned}$$

where $\log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- θ with $\mu 1$
- Σ with the covariance matrix from the previous slide.

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$\begin{aligned} L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ &= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\ &= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2), \end{aligned}$$

where $\log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- θ with $\mu 1$
- Σ with the covariance matrix from the previous slide.

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$\begin{aligned} L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ &= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\ &= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2), \end{aligned}$$

where $\log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- $\boldsymbol{\theta}$ with $\mu \mathbf{1}$
- Σ with the covariance matrix from the previous slide.

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$\begin{aligned} L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ &= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\ &= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2), \end{aligned}$$

where $\log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- $\boldsymbol{\theta}$ with $\mu \mathbf{1}$
- Σ with the covariance matrix from the previous slide.

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$
\begin{aligned}
L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\
&= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\
&= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2),
\end{aligned}
$$

where $\log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- $\boldsymbol{\theta}$ with $\mu \mathbf{1}$
- Σ with the covariance matrix from the previous slide.

Computing the log-likelihood

MLEs of (μ, σ^2, τ^2) can be found by maximizing the log likelihood.

Log likelihood:

$$\begin{aligned} L(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ l(\boldsymbol{y} : \mu, \sigma^2, \tau^2) &= \log p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \mu, \sigma^2, \tau^2) \\ &= \log \prod_{j=1}^{m} p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2) \\ &= \sum_{j=1}^{m} \log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2), \end{aligned}$$

where $\log p(\boldsymbol{y}_j | \mu, \sigma^2, \tau^2)$ is the log of a multivariate normal density.

For the HNM, we replace

- $\boldsymbol{\theta}$ with $\mu \mathbf{1}$
- Σ with the covariance matrix from the previous slide.

Computing the (minus) log-likelihood

```
mll.oneway

## function(mus2t2,y,g)
##
## {
##   mu<-mus2t2[1] ; s2<-mus2t2[2] ; t2<-mus2t2[3]
##
##   ll<-0
##
##   for(gj in sort(unique(g)))
##
##   {
##
##     nj<-sum(g==gj)
##
##     S<-diag(s2,nj) + matrix(t2,nj,nj)
##
##     ll<-ll+ldmvnorm(y[g==gj],mu,S)
##
##   }
##
## -ll
##
## }
```

Example: Wheat data

```
mll.oneway( c(16.3, 1.787, 0.31 ), y,g)

##          [,1]
## [1,] 88.58541

mll.oneway( c(15, 1.787, 0.31 ), y,g)

##          [,1]
## [1,] 101.3711

mll.oneway( c(16.3, 2, 0.31 ), y,g)

##          [,1]
## [1,] 88.71672

mll.oneway( c(16.3, 1.787, 0.4 ), y,g)

##          [,1]
## [1,] 88.62378
```

Optimization in R

```
fit.ml<-optim(c(15,1,1),mll.oneway,gr=NULL,y=y,g=g,lower=c(-Inf,0,0),method="L-BFGS-B",hessian=TRUE)

fit.ml

## $par
## [1] 16.3063995  1.7872063  0.3099255
##
## $value
## [1] 88.5851
##
## $counts
## function gradient
##       16       16
##
## $convergence
## [1] 0
##
## $message
## [1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
##
## $hessian
##              [,1]         [,2]         [,3]
## [1,] 1.498426e+01 2.186695e-06 1.090683e-05
## [2,] 2.186695e-06 6.710598e+00 2.245294e+00
## [3,] 1.090683e-05 2.245294e+00 1.122654e+01
```

The MLEs are

$$\hat{\mu} = 16.3063995\ ,\ \hat{\sigma}^2 = 1.7872063\ ,\ \hat{\tau}^2 = 0.3099255$$

Confidence intervals via the Information matrix

For maximum likelihood estimation in general,

- $\hat{\gamma}_{MLE} \to \theta$ as the sample size goes to infinity (if the model is correct);
- $\hat{\gamma} \dot{\sim}$ normal$(\gamma, \text{Var}[\hat{\gamma}])$, where
- $\text{Var}[\hat{\gamma}] \approx I_n^{-1}$ for large sample sizes.

For our hierarchical normal model, this means that approximate 95% confidence intervals for (μ, τ^2, σ^2) can be obtained from the curvature of the log likelihood.

Confidence intervals via the Information matrix

For maximum likelihood estimation in general,

- $\hat{\gamma}_{MLE} \to \theta$ as the sample size goes to infinity (if the model is correct);
- $\hat{\gamma} \dot{\sim}$ normal$(\gamma, \text{Var}[\hat{\gamma}])$, where
- $\text{Var}[\hat{\gamma}] \approx I_n^{-1}$ for large sample sizes.

For our hierarchical normal model, this means that approximate 95% confidence intervals for (μ, τ^2, σ^2) can be obtained from the curvature of the log likelihood.

Confidence intervals via the Information matrix

For maximum likelihood estimation in general,

- $\hat{\gamma}_{MLE} \to \theta$ as the sample size goes to infinity (if the model is correct);
- $\hat{\gamma} \dot{\sim}$ normal$(\gamma, \text{Var}[\hat{\gamma}])$, where
- Var$[\hat{\gamma}] \approx I_n^{-1}$ for large sample sizes.

For our hierarchical normal model, this means that approximate 95% confidence intervals for (μ, τ^2, σ^2) can be obtained from the curvature of the log likelihood.

Confidence intervals via the Information matrix

For maximum likelihood estimation in general,

- $\hat{\gamma}_{MLE} \rightarrow \theta$ as the sample size goes to infinity (if the model is correct);
- $\hat{\gamma} \dot{\sim}$ normal$(\gamma, \text{Var}[\hat{\gamma}])$, where
- $\text{Var}[\hat{\gamma}] \approx I_n^{-1}$ for large sample sizes.

For our hierarchical normal model, this means that approximate 95% confidence intervals for (μ, τ^2, σ^2) can be obtained from the curvature of the log likelihood.

Confidence intervals via the Information matrix

For maximum likelihood estimation in general,

- $\hat{\gamma}_{MLE} \rightarrow \theta$ as the sample size goes to infinity (if the model is correct);
- $\hat{\gamma} \dot{\sim}$ normal$(\gamma, \text{Var}[\hat{\gamma}])$, where
- $\text{Var}[\hat{\gamma}] \approx I_n^{-1}$ for large sample sizes.

For our hierarchical normal model, this means that approximate 95% confidence intervals for (μ, τ^2, σ^2) can be obtained from the curvature of the log likelihood.

Confidence intervals via the Information matrix

The *observed information matrix* is the (matrix of) second derivative(s) of the negative log-likelihood function at the MLE (aka the *Hessian*):

$$I_n(\hat{\gamma} : \boldsymbol{y}) = \{-\frac{\partial^2 l(\gamma : y)}{\partial \gamma_j \partial \gamma_k}\}|_{\gamma=\hat{\gamma}}$$

The inverse of the information matrix gives an estimate of the variance/covariance of the MLE's:

$$\text{Var}[\hat{\gamma} : y] \approx I_n^{-1}(\hat{\gamma} : \boldsymbol{y})$$

From this, we can get confidence intervals:

- $\sqrt{I_{jj}^{-1}}$ gives an approximate standard error for γ_j.
- The MLE plus and minus 2 standard errors gives a rough confidence interval for the parameters.

$$\Pr(\gamma_j \in \hat{\gamma}_j \pm 2 \times \text{se}[\hat{\gamma}_j]) \approx 0.95$$

Confidence intervals via the Information matrix

The *observed information matrix* is the (matrix of) second derivative(s) of the negative log-likelihood function at the MLE (aka the *Hessian*):

$$I_n(\hat{\gamma} : \boldsymbol{y}) = \{-\frac{\partial^2 l(\gamma : y)}{\partial \gamma_j \partial \gamma_k}\}|_{\gamma=\hat{\gamma}}$$

The inverse of the information matrix gives an estimate of the variance/covariance of the MLE's:

$$\mathrm{Var}[\hat{\gamma} : y] \approx I_n^{-1}(\hat{\gamma} : \boldsymbol{y})$$

From this, we can get confidence intervals:

- $\sqrt{I_{jj}^{-1}}$ gives an approximate standard error for γ_j.
- The MLE plus and minus 2 standard errors gives a rough confidence interval for the parameters.

$$\Pr(\gamma_j \in \hat{\gamma}_j \pm 2 \times \mathrm{se}[\hat{\gamma}_j]) \approx 0.95$$

Confidence intervals via the Information matrix

The *observed information matrix* is the (matrix of) second derivative(s) of the negative log-likelihood function at the MLE (aka the *Hessian*):

$$I_n(\hat{\gamma} : \boldsymbol{y}) = \{-\frac{\partial^2 l(\gamma : y)}{\partial\gamma_j \partial\gamma_k}\}|_{\gamma=\hat{\gamma}}$$

The inverse of the information matrix gives an estimate of the variance/covariance of the MLE's:

$$\mathrm{Var}[\hat{\gamma} : y] \approx I_n^{-1}(\hat{\gamma} : \boldsymbol{y})$$

From this, we can get confidence intervals:

- $\sqrt{I_{jj}^{-1}}$ gives an approximate standard error for γ_j.
- The MLE plus and minus 2 standard errors gives a rough confidence interval for the parameters.

$$\Pr(\gamma_j \in \hat{\gamma}_j \pm 2 \times \mathrm{se}[\hat{\gamma}_j]) \approx 0.95$$

Confidence intervals via the Information matrix

The *observed information matrix* is the (matrix of) second derivative(s) of the negative log-likelihood function at the MLE (aka the *Hessian*):

$$I_n(\hat{\gamma} : \boldsymbol{y}) = \{-\frac{\partial^2 l(\gamma : y)}{\partial \gamma_j \partial \gamma_k}\}|_{\gamma=\hat{\gamma}}$$

The inverse of the information matrix gives an estimate of the variance/covariance of the MLE's:

$$\mathrm{Var}[\hat{\gamma} : y] \approx I_n^{-1}(\hat{\gamma} : \boldsymbol{y})$$

From this, we can get confidence intervals:

- $\sqrt{I_{jj}^{-1}}$ gives an approximate standard error for γ_j.
- The MLE plus and minus 2 standard errors gives a rough confidence interval for the parameters.

$$\Pr(\gamma_j \in \hat{\gamma}_j \pm 2 \times \mathrm{se}[\hat{\gamma}_j]) \approx 0.95$$

Confidence intervals via the Information matrix

The *observed information matrix* is the (matrix of) second derivative(s) of the negative log-likelihood function at the MLE (aka the *Hessian*):

$$I_n(\hat{\gamma} : \boldsymbol{y}) = \{-\frac{\partial^2 l(\gamma : y)}{\partial \gamma_j \partial \gamma_k}\}|_{\gamma=\hat{\gamma}}$$

The inverse of the information matrix gives an estimate of the variance/covariance of the MLE's:

$$\mathrm{Var}[\hat{\gamma} : y] \approx I_n^{-1}(\hat{\gamma} : \boldsymbol{y})$$

From this, we can get confidence intervals:

- $\sqrt{I_{jj}^{-1}}$ gives an approximate standard error for γ_j.
- The MLE plus and minus 2 standard errors gives a rough confidence interval for the parameters.

$$\Pr(\gamma_j \in \hat{\gamma}_j \pm 2 \times \mathrm{se}[\hat{\gamma}_j]) \approx 0.95$$

Confidence intervals via the Information matrix

The *observed information matrix* is the (matrix of) second derivative(s) of the negative log-likelihood function at the MLE (aka the *Hessian*):

$$I_n(\hat{\gamma} : \boldsymbol{y}) = \{-\frac{\partial^2 l(\gamma : y)}{\partial \gamma_j \partial \gamma_k}\}|_{\gamma=\hat{\gamma}}$$

The inverse of the information matrix gives an estimate of the variance/covariance of the MLE's:

$$\mathrm{Var}[\hat{\gamma} : y] \approx I_n^{-1}(\hat{\gamma} : \boldsymbol{y})$$

From this, we can get confidence intervals:

- $\sqrt{I_{jj}^{-1}}$ gives an approximate standard error for γ_j.
- The MLE plus and minus 2 standard errors gives a rough confidence interval for the parameters.

$$\Pr(\gamma_j \in \hat{\gamma}_j \pm 2 \times \mathrm{se}[\hat{\gamma}_j]) \approx 0.95$$

Confidence intervals via the Information matrix

```
gamma.wheat<-fit.ml$par

gamma.wheat

## [1] 16.3063995  1.7872063  0.3099255

I<-fit.ml$hessian

V.wheat<-solve(I)

V.wheat

##               [,1]          [,2]          [,3]
## [1,]  6.673668e-02 -5.694851e-11 -6.482475e-08
## [2,] -5.694851e-11  1.597051e-01 -3.194081e-02
## [3,] -6.482475e-08 -3.194081e-02  9.546274e-02

sqrt(diag(V.wheat))

## [1] 0.2583344 0.3996312 0.3089705

gamma.wheat-2*sqrt(diag(V.wheat))

## [1] 15.7897307  0.9879440 -0.3080154

gamma.wheat+2*sqrt(diag(V.wheat))

## [1] 16.8230684  2.5864686  0.9278664
```

Comparison to what is known

```
sqrt( gamma.wheat[2]/(m*n) + gamma.wheat[3]/m )

##         1
## 0.2583344
```

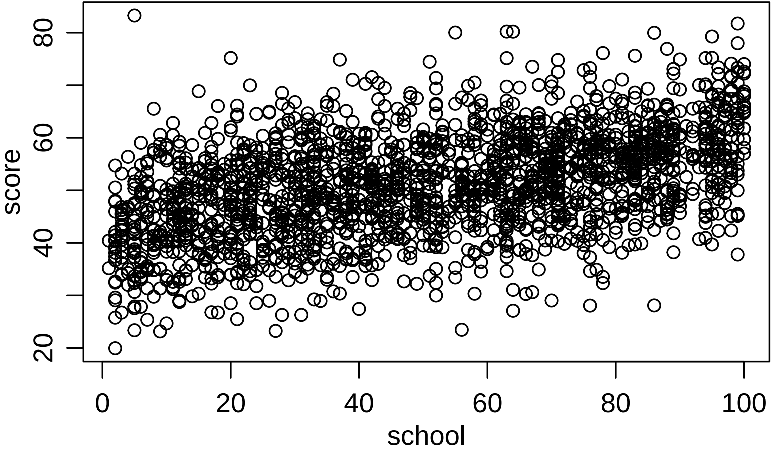
Fitting via lme4: Wheat

```
fit.wheat<-lmer(y~1+(1|g),REML=FALSE)
summary(fit.wheat)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: y ~ 1 + (1 | g)
##
##      AIC      BIC   logLik deviance df.resid
##    183.2    188.9    -88.6    177.2       47
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7913 -0.6035  0.1311  0.6520  1.7262
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  g        (Intercept) 0.3099   0.5567
##  Residual             1.7872   1.3369
## Number of obs: 50, groups:  g, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  16.3064     0.2583   63.12

gamma.wheat

## [1] 16.3063995  1.7872063  0.3099255

sqrt(diag(V.wheat))

## [1] 0.2583344 0.3996312 0.3089705
```

```
CIs<-confint(fit.wheat)

CIs

##                  2.5 %    97.5 %
## .sig01        0.000000  1.228364
## .sigma        1.089911  1.693331
## (Intercept) 15.747322 16.865478

CIs[1:2,]^2

##           2.5 %   97.5 %
## .sig01 0.000000 1.508877
## .sigma 1.187907 2.867369
```

NELS example

100 randomly sampled schools from the NELS dataset



Analysis of all schools

```
fit.ml.nels<-optim(c(50, 1, 1), mll.oneway, gr = NULL, y = nels$mscore, g = nels$school, lower = c(-Inf, 0

fit.ml.nels

## $par
## [1] 50.93914 73.70881 23.63382
##
## $value
## [1] 46956.63
##
## $counts
## function gradient
##       27       27
##
## $convergence
## [1] 0
##
## $message
## [1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
##
## $hessian
##             [,1]        [,2]       [,3]
## [1,] 24.35837087 -0.01576882 0.04913818
## [2,] -0.01576882  1.13128044 0.03026526
## [3,]  0.04913818  0.03026526 0.42089960
```

The MLEs are

$$\hat{\mu} = 50.9391407 \ , \ \hat{\sigma}^2 = 73.708808 \ , \ \hat{\tau}^2 = 23.6338229$$

Confidence intervals via the Information matrix

```
gamma.nels<-fit.ml.nels$par

gamma.nels

## [1] 50.93914 73.70881 23.63382

I<-fit.ml.nels$hessian

V.nels<-solve(I)

V.nels

##               [,1]          [,2]         [,3]
## [1,]  0.0410638760  0.0007019913 -0.004844505
## [2,]  0.0007019913  0.8856698641 -0.063767034
## [3,] -0.0048445047 -0.0637670344  2.381014344

sqrt(diag(V.nels))

## [1] 0.2026422 0.9411003 1.5430536

gamma.nels-2*sqrt(diag(V.nels))

## [1] 50.53386 71.82661 20.54772

gamma.nels+2*sqrt(diag(V.nels))

## [1] 51.34443 75.59101 26.71993
```

Fitting via lme4: Math scores

```
fit.nels<-lmer(mscore~1+(1|school),REML=FALSE,data=nels)
summary(fit.nels)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: mscore ~ 1 + (1 | school)
##    Data: nels
##
##      AIC      BIC   logLik deviance df.resid
##  93919.3  93941.7 -46956.6  93913.3    12971
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8112 -0.6534  0.0093  0.6732  4.6999
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 23.63    4.861
##  Residual             73.71    8.585
## Number of obs: 12974, groups:  school, 684
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  50.9391     0.2026   251.4

gamma.nels

## [1] 50.93914 73.70881 23.63382

sqrt(diag(V.nels))

## [1] 0.2026422 0.9411003 1.5430536
```

```
CIs<-confint(fit.nels)

CIs

##                  2.5 %    97.5 %
## .sig01       4.562275  5.185387
## .sigma       8.479051  8.693913
## (Intercept) 50.541015 51.336528

CIs[1:2,]^2

##           2.5 %   97.5 %
## .sig01 20.81435 26.88823
## .sigma 71.89431 75.58412
```

Our technology so far

ANOVA, method of moments:

- Estimation: $\hat{\mu} = \bar{y}_{\cdot\cdot}$, $\hat{\sigma}^2 = MSE$, $\hat{\tau}^2 = (MSG - MSE)/n$
- Inference: F-test for across-group differences.

Maximum likelihood:

- Estimation: MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2)$
- Inference: CIs for population parameters via likelihood curvature.

What about estimation and inference for a_j's or θ_j's ?

Our technology so far

ANOVA, method of moments:

- Estimation: $\hat{\mu} = \bar{y}_{\cdot\cdot}$, $\hat{\sigma}^2 = MSE$, $\hat{\tau}^2 = (MSG - MSE)/n$
- Inference: F-test for across-group differences.

Maximum likelihood:

- Estimation: MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2)$
- Inference: CIs for population parameters via likelihood curvature.

What about estimation and inference for a_j's or θ_j's ?

Our technology so far

ANOVA, method of moments:

- Estimation: $\hat{\mu} = \bar{y}_{\cdot\cdot}$, $\hat{\sigma}^2 = MSE$, $\hat{\tau}^2 = (MSG - MSE)/n$
- Inference: F-test for across-group differences.

Maximum likelihood:

- Estimation: MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2)$
- Inference: CIs for population parameters via likelihood curvature.

What about estimation and inference for a_j's or θ_j's ?

Our technology so far

ANOVA, method of moments:

- Estimation: $\hat{\mu} = \bar{y}_{\cdot\cdot}$, $\hat{\sigma}^2 = MSE$, $\hat{\tau}^2 = (MSG - MSE)/n$
- Inference: F-test for across-group differences.

Maximum likelihood:

- Estimation: MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2)$
- Inference: CIs for population parameters via likelihood curvature.

What about estimation and inference for a_j's or θ_j's ?

Our technology so far

ANOVA, method of moments:

- Estimation: $\hat{\mu} = \bar{y}_{\cdot\cdot}$, $\hat{\sigma}^2 = MSE$, $\hat{\tau}^2 = (MSG - MSE)/n$
- Inference: F-test for across-group differences.

Maximum likelihood:

- Estimation: MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2)$
- Inference: CIs for population parameters via likelihood curvature.

What about estimation and inference for a_j's or θ_j's ?

Our technology so far

ANOVA, method of moments:

- Estimation: $\hat{\mu} = \bar{y}_{\cdot\cdot}$, $\hat{\sigma}^2 = MSE$, $\hat{\tau}^2 = (MSG - MSE)/n$
- Inference: F-test for across-group differences.

Maximum likelihood:

- Estimation: MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2)$
- Inference: CIs for population parameters via likelihood curvature.

What about estimation and inference for a_j's or θ_j's ?

Our technology so far

ANOVA, method of moments:

- Estimation: $\hat{\mu} = \bar{y}_{\cdot\cdot}$, $\hat{\sigma}^2 = MSE$, $\hat{\tau}^2 = (MSG - MSE)/n$
- Inference: F-test for across-group differences.

Maximum likelihood:

- Estimation: MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2)$
- Inference: CIs for population parameters via likelihood curvature.

What about estimation and inference for a_j's or θ_j's ?