Overview of Maximum Likelihood Estimation

Andrew Zieffler

Conceptual Overview

- Question: For which parameter values are the data most likely?
  - Given the data and model, what are the most likely value of the parameters
  - Maximum likelihood (ML) provides framework for answering this question
  - ML provides results which have attractive properties and are favorable for inference

Optimization Criterion

- Optimization criterion provides basis for computing estimates of parameters
  - Optimization criterion for OLS is SSE
  - SSE is optimal at its minimum (least) value
  - Fixed effects estimates chosen to minimize SSE
- Optimization criterion for ML is likelihood function or deviance function
  - Fixed effects (and other) estimates chosen to minimize likelihood function

Advantages of ML

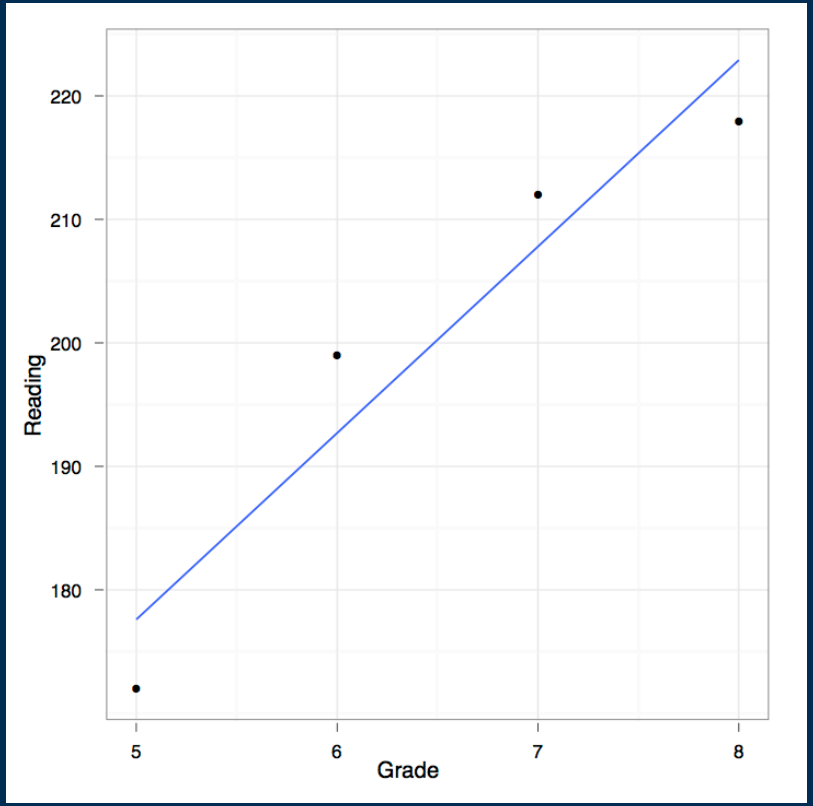
- ML yields a global fit index for the model on top of the parameter estimates
  - Index is minimum of the deviance function
  - Can be used to compare models
- Estimators produced under ML have desirable large-sample (asymptotic) properties
  - Consistent, asymptotically normally distributed
  - With small samples this may not hold

- ML estimates are always approximate
  - Samples are finite
  - Often have missing data
- Approximations may imply researchers should not get too hung up on rigid rules of practice
  - Use of 2(SE) vs 1.96(SE)
  - Inflexible cutoffs (0.05) are not warranted
- Notes only consider *regular problems*
  - ML solutions are possible
  - Assumption of sample data coming from hypothetical, infinitely sized population (repeated sampling scenario)

ML and LM

- We use traditional regression as a familiar starting point
  - Use `Minneapolis-ML.csv`
  - Cross-section of Minneapolis-Long.csv data (one grade from Subjects 1, 3, 5, and 7)
  - Pedagogically small

```
> mpls.ml
  subid read grade
1     1  172     5
2     3  199     6
3     5  212     7
4     7  218     8
```




```
> library( ggplot2 )
```

```
> ggplot( data = mpls.ml, aes( x = grade, y = read ) ) +
    geom_point() +
    stat_smooth( method = "lm", se = FALSE ) +
    theme_bw() +
    scale_x_continuous( name = "Grade", breaks = 5:8 ) +
    scale_y_continuous( name =  "Reading" )
```

- Suppose goal is to estimate slope of line (and that we are ignorant of formula)
- ML estimation
  - Given the data and the model, what is the most likely value of the unknown slope parameter?
- Obviously data is given–it exists in data frame
- What is "given" model?

- Model for this example
  - Traditional regression equation along with the enhancement of an explicit probability model
- Regression equation
  - Grade as single predictor of reading achievement
- Explicit probability model
  - Normal probability model
- Enhancement
  - Traditional regression equation is *embedded* in a normal probability model

- Probability model specified before the analysis
  - Gives probability for all possible samples for the parameters in the model
  - Inferences made according to this model after sample data selected
- Inferences based on likelihood (not probability)
  - Likelihood supplies order of preference (plausibility) among possible parameter values given the data and model
  - Denoted *Lik*

- ML begins with probability model for each and every observed score
  - In traditional regression this is a LM with distributional assumption on the errors

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

where ϵ_i are normally distributed with $\mu_\epsilon = 0$ and $\sigma^2_\epsilon > 0$

- LM model then expressed in terms of probability model

$$\epsilon_i \sim \mathbb{N}(\mu_\epsilon, \sigma_\epsilon^2) \qquad f(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp\left[-\frac{(\epsilon_i - \mu_\epsilon)^2}{2\sigma_\epsilon^2}\right]$$

where $\pi = 3.14159\ldots$ and $exp(a) = e^a$

With the usual assumption that $\mu_\epsilon = 0$ this simplifies to

$$f(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp\left[-\frac{\epsilon_i^2}{2\sigma_\epsilon^2}\right]$$

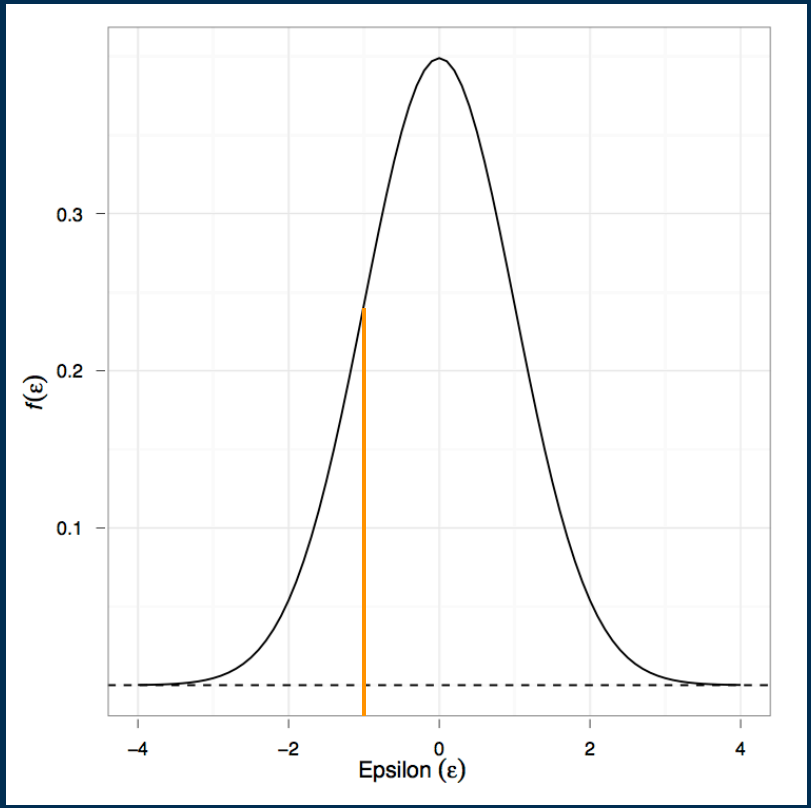
- Probability model associates different probabilities with different individual errors
- `dnorm()` function computes probabilities

Suppose $\sigma_\epsilon^2 = 1$ and $\epsilon_i = -1$, then the probability $f(\epsilon_i)$ =

$\epsilon_i = -1$ $\mu_\epsilon = 0$ $\sigma_\epsilon^2 = 1$

```
> dnorm( -1 , mean = 0 , sd = sqrt(1) )

[1] 0.2419707
```



Likelihood Function

- Basis for all inference in ML estimation
- Consists of probability function for each and every potential observation
- For LM probability function defined for each ϵ_i
  - Assumption of independence allows us to multiply each probability function to obtain joint probability

$$Lik = [f(\epsilon_1)] \cdot [f(\epsilon_2)] \cdot [f(\epsilon_3)] \cdot \ldots \cdot [f(\epsilon_N)]$$

Substitute the function for the normal probability model in for each term in the likelihood function

$$Lik = \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^N exp\left(-\frac{\epsilon_1^2}{2\sigma_\epsilon^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp\left(-\frac{\epsilon_2^2}{2\sigma_\epsilon^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp\left(-\frac{\epsilon_3^2}{2\sigma_\epsilon^2}\right) \cdot \ldots \cdot \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp\left(-\frac{\epsilon_N^2}{2\sigma_\epsilon^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^N exp\left(-\frac{\sum_{i=1}^{N} \epsilon_i^2}{2\sigma_\epsilon^2}\right)$$

Easier to take natural logarithm of both sides of equation

$$log(Lik) = log\left(\left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^N exp\left(-\frac{\sum_{i=1}^{N}\epsilon_i^2}{2\sigma_\epsilon^2}\right)\right)$$

$$= log\left(\left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^N\right) + log\left(exp\left(-\frac{\sum_{i=1}^{N}\epsilon_i^2}{2\sigma_\epsilon^2}\right)\right)$$

$$= N \cdot log\left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right) - \frac{\sum_{i=1}^{N} \epsilon_i^2}{2\sigma_\epsilon^2}$$

$$= N \cdot log\left(\left(2\pi\sigma_\epsilon^2\right)^{-1/2}\right) - \frac{1}{2\sigma_\epsilon^2} \cdot \sum_{i=1}^{N} \epsilon_i^2$$

$$= -\frac{N}{2} \cdot log\left(2\pi\sigma_\epsilon^2\right) - \frac{1}{2\sigma_\epsilon^2} \cdot \sum_{i=1}^{N} \epsilon_i^2$$

This is called the log likelihood function

Can multiply both sides by -2 (called the *deviance function*)

$$-2log(Lik) = N \cdot log\left(2\pi\sigma_\epsilon^2\right) + \frac{1}{\sigma_\epsilon^2} \cdot \sum_{i=1}^{N} \epsilon_i^2$$

We can write ϵ_i as $y_i - \beta_0 - \beta_1 \cdot \text{grade}_i$

$$= N \cdot log\left(2\pi\sigma_\epsilon^2\right) + \frac{1}{\sigma_\epsilon^2} \cdot \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta_1 \cdot \text{grade}_i\right)^2$$

The model is embedded in the specified probability function, by way of the deviance function. Now the deviance function can be minimized (β_0 and β_1 are chosen so that the overall sum is as small as possible.)

Intuitive Idea of Minimization

- Goal is to minimize the value of the deviance
- Recall our data

Subject	Grade	Read
1	5	172
2	6	199
3	7	212
4	8	218

Implausible case (for pedagogical purposes)

- Assume that the parameters of the error variance and the intercept are known, and that the only unknown parameter to be estimated is the slope

$$\beta_0 = 102 \qquad \sigma_\epsilon^2 = 49$$

- Given the model (with known parameter values) and the data, object is to choose "best" value for the slope
- "Best" value here means that after substituting values in to the deviance function equation, the smallest deviance possible has been obtained
- In ML theory, this is the estimate (value) of β_1 that is maximally likely given the data and the model–hence *maximum likelihood*

Everything in the deviance function is known, except for β_1

$$deviance = N \cdot log\left(2\pi\sigma_\epsilon^2\right) + \frac{1}{\sigma_\epsilon^2} \cdot \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta_1 \cdot \text{grade}_i\right)^2$$

Substitute in values

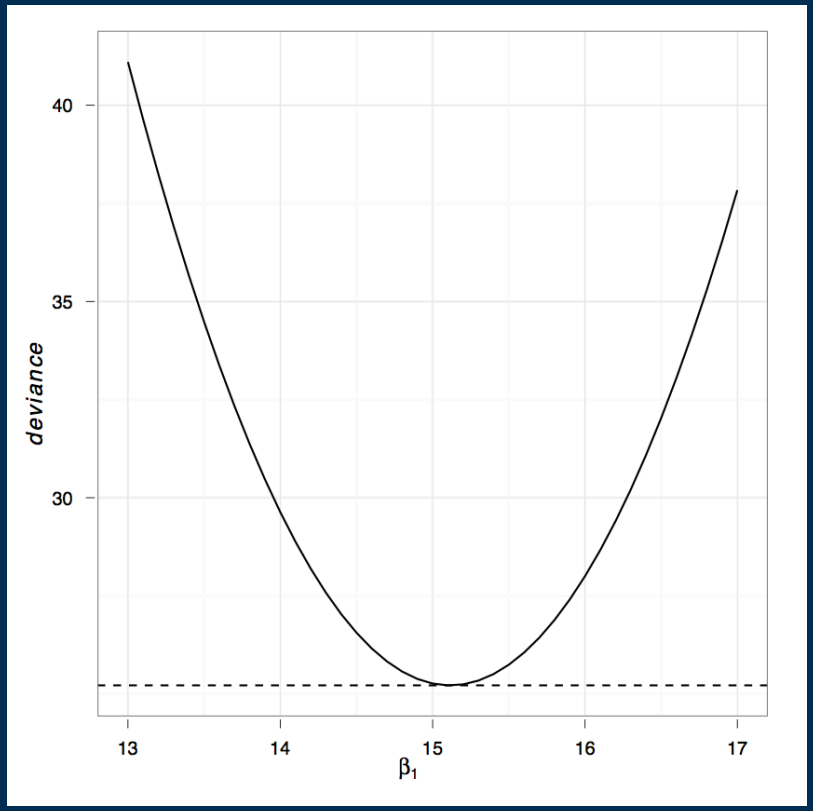
$$\begin{aligned} deviance = 4 \cdot log\left(2 \cdot \pi \cdot 49\right) + \frac{1}{49} \cdot \left(172 - 102 - \beta_1 \cdot 5\right)^2 \\ + \left(199 - 102 - \beta_1 \cdot 6\right)^2 + \left(212 - 102 - \beta_1 \cdot 7\right)^2 \\ + \left(218 - 102 - \beta_1 \cdot 8\right)^2 \end{aligned}$$

Given we want the smallest deviance what value should we choose for β_1 ?

Guess and check

- Pick several candidate values for and substitute them into the equation. Solve for deviance. Choose the one with the smallest deviance value.
- For example, consider candidate values for β_1 between 13 and 17

β_1	deviance
13.0	41.10246
13.1	39.63593
13.2	38.24042
13.3	36.91593
⋮	⋮



Use calculus

- The derivative of the deviance function can be analytically computed and set equal to zero. Solving for β_1 will give the value that minimizes the deviance
- Avoids exhaustive search
- In more complex models (e.g., LMER) estimation is not so straightforward. Exhaustive search methods are required (numerical analysis)
- Once deviance function becomes more complex, the multidimensional form of the tangent line most be discovered

Several Unknown Parameters

- Assume that the parameters of the error variance is known, and that both the intercept and the slope are unknown parameters to be estimated

$$\sigma_\epsilon^2 = 49$$

Substitute in values

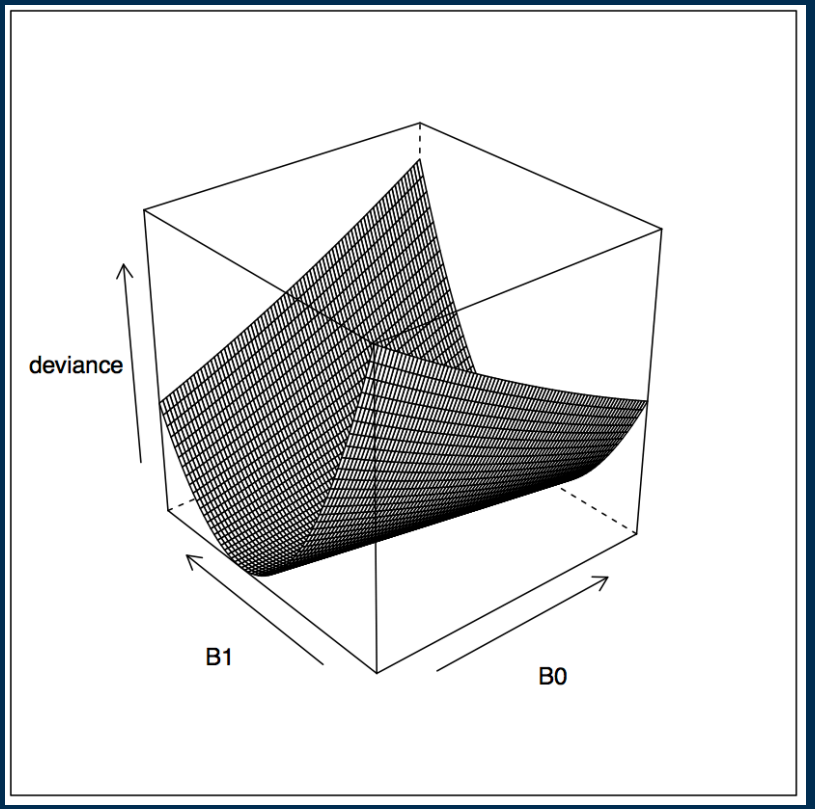
$$deviance = 4 \cdot log\left(2 \cdot \pi \cdot 49\right) + \frac{1}{49} \cdot \left(172 - \beta_0 - \beta_1 \cdot 5\right)^2 + \left(199 - \beta_0 - \beta_1 \cdot 6\right)^2 + \left(212 - \beta_0 - \beta_1 \cdot 7\right)^2 + \left(218 - \beta_0 - \beta_1 \cdot 8\right)^2$$

With two unknowns, the deviance function is a plane in three-dimensional space ($x = \beta_0, y = \beta_1, z = deviance$)

Candidate values for β_0 and β_1 are considered simultaneously

Search grid includes (89, 106) for β_0 and (12, 18) for β_1 in increments of 0.1

β_0	β_1	deviance
98.0	12	74.20450
98.1	12	73.80940
98.2	12	73.41593
98.3	12	73.02410
⋮	⋮	⋮
102.1	15.1	25.21879
⋮	⋮	⋮



Minimum deviance occurs at floor of the plot (horizontal plane on which the graph rests)

If the error variance is also unknown, all three parameters are simultaneously estimated and the deviance is again minimized.

```
> lm.1 <- lm( read ~ 1 + grade, data = mpls.ml )

> logLik( lm.1 )
'log Lik.' -12.35262 (df=3)

> -2 * -12.35262
[1] 24.70524
```

Exhaustive Search vs. Numerical Methods

- The exhaustive searches carried out in the examples were convenient
  - Ranges were used where the ML estimates were known to reside
  - In practice, these are not known
  - Increments are usually finer than 0.1 in practice
- Computing time and memory becomes a valuable resource

- Numerical methods, based on calculus are usually employed
  - Newton-Raphson method is most common algorithm
  - Deviance function assumed to be smooth and continuous with only one minimum (regularity assumption)
- Functions such as lmer() generally combine both methods
  - Numerical methods are used to get in the neighborhood of the minimum deviance
  - Once more limited search space has been defined an iterative method can hone in on the minimum deviance until it is "good enough"

```
> lmer.0 <- lmer( read ~ 1 + grade + ( 1 | subid ),
                  data = mpls.l, REML = FALSE, verbose = TRUE )

  0:     635.00580: 0.856349
  1:     591.28233:  1.85635
  2:     584.56323:  2.26758
  3:     580.59691:  2.76574
  4:     579.53345:  3.09988
  5:     579.30814:  3.30318
  6:     579.28909:  3.37498
  7:     579.28867:  3.38694
  8:     579.28867:  3.38751
  9:     579.28867:  3.38751
 10:     579.28867:  3.38751
```

deviance

ratio of SD of intercept to SD of random effects

$$\frac{\sqrt{Var(b_{0i})}}{\sqrt{\sigma^2_\epsilon}}$$

Restricted Maximum Likelihood

- 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

- To correct the bias, a different denominator is used
  - Called restricted maximum likelihood (REML) estimator

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

where p is the number of predictors

- Square root of error variance is *residual standard error*
  - Printed in `summary()` output of `lm()`

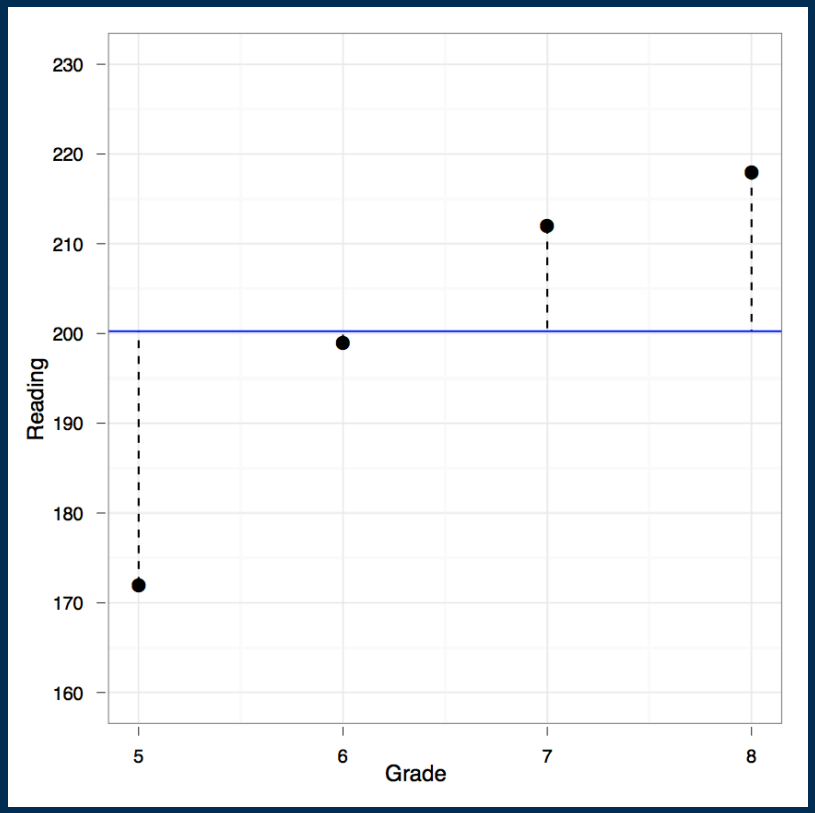
- Estimation without correction for bias is called *full ML* or just ML
- When sample size is large, ML and REML results are similar
  - As sample size increases without bound results converge
  - Most ML results used for inference based on large-sample theory
- REML is correction for variances
  - Nature of correction depends on fixed effects structure of model
  - Nested models can therefore differ not only in fixed effects, but also in correction terms
  - REML should not be used for comparing models

Comparing Models

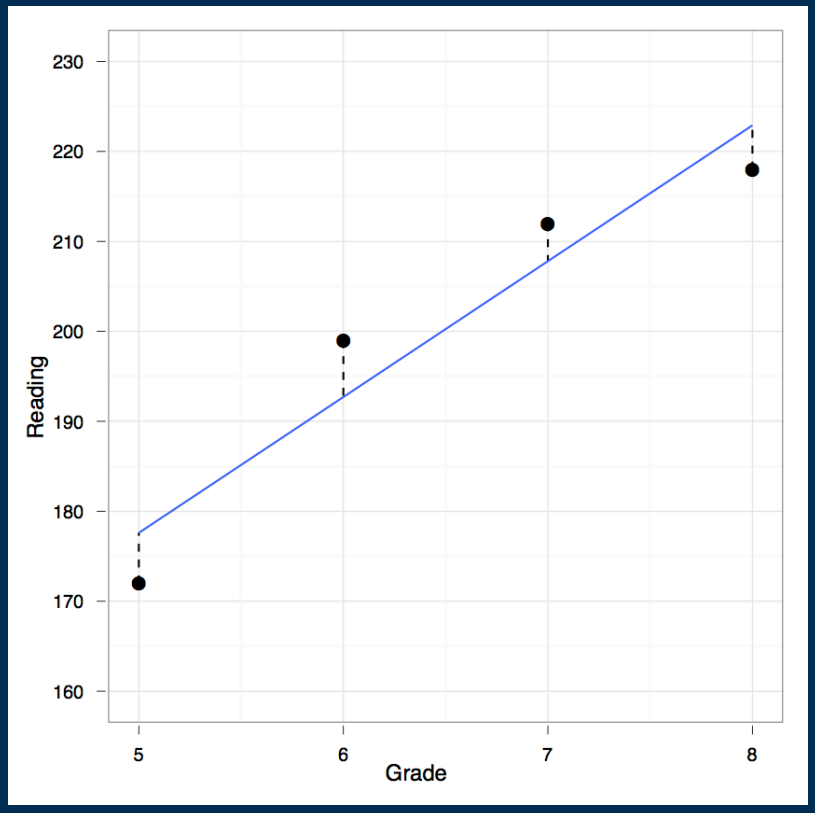
- Deviance is solid foundation for comparing models
  - Basis for AIC and LRT
- Consider intercept-only model

$$y_i = \beta_0 + \epsilon_i$$

$$y_i = \beta_0 + \epsilon_i$$



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



- Residuals are larger for the intercept-only model
- Slope model fits better

- Using the ML estimator

$$deviance = N \cdot log\left(2 \cdot \pi \cdot \hat{\sigma}^2_{\epsilon,ML}\right) + \frac{1}{\hat{\sigma}^2_{\epsilon,ML}} \cdot \sum_{i=1}^{N} \epsilon_i^2$$

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

$$= N \cdot log\left(2 \cdot \pi \cdot \hat{\sigma}^2_{\epsilon,ML}\right) + \frac{N}{\sum_{i=1}^{N} \hat{\epsilon}_i^2} \cdot \sum_{i=1}^{N} \epsilon_i^2$$

$$= N \cdot log\left(2 \cdot \pi \cdot \hat{\sigma}^2_{\epsilon,ML}\right) + N$$

$$= N\left[log\left(2 \cdot \pi \cdot \hat{\sigma}^2_{\epsilon,ML}\right) + 1\right]$$

$$= N\left[log\left(2 \cdot \pi \cdot N^{-1} \sum_{i=1}^{N} \hat{\epsilon}_i^2\right) + 1\right]$$

Sum of squared residuals (SSR)

- SSR is only term in deviance that is not a constant
- For a fixed sample size (N), the SSR is the only influence on the size of the deviance
- Minimizing the deviance is equivalent to minimizing the SSR
- OLS is a special case of ML, but only for LM...not LMER

```
> logLik( lm.0 )
'log Lik.' -17.16936 (df=2)
> -2*-17.16936
[1] 34.33872

> logLik( lm.1 )
'log Lik.' -12.35262 (df=3)
> -2*-12.35262
[1] 24.70524
```

- Deviance is smaller for slope model, model fits better to data
- Any model with more parameters will fit the data better
- SSR and deviance will always decrease as more predictors added to the model
- Worthless predictors will still decrease the SSR and deviance

Information Criteria

- To guard against adding potentially worthless predictors, the deviance is "penalized"
- Penalized indexes are known generally as *information criteria* (IC)
- IC = *deviance* + *penalty*

- Smaller IC values indicate better fit
  - Penalty term is always non-negative
  - 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)

Information Criterion	LM	LMER
AIC	$deviance + 2 \cdot K$	$deviance + 2 \cdot K$
BIC	$deviance + K \cdot log(N)$	$deviance + K \cdot log(\sum_{i}^{N} n_i)$

K is the total number of estimated parameters.

- Debate about which IC should be used in practice
  - Each has advantages, depending on goals of analysis
  - Within this course the use of AIC is emphasized

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 parameters

$$\chi^2 = deviance_R - deviance_F$$

- Larger values of χ^2 indicate better fit
  - Parallels to AIC (discussed in future notes)

ML and LMER

- Deviance function found in similar manner to LM
- Models can be compared using deviance, AIC, BIC or the LRT
  - It is more complex because of the additional parameters in the LMER model, but concepts are all similar

```
> lmer.1 <- lmer( read ~ 1 + grade + ( 1 + grade | subid ), mpls.l,
                  REML = FALSE )

> summary(lmer.1)@AICtab

     AIC      BIC    logLik deviance  REMLdev
583.6975 597.9896 -285.8487 571.6975 565.8497
```

- REMLdev is deviance computed with REML error variance (not useful for model comparison)
- Consider model that adds dadv as predictor

```
> lmer.2 <- lmer( read ~ 1 + grade + dadv + ( 1 + grade | subid ),
                  data = mpls.l, REML = FALSE )
```

Likelihood Ratio Test (LRT)

```
> anova( lmer.1, lmer.2 )

Data: mpls.l
Models:
lmer.1: read ~ 1 + grade + (1 + grade | subid)
lmer.2: read ~ 1 + grade + dadv + (1 + grade | subid)
       Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
lmer.1  6 583.70 597.99 -285.85
lmer.2  7 577.91 594.58 -281.95 7.7916      1   0.005249 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ML Standard Errors

- Emphasis has been on computing ML point estimates via minimum deviance
- Sampling fluctuation suggests uncertainty about every part of any analysis
  - Uncertainty in point estimates indexed by the standard error (SE)
  - Precision refers to extent of uncertainty indexed by SE

- Precision is numerically indexed in many ways
  - Compute ratio of estimate to its estimated SE (i.e., *t*-ratio)

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

- Absolute values of *t* close to zero indicate relatively low precision
  - High precision is desirable
- *t* ratio is a type of standardized effect
  - Use of *t* as a relative measure (without cutoffs or statistical tests) is emphasized

- Consider LMER model with risk and ethnicity intercept effects

Create dummy coded ethnicity variable

```
> mpls.l$white <- ifelse( mpls.l$eth == "Whi", 1, 0 )
```

Fit model

```
> lmer.3 <- lmer( read ~ 1 + grade + dadv + white + ( 1 + grade | subid ),
                  data = mpls.l, REML = FALSE)
```

Examine coefficients for fixed effects

```
> summary(lmer.3)@coefs
```

```
             Estimate Std. Error   t value
(Intercept) 179.264104  8.0006159 22.406288
grade         4.876703  0.7491212  6.509899
dadv        -10.469852  6.8212285 -1.534892
white        19.096176  6.9074093  2.764593
```

- Precision is can also be numerically indexed using a CI
  - CI for fixed effect

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

- Interval offers applied researchers plausible estimates of the parameter values

Make the coefficients slot a stand-alone data frame

```
> mytable <- as.data.frame( summary( lmer.2 )@coefs )
```

Compute the lower limit on the CI and add it into the data frame

```
> mytable$LCI <- mytable$Estimate - 2 * mytable$"Std. Error"
```

Compute the upper limit on the CI and add it into the data frame

```
> mytable$UCI <- mytable$Estimate + 2 * mytable$"Std. Error"
```

Examine data frame

```
> mytable

             Estimate Std. Error   t value       LCI       UCI
(Intercept) 192.453130  7.0479044 27.306433 178.357321 206.548939
grade         4.883554  0.7465818  6.541219   3.390391   6.376718
dadv        -20.398777  6.6514761 -3.066804 -33.701729  -7.095824
```

- God forbid...you need *p*-values
  - *t*-ratio is a special case of general form of statistic known as *Wald statistic*
  - Two-tailed test for $H_0 : \beta_k = 0$

```
> mytable$p.value <- 2 * pnorm( q = abs( mytable$"t value" ),
      lower.tail = FALSE )

> round( mytable, 4 )

            Estimate Std. Error t value      LCI      UCI p.value
(Intercept) 192.4531     7.0479 27.3064 178.3573 206.5489  0.0000
grade         4.8836     0.7466  6.5412   3.3904   6.3767  0.0000
dadv        -20.3988     6.6515 -3.0668 -33.7017  -7.0958  0.0022
```

- Multiparameter Wald test
  - Two-tailed test for $H_0 : \beta_k = \beta_{k'} = 0$
  - Wald test and LRT agree for large sample sizes
  - Wald test and LRT do not always agree for small sample sizes
  - Should use one or the other, not both
- Recommendation is to use the LRT
  - No specific alternative hypothesis with Wald test
  - LRT has explicit null (reduced model) and alternative (full model)
  - This is favorable for calculating *p*-values
  - Furhtermore, the Wald test can give different results when varied forms of H_o are tested

- Size of an SE is determined by curvature of the deviance function
  - Relatively flat deviance functions indicate low precision (i.e., greater uncertainty)
  - Relatively high curvature indicates high precision

- Estimated variance-covariance matrix of fixed effects is obtained from deviance function
  - SEs are the square roots of the variances

Varieties of variances and covariances in LMER

Object	Notation	Importance
Random effects	$Var(b_{0i}), Var(b_{1i}), Cov(b_{0i}, b_{1i})$	Indexes individual variation and covariation of change curve terms
Response	$Var(y_{i1}), Var(y_{i2}), Cov(y_{i1}, y_{i2})$	Indexes variability and dependence of the observed responses
Fixed effects	$Var(\beta_0), Var(\beta_1), Cov(\beta_0, \beta_1)$	Square root of variance is the SE; basis for inference with fixed effect

- In formulas for SE the values for grade appear in the formula for $SE_{\hat{\beta}_0}$, but not for $SE_{\hat{\beta}_1}$
  - Linear transformations of time predictor affect SE for intercept, but not SE for slope

Default `lmer()` Output

- Default output does not include *p*-values or CIs
  - Also no SEs for variance components
- Traditional cure for small sample problems is to use the *t*-distribution
  - For simple problems, *df* can be estimated from the data
  - No clear method (if any is appropriate at all) for estimating *df* for LMER

- Not clear what constitutes effective sample size when there is imbalance and missing data
  - Known that effective sample size is between *N* and the number of observations ($\sum n_i$)
  - No agreed upon approach, so Bates left *p*-values and CIs out of `lmer()`
  - Wald statistic can be used for most empirical situations
  - Later we will discuss bootstrap methods for LRT

- SEs for variance components have been omitted because they are prone to misuse
  - SEs are used to compute CIs
  - Variances are bounded at 0 and unbounded above, which leads to asymmetric sampling distributions
  - Leads to negative values in the CI
  - *p*-values based on the Wald test can be wildly inaccurate, since they are not based on the chi-squared distribution regardless of sample size
  - Actually based on mixture distributions
  - Later bootstrap methods of inference will be introduced

Missing Data

- Strategy of handling missing data is to omit it
  - Any row with NA is removed from analysis
- ML provides unbiased estimates as long as the mechanism for missing data is MCAR or MAR

- Missing data can arise from the researcher omitting observations
  - Outliers found in EDA
  - This falls under the MAR taxonomy of missing data mechanisms
  - In theory, ML should provide unbiased estimates when observations are omitted by the researcher