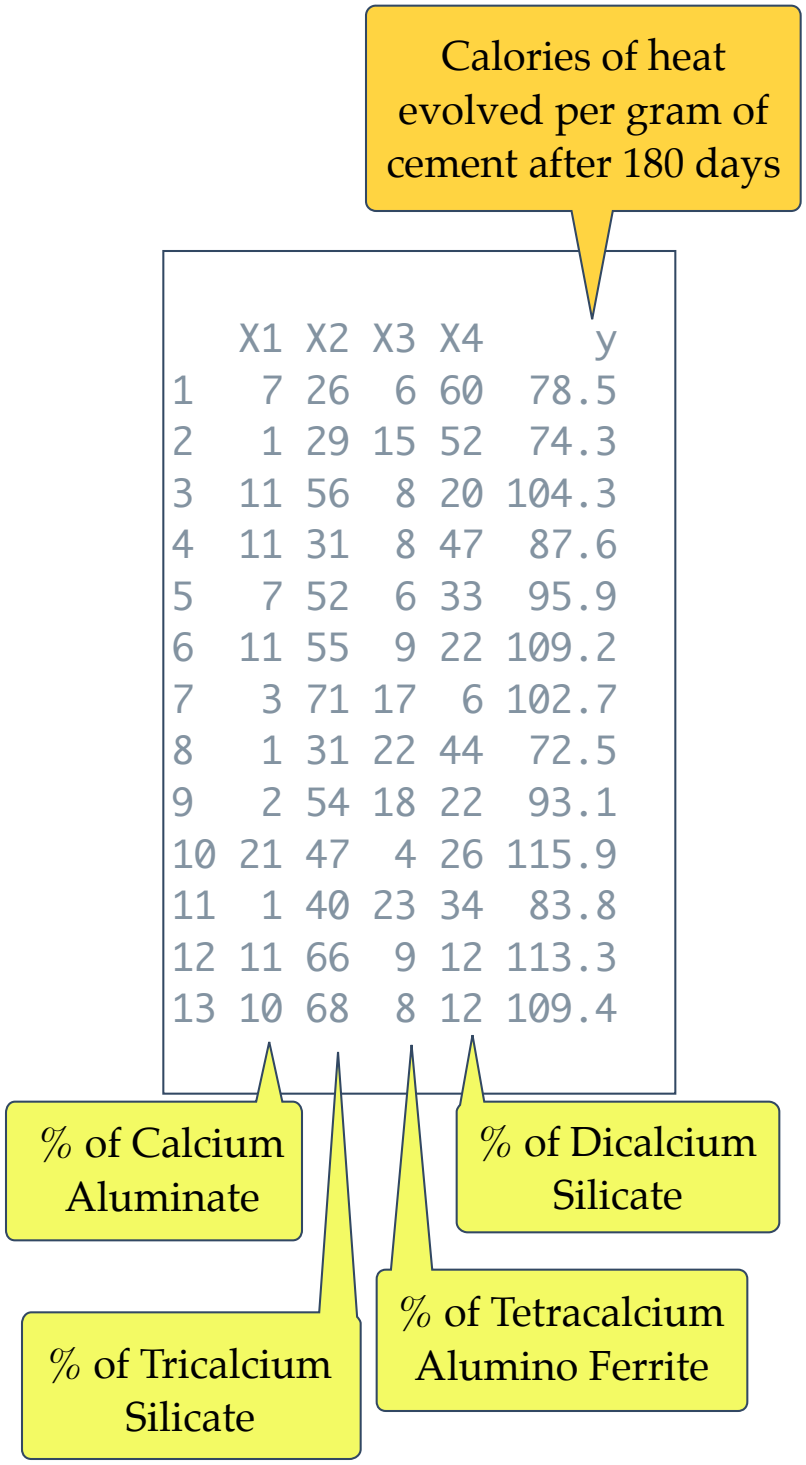
Restricted Maximum Likelihood

```
# Load MuMIN library
> library(MuMIN)

# Load Cement data
> data(Cement)

# Print Cement data
> Cement
```

```
   X1 X2 X3 X4     y
1   7 26  6 60  78.5
2   1 29 15 52  74.3
3  11 56  8 20 104.3
4  11 31  8 47  87.6
5   7 52  6 33  95.9
6  11 55  9 22 109.2
7   3 71 17  6 102.7
8   1 31 22 44  72.5
9   2 54 18 22  93.1
10 21 47  4 26 115.9
11  1 40 23 34  83.8
12 11 66  9 12 113.3
13 10 68  8 12 109.4
```



Maximum Likelihood Estimators

★ It can be shown that the ML estimator for the **error variance** is

$$\hat{\sigma}^2_{\epsilon,ML} = \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^2}{N}$$

★ This is a **biased estimator** of the population variance

✓ Underestimates the population value in repeated sampling

Likelihood

$$\mathcal{L}(\boldsymbol{\beta}; \mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \right)^N \times \exp\left[-\frac{\sum (y_i - \beta_0)^2}{2\sigma_\epsilon^2} \right]$$

Likelihood function

$$\ell(\boldsymbol{\beta}; \mathbf{y}) = -\frac{N}{2} \times \ln\left(2\pi\sigma_\epsilon^2\right) - \frac{1}{2\sigma_\epsilon^2} \times \sum (y_i - \beta_0)^2$$

Log-likelihood function

We can use calculus to show that the estimate for β_o that maximizes the log-likelihood is

$$\hat{\beta}_0 = \frac{\sum y_i}{n}$$

Similarly the estimate for σ^2_e that maximizes the log-likelihood is

$$\hat{\sigma}_\epsilon^2 = \frac{\sum \hat{\epsilon}_i^2}{n}$$

```
# Sum of y
> sum(Cement$y)

[1] 1240.5
```

$$\hat{\beta}_0 = \frac{1240.5}{13} = 95.42$$

```
# Fit mean model
> lm.1 = lm(y ~ 1, data = Cement)

# ANOVA decomposition
> anova(lm.1)

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 12 2715.8  226.31
```

$$\hat{\sigma}^2_\epsilon = \frac{2715.8}{13} = 208.91$$

Maximum Likelihood Variance Estimates

SSE

$$\hat{\beta}_0 = \frac{1240.5}{13} = 95.42 \qquad \hat{\sigma}^2_\epsilon = \frac{2715.8}{13} = 208.91$$

```
# Mean of y
> mean(Cement$y)

[1] 95.42308


# Variance of y
> var(Cement$y)

[1] 226.3136
```

The output from the `var()` function is **not** the maximum likelihood estimate for the variance

$$\hat{\beta}_0 = \frac{1240.5}{13} = 95.42 \qquad \hat{\sigma}^2_\epsilon = \frac{2715.8}{13} = 208.91$$

```
# summary from model
> summary(lm.1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   95.423      4.172   22.87  2.9e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.04 on 12 degrees of freedom
```

$$15.04^2 = 226.2$$

The coefficient output from the model summary are the maximum likelihood estimates. The estimate for the residual standard error (square root of the variance) is **not** the maximum likelihood estimate for the variance.

The maximum likelihood estimate for the variance,

$$\hat{\sigma}_\epsilon^2 = \frac{\sum \hat{\epsilon}_i^2}{n}$$

is a **biased estimator** of the population variance. It *underestimates* the population value in repeated sampling.

$$E\left(\hat{\sigma}_\epsilon^2\right) < \sigma_\epsilon^2$$

To correct for the bias, a different denominator is used,

$$\hat{\sigma}_\epsilon^2 = \frac{\sum \hat{\epsilon}_i^2}{n-k}$$

where *k* is the number of regression coefficients being estimated. This is referred to as the **restricted maximum likelihood (REML)** estimator

$$\hat{\sigma}^2_\epsilon = \frac{\sum \hat{\epsilon}^2_i}{n-k} = \frac{2715.8}{13-1} = 226.3$$

★ Estimation without correction for
ML

The var() function and the model summary output give REML estimates.

ge, ML and

✓ As sample size increases without bound r

Note that REML only affects the variance estimates…the regression coefficients are exactly the same.

erence based

★ REML is correction for variances

✓ Nature of correction depends on fixed ef

✓ Nested models can therefore differ not o
correction terms

✓ REML should not be used for comparing

ML

$$\hat{\sigma}^2_\epsilon = \frac{\sum \hat{\epsilon}^2_i}{n}$$

REML

$$\hat{\sigma}^2_\epsilon = \frac{\sum \hat{\epsilon}^2_i}{n-k}$$

- When *n* is very large the ML and REML estimates for variance will be pretty much the same.
- The magnitude of the difference in the ML and REML estimates for variance are dependent on the value of *k* (the number of fixed-effects in the model)

Consider comparing two nested models,

$$Y_i = \beta_0 + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1(X_i) + \epsilon_i$$

In this situation, we are often interested in examining the effect of X on Y. To do that we assume the only difference between the models is the added effect of X, the fixed-effects structure.

If we use REML, not only are the fixed-effects structures of the two models different, but so are the variances of the errors!

REML should **never be used** to compare the fixed-effects structures of two models!

$$\ell(\boldsymbol{\beta};\mathbf{y}) = -\frac{N}{2} \times \ln\left(2\pi\sigma_\epsilon^2\right) - \frac{1}{2\sigma_\epsilon^2} \times \sum (y_i - \beta_0)^2$$

$$\text{deviance} = 2 \times \ell(\boldsymbol{\beta};\mathbf{y})$$

Solving algebraically, and substituting the sample estimates gives…

$$\text{deviance} = N \cdot \ln\left(2\pi\hat{\sigma}_\epsilon^2\right) + \frac{1}{\hat{\sigma}_\epsilon^2} \cdot \sum \hat{\epsilon}_i^2$$

Substituting the ML estimate for error variance into the deviance equation

$$\hat{\sigma}_\epsilon^2 = \frac{\sum \hat{\epsilon}_i^2}{n}$$

$$\text{deviance} = n \cdot \ln\left(2\pi\frac{\sum \hat{\epsilon}_i^2}{n}\right) + \frac{n}{\sum \hat{\epsilon}_i^2} \cdot \sum \hat{\epsilon}_i^2$$

$$\text{deviance} = n \cdot \ln\left(2\pi\frac{\sum \hat{\epsilon}_i^2}{n}\right) + n$$

$$\text{deviance} = n \cdot \ln\left(2\pi \frac{\sum \hat{\epsilon}_i^2}{n}\right) + n$$

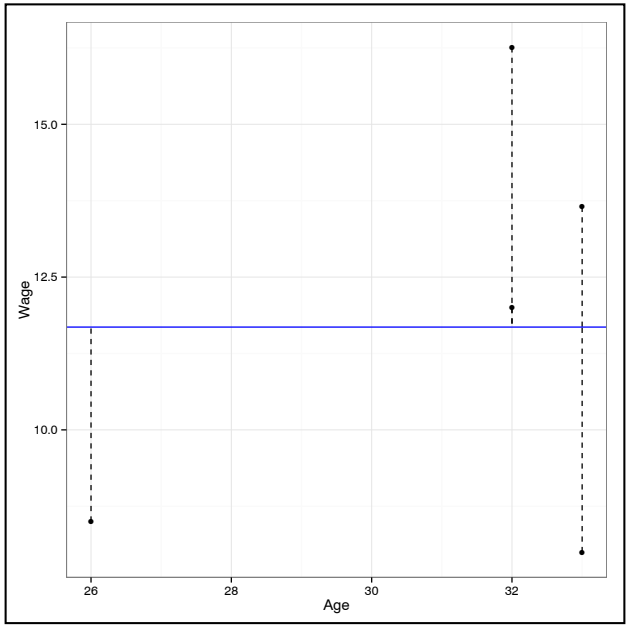
In ML estimation, everything in this equation is given except for the SSE.

For a fixed sample size (n), the SSE is the **only influence** on the size of the deviance. This makes the deviance a good basis for comparing models. Models with smaller deviances will be a better fit to a given set of data.

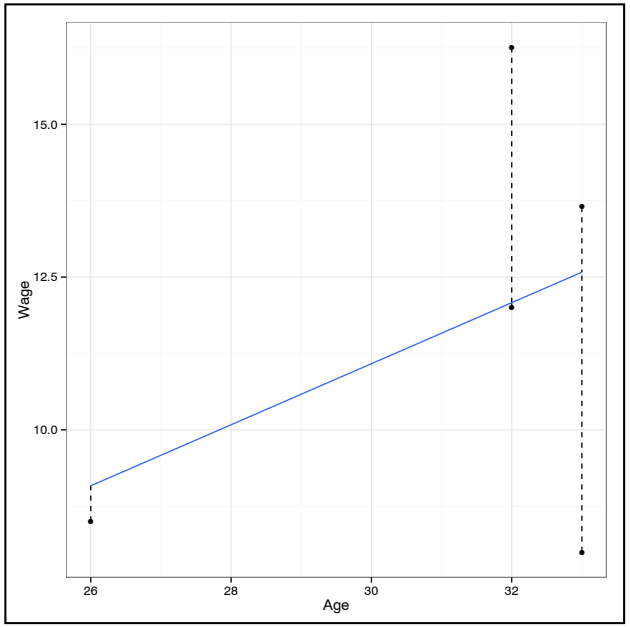
We can also plainly see that minimizing the deviance (which is the same as maximizing the likelihood, or log-likelihood) is equivalent to minimizing the SSE. Now we can see that OLS is a special case of ML (but only for linear models).

Comparing Models

$$y_i = \beta_0 + \epsilon_i$$



$$y_i = \beta_0 + \beta_1(\text{age}) + \epsilon_i$$



- ★ Residuals are larger for the intercept-only model (i.e., larger SSE)
- ★ Slope model fits the data better

```
> -2 * logLik(lm.1)[1]                    ## Compute deviance
[1] 106.3368


> lm.2 = lm(y ~ X1, data = Cement)        ## Fit slope model
> -2 * logLik(lm.2)[1]                    ## Compute deviance
[1] 96.41187
```

The deviance is smaller for slope model (96.4 vs. 106.3). The slope model fits the data better than the intercept-only model.

It is important to note, however, that any model with more parameters will produce a smaller SSE and smaller deviance, and will fit the data better. These two measures will **always decrease** as more predictors added to the model. Even adding statistically worthless predictors will decrease the SSE and deviance.

Information Criteria

To guard against adding potentially worthless predictors, the deviance is "penalized". Penalized indexes are known generally as *information criteria* (IC). Smaller IC values indicate better fit to the data.

$$IC = deviance + penalty$$

- Penalty term is always non-negative
- IC increases as parameters are added to the model

★ Two popular IC are AIC and BIC

- ✓ Akaike information criteria (Akaike, 1973, 1974, 1981)
- ✓ Schwartz's Bayesian information criteria (Schwartz, 1978)

$$\text{AIC} = \text{deviance} + 2 \cdot K$$

$$\text{BIC} = \text{deviance} + K \ln(N)$$

K is the number of estimated parameters in the model

★ Debate about which IC should be used in practice

- ✓ Each has advantages, depending on goals of analysis
- ✓ See Burnham, K. C., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261–304.

Likelihood Ratio Test

- ★ Statistical test based on the deviance for **comparing two nested models**
  - ✓ Models are nested when parameters in more complex model, referred to as **full model**, can be set equal to 0 to obtain **reduced model**
  - ✓ Intercept-only model (reduced model) is nested in the slope model (full model)
- ★ Test statistic: difference in deviances between full and reduced models
  - ✓ Distributed as chi-squared, with *df* equal to the difference in the number of residual *df*

$$\chi^2 = \text{deviance}_R - \text{deviance}_F$$

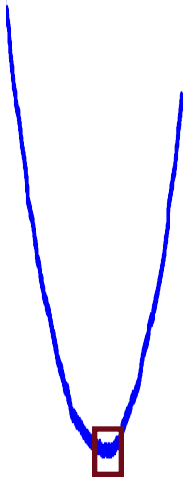
The use of the likelihood test requires ML estimation as noted earlier.

ML Standard Errors

So far, the emphasis has been on computing point estimates (e.g., B_0, B_1) via maximizing the likelihood (or minimizing the deviance).

Sampling variation suggests uncertainty about any estimate (regardless of how it is estimated). We can also compute the SE for an ML estimate.

minimum deviance ???



minimum deviance ???

The SE, which is a measure of estimate uncertainty, is determined by curvature of the deviance function. Relatively **flat deviance functions** indicate **low precision** (i.e., greater uncertainty and a large SE), while relatively **high curvature** indicates **high precision** (small SE).

Indexing Precision of the Estimates

Precision refers to *extent of uncertainty* indexed by SE

Precision is numerically indexed in many ways. One way to index precision is to compute the ratio of an estimate to its estimated SE (i.e., *t*-ratio)

$$t = \frac{\hat{\beta}_k}{\hat{SE}_{\hat{\beta}_k}}$$

Absolute values of $t \approx 0$ indicate **relatively low precision**. It is more desirable to have high precision which correspond to large values of *t*.

Since the *t*-ratio is a **standardized effect**, it is completely appropriate to report *t* as a relative measure and not use cutoffs or statistical tests to evaluate it.

Precision can also be numerically indexed using a confidence interval.

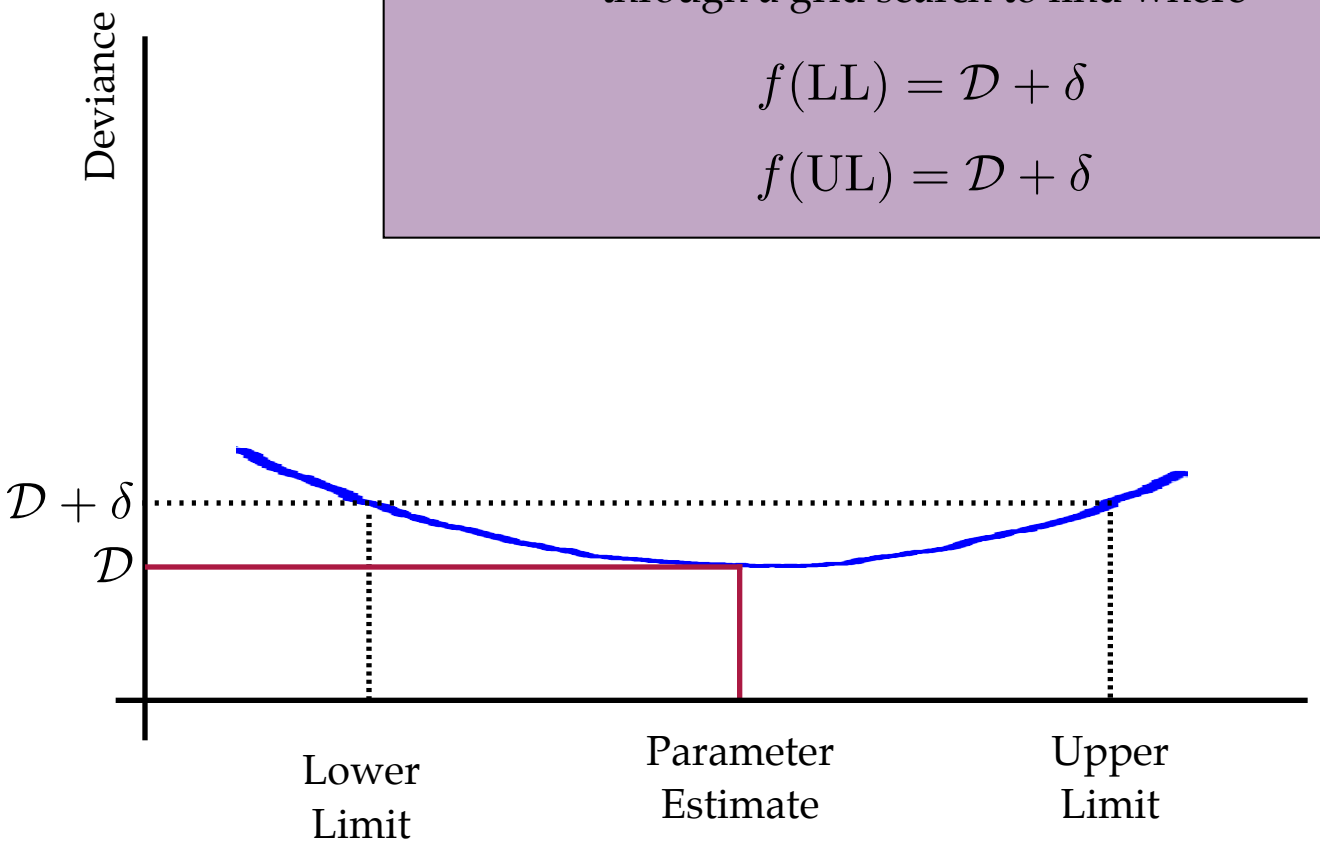
$$\hat{\beta}_k \pm 2 \cdot \hat{SE}_{\hat{\beta}_k}$$

The interval offers applied researchers plausible estimates of the parameter values by using the information in both the estimate and SE of the estimate to give a range of candidate values for the parameter.

The method to compute the limits is called profiling. We profile the likelihood function or the deviance function through a grid search to find where

$$f(\mathrm{LL}) = \mathcal{D} + \delta$$

$$f(\mathrm{UL}) = \mathcal{D} + \delta$$



When the deviance function has more curvature (smaller SE) the CI is smaller indicating more precision on the estimate.