Multimodal Inference

Andrew Zieffler

Objects of Inference

- Statistical inference is reasoned judgment about population unknowns using sample data
  - 'Unknowns' is broadly defined (e.g., parameters, characteristics of parameters)
- All statistical inference based on ML methods

Parameters of Interest in LMER

- LMER models have three additive components
  - Fixed effects structure
  - Random effects structure
  - Random error structure

} Random variables typically not estimated–referred to as *variance components*

- Fixed effects define aggregate change

Focus of most applied research

- Variance components define individual change and differences Secondary importance
  - In many analyses variance components are treated as nuisance parameters
  - Included only to account for dependency in the repeated measures

- Complexity of the statistical inference is dependent on the model
  - Consider following goals research: Evaluate whether there are intercept and slope effects for the risk group static predictor.
- Situation 1
  - It is known that linear change curve and two random effects are adequate to model change
  - Statistical inference and model evaluation focused on fixed effects of the static predictor
  - Fixed effects for change curve (grade) and variance components are estimated, but are *not* objects of inference (included similar to including intercept and slope in traditional regression)

- Situation 2
  - Shape of change curve is unknown (implies number of random effects are unknown)
  - Statistical inference and model evaluation involve many models and different terms
    - Subject-specific cubic change curves? (most complex model): grade, grade^2, grade^3 fixed-effects for change curve and associated random effects
    - Simpler models may drop grade transformations or random effects or combinations
    - Number of potential models is quite large

- Adding a single random effect adds several variance components to the model
  - Consider addition of random slope effect to a model that already includes a random intercept effect
  - Variance–covariance matrix is unconstrained implying that every unique element is estimated
  - Addition of random slope effect adds two variance components (variance of slope effects and covariance between intercept and slope)
  - Addition of third random effects adds three more variance components, and so on

- Statistical inference involving fixed and random effects is not a straight-forward problem
  - Preselection of time predictors and their associated random effects is recommended to make statistical inference more manageable within LMER analysis
  - 'Preselection' means change curve effects are constant across models under consideration
  - Formal inference only includes fixed effects of static predictors
  - Most justifiable when change curve is dictated by theory
  - Random effects can also be suggested by theory, or included automatically for each fixed effect in the change curve

Data-Driven Analyses

- Shape of change curve is suggested by the data
  - Random effects are also suggested by data
  - Introduces uncertainty (this is not accounted for in the inferences carried out later)
  - Currently believed that introduction of uncertainty is worth the simplicity gained in later inferences (in many situations)
  - Chatfield (1995) and Faraway (1992) discuss formal methods of inference for selecting fixed and random effects for change curves

Model uncertainty is underestimated
Extent of bias is unknown

Statistical Strategy

- Applied research begins with a substantive problem
  - Theoretical notions about how the system under study works including proposed influences that may be responsible for the patterns seen in the sample data (i.e., research questions)

    What is the nature of risk group differences in reading achievement over time? What type of longitudinal risk group differences exist when ethnicity is considered?

- Research questions give rise to *working hypotheses*
  - Specific assertions about phenomenon under study

    The advantaged group has a different mean intercept than the disadvantaged group. The growth rate of the advantaged group is different than the disadvantaged group after controlling for ethnicity.

  - Lines are blurry between research questions and working hypotheses
  - Latter is often considered more related to statistical models used in analysis

- Statistical models are translation of working hypotheses into mathematical equations
  - This is often inexact
  - Statistical models can have additional aspects and assumptions not explicitly stated in working hypotheses (e.g., normality)
  - The converse can also be true
  - One hallmark of statistical philosophy of multimodal inference is that the parameters of the models are assumed to vary but never include the value 0
    - This is akin to the alternative models in NHST
    - Unlike NHST, multimodal inference does not include null models...such models do not include hypotheses of interest

- Typical for there to be multiple statistical models of interest since there are often several working hypotheses
  - Formulation of many working hypotheses is encouraged
  - Long history in science
  - RQ: What is the nature of risk group intercept differences in reading achievement when ethnicity is considered?
  - Consider three working hypotheses:
    1. Change curve is linear; intercept difference b/w advantaged and disadvantaged risk groups
    2. Change curve is linear; intercept difference b/w white and non-white ethnic groups
    3. Change curve is linear; intercept difference b/w advantaged and disadvantaged risk groups controlling for ethnicity

- Change curve is linear; intercept difference b/w advantaged and disadvantaged risk groups

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})(\texttt{grade5}_{ij}) + \beta_2(\texttt{dadv}_i) + \epsilon_{ij}$$

- Change curve is linear; intercept difference b/w white and non-white ethnic groups

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})(\texttt{grade5}_{ij}) + \beta_3(\texttt{eth}_i) + \epsilon_{ij}$$

- Change curve is linear; intercept difference b/w advantaged and disadvantaged risk groups controlling for ethnicity

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})(\texttt{grade5}_{ij}) + \beta_2(\texttt{dadv}_i) + \beta_3(\texttt{eth}_i) + \epsilon_{ij}$$

- Compare Model 1 to Model 3

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})(\texttt{grade5}_{ij}) + \beta_2(\texttt{dadv}_i) + \epsilon_{ij}$$

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})(\texttt{grade5}_{ij}) + \beta_2(\texttt{dadv}_i) + \beta_3(\texttt{eth}_i) + \epsilon_{ij}$$

- Model 1 implies that $\beta_2 \neq 0$ and $\beta_3 = 0$
- Model 3 implies that $\beta_2 \neq 0$ and $\beta_3 \neq 0$
- If Model 1 is equally plausible as Model 3, then ethnicity does not account for risk group differences

- Compare Model 2 to Model 3

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})(\texttt{grade5}_{ij}) + \beta_3(\texttt{eth}_i) + \epsilon_{ij}$$

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})(\texttt{grade5}_{ij}) + \beta_2(\texttt{dadv}_i) + \beta_3(\texttt{eth}_i) + \epsilon_{ij}$$

- Model 2 implies that $\beta_3 \neq 0$ and $\beta_2 = 0$
- Model 3 implies that $\beta_2 \neq 0$ and $\beta_3 \neq 0$
- If Model 2 is equally plausible as Model 3, then risk group differences are negligible above and beyond differences accouted for by ethnicity

- What does it mean for one model to be more "plausible" than another model?
  - Greater statistical evidence (based on the sample data) for one model over another
  - Need to quantify the statistical evidence
- ML methods provide a solid foundation for the quantification of statistical evidence
  - Deviance function is the quantification of the statistical evidence
  - All inferences can be gleaned from the deviance function
  - Minimum deviance can be used for assessing relative plausibility of a model

- Drawback is deviance will decrease when number of estimated parameters increase
  - Deviance will indicate better fit (more plausible model) even if added predictors are worthless
  - Deviance is often penalized based on number of predictors used
  - Penalty is tradeoff between model fit and model complexity

- Penalization Approach 1: AIC
  - Deviance adjusted upward to account for number of estimated parameters
- Penalization Approach 2: LRT
  - Penalty is realized in the *df* of the χ^2 sampling distribution
  - Takes into account difference in number of estimated parameters between two nested models

AIC

- Basis in framework known as Kullback–Leibler information theory (Kullback & Leibler, 1951)
- AIC is also estimate of predictive accuracy
  - Ability of model to predict new data
  - This is often of interest to applied researchers

Recall

Penalty guards against improved fit just by adding predictors

$$AIC = deviance + 2 \cdot K$$

where *K* is the number of estimated parameters (number of fixed effects and variance components in model)

- Smaller deviance = better fit (same is true for AIC)
- If worthless predictors are added (i.e., deviance does not improve much), AIC will increase because 2*K* will go up
- If non-worthless predictors are added, 2*K* will still go up, but the decrease in deviance will outweigh this

Illustration of Predictive Accuracy

- Suppose researcher has knowledge of model underlying sample data
  - Population model with known form and parameter values
  - This is the *data generating model* (not the *true model*)
  - True model is never known (if it were known analysis of sample data would be pointless)
- Let true model be cross-sectional regression model with reading achievement as response and attendance as predictor

True model

$$\begin{aligned} y_i &= \beta_0 + \beta_1(\texttt{att}_i) + \epsilon_i \\ &= \beta_0 + 0(\texttt{att}_i) + \epsilon_i \\ &= \beta_0 + \epsilon_i \end{aligned}$$

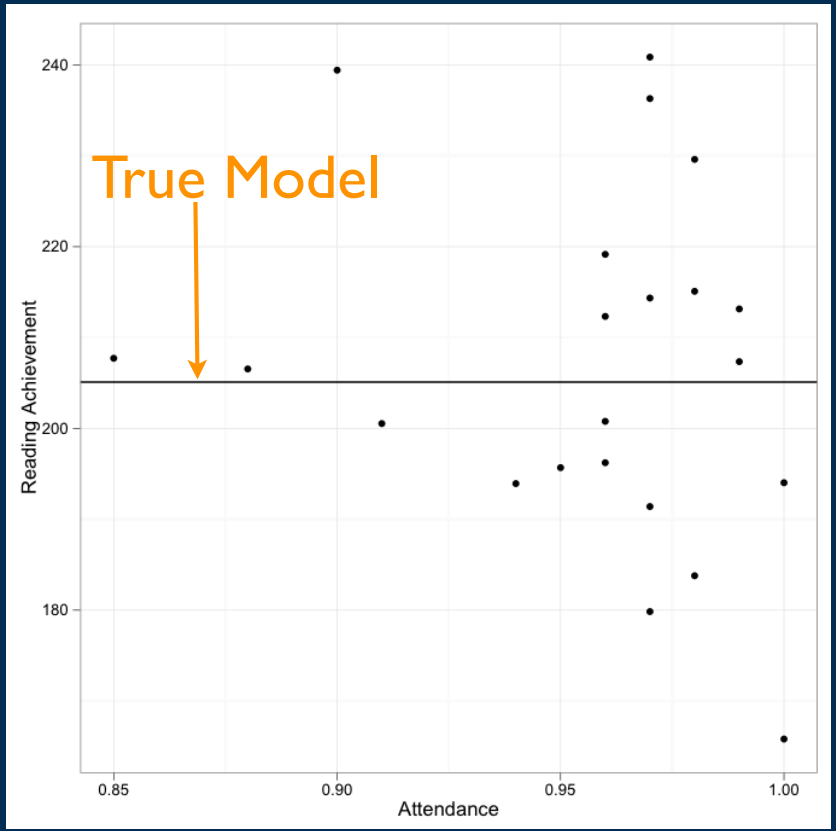
The ε_i are normally distributed with mean of 0 and constant variance.

- There is no relationship between reading achievement scores and attendance
- Despite this, there each subject in the population still has a value for attendance ($\texttt{att}_i$)
- True regression line is horizontal line with intercept β_1
- Because of random error term, random scatter around this line

- When the true model is unknown (typical) goal is to make a knowledgable guess about the form of the regression line
  - Best guess is based on evaluation of different models of the sample data
- To make things concrete, we will use the `simulate()` function from the **lme4** library to generate a sample of response scores from the true model
  - Data set will be called Sample A
  - Associated data frame called `sample.a`
  - New response data from `simulate()` generated under assumption of normality of error terms
  - Predictor scores treated as fixed (values from original sample)

```
## select grade 5 (cross-sectional) data
> mysample <- subset( mpls.l, grade == 5 )

## fit true model
> true.model <- lm( read ~ 1, data = mysample )

## reader can reproduce results
> set.seed( 123 )

## simulate data
> sim.dv <- unlist( simulate( true.model ) )

## store simulated values in sample.a and add attendance
## scores from original data frame
> sample.a <- data.frame( read = sim.dv, att = mysample$att )
```

```
> head( sample.a )
             read  att
sim_11 193.9305 0.94
sim_12 200.5343 0.91
sim_13 236.3005 0.97
sim_14 206.5461 0.88
sim_15 207.7213 0.85
sim_16 239.4266 0.90
```



Suppose you are given Sample A and asked to investigate the relationship between reading achievement and attendance, *with no knowledge of the true model.*

You decide to fit two models

- A linear model using attendance as a predictor
- A non-linear model using both attendance and attendance squared

These two models are referred to as *candidate models*

$$y_i = \beta_0 + \epsilon_i$$ True model

$$y_i = \beta_0 + \beta_1(\texttt{att}_i) + \epsilon_i$$ Candidate model 1

$$y_i = \beta_0 + \beta_1(\texttt{att}_i) + \beta_2(\texttt{att}_i^2) + \epsilon_i$$ Candidate model 2

Typically, one of the candidate models is 'chosen' as the best guess for the true model

We note that neither candidate model is correct

- Both have extraneous predictors and parameters
- Model 1 is 'less false' since it has fewer extraneous parameters

Model 1 is a better *approximating model* than model 2 as it is closer to the truth.

- Since the true model is never known in practice, neither is the better approximating model
- However, an informed guess can be made

Suppose we fit our candidate models to the data in Sample A

```
## Candidate model 1
> lm.1a <- lm( read ~ att, data = sample.a )

## Candidate model 2
> lm.2a <- lm( read ~ att + I( att ^ 2 ),
               data = sample.a )
```

Suppose we compute the deviance for each candidate model

```
> deviance( lm.1a )
[1] 7515.559

> deviance( lm.2a )
[1] 7219.297
```

Indicates better fit for candidate model 2

Since true model is flat, both predictors are fitting random error and the spurious superiority is due to *over-fitting* of the model to the data

More complex models will always have superior fit to the data at hand

- Added complexity is only desired if it constitutes genuine effect

How can you know if the particular fitted model is genuine or a case of over-fitting?

- One way to proceed is to examine how well the fitted models predict new data generated by the true model
- If a model is closer to the truth than another model, it will have better average fit to the new data

What is meant by 'prediction of new data'?

- Each candidate model can be used to produce a set of predicted (fitted) values based on the x_i's
- Suppose a second sample (Sample B) is generated from the true model
- The deviance using the fitted values from the candidate model in question and the observed values from Sample B can be computed
  - In this case, deviance is not solely based on data at hand
  - Can be thought of as a cross-sample *predictive deviance*

Let's generate Sample B and fit the same two candidate models and compute the deviances

```
## Generate data for sample.b
> set.seed( 12 )
> sim.dv <- unlist( simulate( true.model ) )
> sample.b <- data.frame( read = sim.dv, att = mysample$att )

## Fit candidate models
> lm.1b <- lm( read ~ att, data = sample.b )
> lm.2b <- lm( read ~ att + I( att ^ 2 ), data = sample.b )

## Compute deviance
> deviance( lm.1b )
[1] 7264.348

> deviance( lm.2b )
[1] 6714.556
```

Once again, the deviance is smaller for model 2, which indicates better fit for the data at hand

Goal, however, was to compare the predictive deviance for the models

Recall that the deviance, introduced in the last set of notes, included the sum of squared residuals (SSR)

For predictive deviance, the residuals are computed using the responses from Sample B and the fitted values based on Sample A

$$\tilde{\epsilon}_i = y_{Bi} - \hat{y}_{Ai}$$

It can be shown that

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

We will compute these below for model 1

```
> N <- nrow( sample.b )
> prdev.1b <- N * (log( 2 * pi *
    sum(( sample.b$read - fitted( lm.1a$ ) ^ 2 ) ) + 1 )
```

ADD CRAP HERE

Average predictive deviance is important concept related to AIC

- Suppose an infinite number of samples are generated from the true model
- For each sample the predictive deviance is calculated

Given a large sample size, the normality assumption, and the endless repeated sampling scenario, the AIC is an unbiased estimator of the average predictive deviance.

The penalty term $2K$ acts as an adjustment for overfitting comparable to the one that occurs naturally in the average predictive deviance

- The AIC from a single sample is an estimate of the average predictive deviance when the true model is not known

Since the true model is unknown, it is unknown which candidate model is closest to the truth

- In practice, AIC is computed for each candidate model
- Since AIC is unbiased estimate of average long-run predictive deviance, candidate model with lowest AIC is best guess as closest to truth
- This model referred to as *best approximating model*

AIC value can be found in the `summary()` output for the `lmer()` object

AIC value can be extracted directly from the `lmer()` object by accessing the slot `@AICtab$AIC` from the `summary()` output

The theory specifies that the predictive deviance will be smaller on average for the best approximating model

- For any pair of samples the AIC might be smaller for model 2
- The AIC can incorrectly misidentify the best approximating model
- The extent of uncertainty in identifying the best approximating model can be quantified
- This uncertainty quantification is the basis for effect size measures

Some details about the AIC to bear in mind

- AIC values are not standardized
- AIC values may be positive or negative
- AIC values may be very large or very small
- AIC does not offer a method of statistical testing (i.e., the term *statistically significant* should never be used with AIC)
- AIC offers a method of rank ordering the models in a given set of candidate models (nested or not)
- AIC is affected by sample size and should not be used to compare models across studies
- AIC is further changed when he response variable changes so it should not be used to compare models where the response is transformed

Extension to LMER

Previous heuristic of AIC focused on traditional regression
but AIC is valid for other methods including LMER

- Requires appropriate deviance function having a probability distribution with relevant parameters that account for dependency of repeated measures
- In this regard, AIC works with LMER models
- AIC is most appropriate for selection of fixed effects
- Only unbiased when the candidate models differ in their number of fixed effects
- Bias is introduced when the models differ in the number of random effects or associated variance components

Predictive accuracy in LMER refers only to fitted values based on fixed effects

- Over-fitting in LMER is more complex
- Worthless predictors conform not only to random error but also to individual variation represented in the random effects
- Implies that random effects model needs to be 'adequate' so as to eliminate a spurious source in over-fitting
- For linear change, two random effects are considered adequate

Extracting the AIC across many LMER models

- The `aictab()` function from the library **AICmodavg** can be used to conveniently extract the AIC value
- Fitted models are supplied to a list using the argument `cand.set=list()`
- Names for the models must also be supplied as a character vector in the argument `mod.names=c()`
- To compute the AIC, the argument `second.ord=FALSE` is also included

```
## create predictor grade5
> mpls.l$grade5 <- mpls.l$grade - 5

## create dadv and eth2
> mpls.l$dadv <- ifelse(mpls.l$risk == "ADV", 0, 1)
> mpls.l$eth2 <- ifelse(mpls.l$eth == "Whi", 1, 0)

## Fit models
> model.1 <- lmer(read ~ grade5 + dadv + (grade5 | subid), data = mpls.l,
                  REML = FALSE)
> model.2 <- lmer(read ~ grade5 + eth2  + (grade5 | subid), data = mpls.l,
                  REML = FALSE)
> model.3 <- lmer(read ~ grade5 + dadv + eth2 + (grade5 | subid),
                  data = mpls.l, REML = FALSE)
```

AIC Corrected

- The correction term 2k is justified for large sample sizes
- For smaller sample sizes, a bias-adjusted form of the AIC should be used

$$AICc = AIC + \frac{2 \cdot K \cdot (K+1)}{\left(\sum_i^N n_i\right) - K - 1}$$

- Summation is the total number of observed time points (in traditional LM, n is replaced by N)
- As correction term approaches 0 (sum in denominator gets big), AICc approaches AIC
- Recommendation: Always use AICc

```
## load library
> library( AICcmodavg )

## Extract AIC values
> aictab( cand.set = list( model.1, model.2, model.3 ),
          modnames = c( "M1", "M2", "M3" ),
          second.ord = FALSE )

Model selection based on AIC :

   K    AIC Delta_AIC AICWt Cum.Wt      LL
M3 8 573.54      0.00  0.50   0.50 -278.77
M2 7 573.78      0.24  0.44   0.94 -279.89
M1 7 577.91      4.36  0.06   1.00 -281.95
```

Extracting the AICc from LMER models

- The `aictab()` function from the library **AICmodavg** can be used to conveniently extract the AICc value
- The argument `second.ord=` is omitted

```
> aictab( cand.set = list( model.1, model.2, model.3 ),
     modnames = c( "M1", "M2", "M3" ) )

Model selection based on AICc :

   K   AICc Delta_AICc AICcWt Cum.Wt      LL
M2 7 575.34       0.00   0.50   0.50 -279.89
M3 8 575.57       0.24   0.44   0.94 -278.77
M1 7 579.46       4.13   0.06   1.00 -281.95
```

AICc and Effect Size

- The AICc can be used as basis for effect size
- Predictive accuracy underlying AICc value is a relative measure (one never knows the true model and can therefore never estimate its difference from candidate model)
- Can be used to judge the relative proximity to true model for two or more models
  - Smallest value corresponds to best approximating model
  - Since AICc is a sample statistic, it is prone to sampling error
  - If values for two (or more) competing models are similar, there is uncertainty regarding the best approximating model
- Effect size is quantification of (dis)similarity among the candidate models in terms of being he best approximating model
- Many ways to express effect size

Delta

- Delta (Δ) is the difference between each AICc and the smallest AICc
- More formally

$$\Delta_h = AICc_h - AICc_{min}$$

where h = 1, ..., H, with H being the number of models and $AICc_{min}$ is the smallest AICc value in the set of H models

- Δ_h is a calibration of model fit using the best fitting model as the standard
  - Best fitting model has Δ_h of 0
  - All other models have $\Delta_h > 0$
  - Δ_h is interpreted as an estimate of the difference in the plausibility of model *h* and the best fitting model in terms of being the best approximating model

- The interpretation just offered is conditional on the data, the models, and the inability to know the true model
- Plausible candidate models include:
  - The model with the smallest AICc
  - Models with relatively small delta values
- Sometimes the delta value may be small for even the worst fitting candidate model
  - This should be taken as an indication of the large amount of uncertainty in the analysis
- Some guidelines for interpreting delta values (Anderson, 2008)
  - $\Delta_h > 4$ constitutes a strong difference between model *h* and the best fitting model
  - $\Delta_h > 4$ constitutes a very strong difference between model *h* and the best fitting model

- The cutoffs in the guidelines were created based on independence of observations
  - Longitudinal data are not independent
  - These values should not be used as strict cutoff values
- The output from the `aictab()` function also includes the delta values

```
> aictab( cand.set = list( model.1, model.2, model.3 ),
    modnames = c( "M1", "M2", "M3" ) )

Model selection based on AICc :

   K   AICc Delta_AICc AICcWt Cum.Wt      LL
M2 7 575.34       0.00   0.50   0.50 -279.89
M3 8 575.57       0.24   0.44   0.94 -278.77
M1 7 579.46       4.13   0.06   1.00 -281.95
```

Weight of Evidence

- Weight of evidence for the h^{th} model, W_h, is a probability scaling of Δ_h
- The scaling bounds the size between 0 and 1
- Computed as

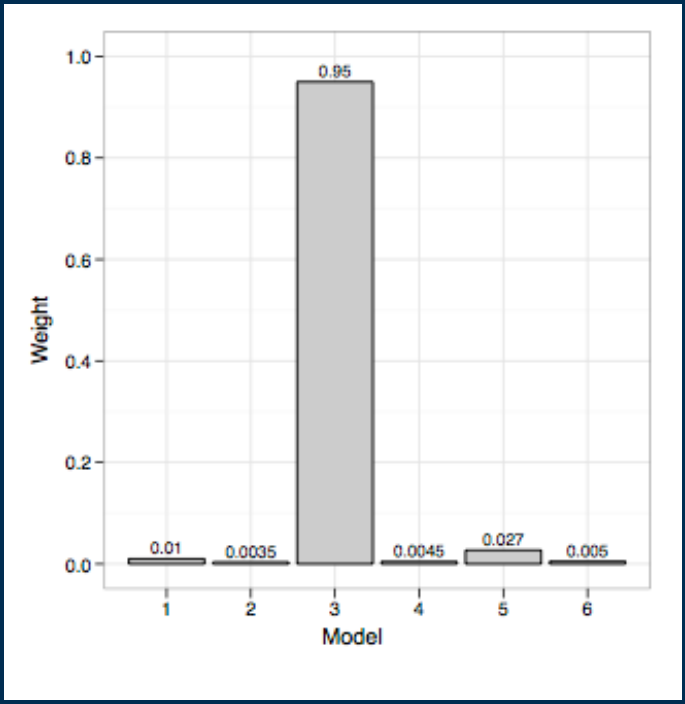
$$W_h = \frac{exp(-0.5 \cdot \Delta_h)}{\sum_h^H exp(-0.5 \cdot \Delta_h)}$$

where the sum is over all models in the set

Given the data, the set of candidate models, and the unknowable true model, W_h indicates the probability that model h is the best approximating model

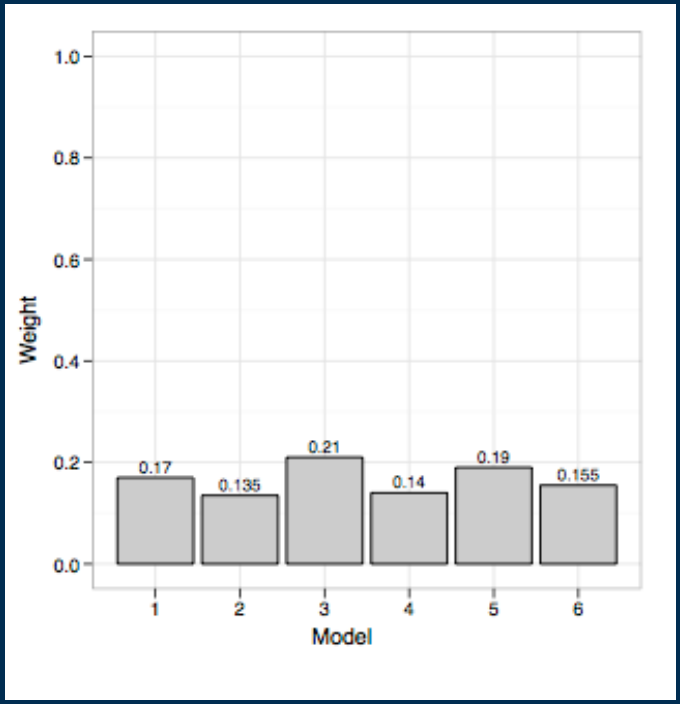
- Larger Δ_h correspond to smaller W_h and indicate that model h is a less likely candidate for the best approximating model
- Weight of evidence can be used as representation of the extent of model uncertainty (uncertainty that model is the best approximating model)
- How large does W_h have to be before a researcher has some degree of certainty about the model?
  - Judgment call by the researcher
  - Anderson (2008) has suggested $W_h = 0.90$ and $W_h = 0.95$ as benchmarks
  - Researcher can be fairly confident that a candidate model is the best approximating model (not the true model) if $W_h \geq 0.90$

- Not always the case that a set of models will have such a large probability
  - Can be convenient to form a confidence set of the models whose probabilities sum to 0.90 or 0.95
  - Confidence statements about the set of models are akin to the statement about an individual model
  - If sum is high, there is confidence that one of the models in the set is the best approximating model



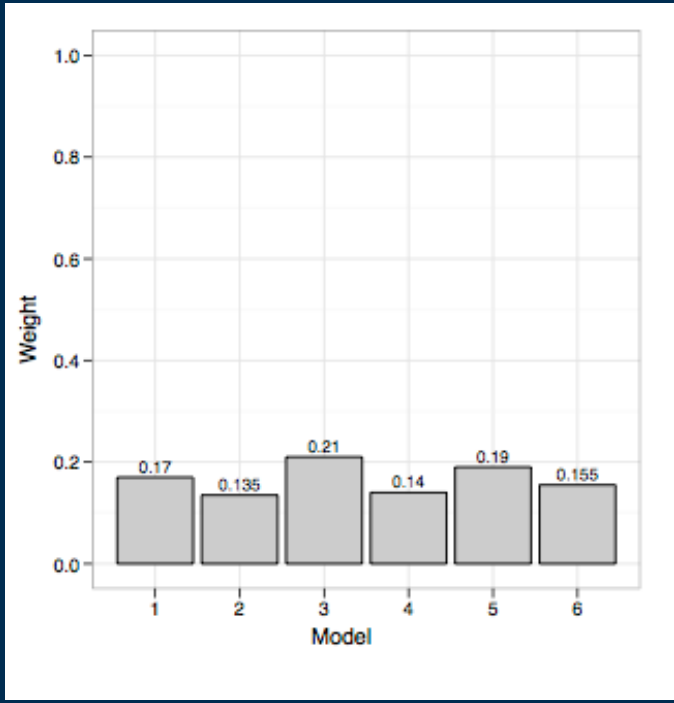
Best fitting model has high weight of evidence

Relative high degree of confidence that if another sample was drawn, Model 3 would again be judged as the best approximating model



Best fitting model has weight of evidence similar to others

Relative low degree of confidence that if another sample was drawn, Model 3 would again be judged as the best approximating model



Confidence Set: Five best fitting models have sum of 0.865. This is still relatively low confidence that if another sample was drawn, Model 3, Model 5, Model 1, Model 6 or Model 4 would be judged as the best approximating model

Neither situation speaks to the worth of the models. They may be close to the true model or very distant. There is no way of knowing.

Model selection based on AICc :

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
M2	7	575.34	0.00	0.50	0.50	-279.89
M3	8	575.57	0.24	0.44	0.94	-278.77
M1	7	579.46	4.13	0.06	1.00	-281.95

Confidence Sets

- With only three models, the output from the `aictab()` function can be used to examine the confidence sets
- With more models it is convenient to use the `confset()` function in the **AICcmodavg** library
- Fitted models are supplied to a list using the argument `cand.set=list()`
- Names for the models must also be supplied as a character vector in the argument `mod.names=c()`

```
> confset(cand.set = list(model.1, model.2, model.3), modnames = c("M1",
"M2", "M3"))

Confidence set for the best model

Method:  raw sum of model probabilities

95% confidence set:
```

```
> confset( cand.set = list( model.1, model.2, model.3 ),
           modnames = c( "M1", "M2", "M3" ) )

Confidence set for the best model

Method:  raw sum of model probabilities

95% confidence set:
   K   AICc Delta_AICc AICcWt
M2 7 575.34       0.00   0.50
M3 8 575.57       0.24   0.44
M1 7 579.46       4.13   0.06

Model probabilities sum to 1
```

- Default confidence is 0.95, but can be changed by including argument `level=`

Evidence Ratio

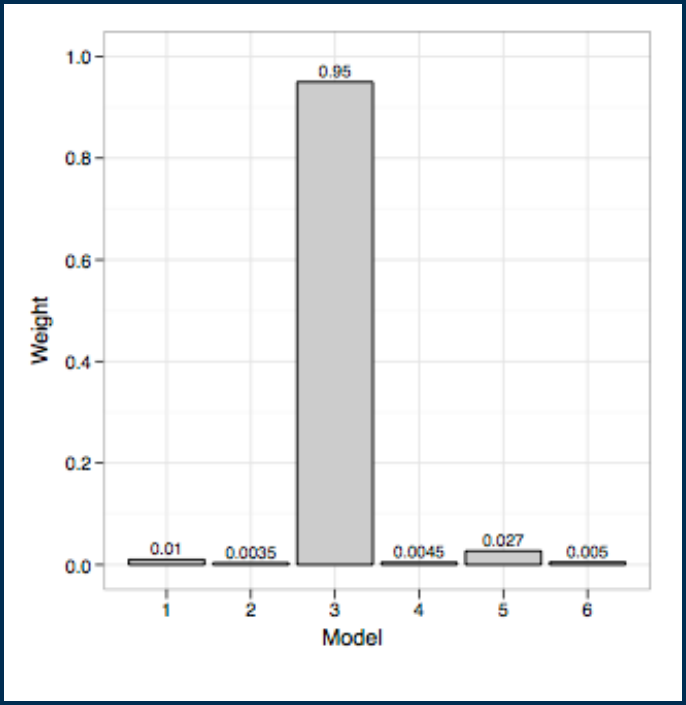
- Evidence ratio expresses difference between best-fitting model and worse-fitting model in terms of odds
- For the h^{th} model this is

$$E_h = \frac{W_{max}}{W_h}$$

Since E_h is ratio of two probabilities, it can be interpreted as the odds that the worse-fitting model is *not* the best approximating model

The higher the odds, the more likely the worse-fitting model is not the best approximating model

- Reference for the evidence ratio is 1 as this is the odds for the best-fitting model
- Again consider the plots



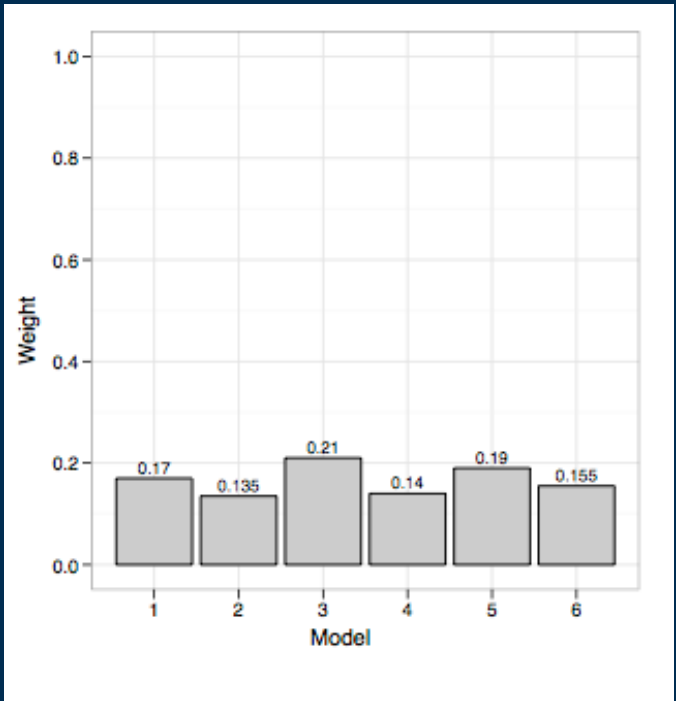
Second best-fitting model is Model 5

$$E_h = \frac{0.95}{0.027} = 35.19$$

Model 3 has a weight of evidence over 35 times as great as Model 5 (the second best-fitting model)

Odds of 35 to 1 that Model 5 is not the best approximating model

- Reference for the evidence ratio is 1 as this is the odds for the best-fitting model
- Again consider the plots



Second best-fitting model is Model 5

$$E_h = \frac{0.21}{0.19} = 1.11$$

Model 3 has a weight of evidence only slightly greater thanModel 5 (the second best-fitting model)

Odds of 1.11 to 1 that Model 5 is not the best approximating model

- A single E_h is computed using the `evidence()` function in **AICcmodavg**
- The input to this function is output from the `aictab()` function

```
> myaicc <- aictab( cand.set = list( model.1, model.2, model.3 ),
                    modnames = c( "M1", "M2", "M3" ), sort = FALSE )
> evidence( myaicc )

Evidence ratio between models ' M2 ' and ' M3 ':
1.12
```

- Default returns evidence ratio between best-fitting and second best-fitting model
- Evidence ratios for other comparisons are made using the `model.high=` and `model.low=` arguments

```
> evidence( myaicc, model.low = "M1" )

Evidence ratio between models ' M2 ' and ' M1 ':
7.87

> evidence( myaicc, model.high = "M3", model.low = "M1" )

Evidence ratio between models ' M3 ' and ' M1 ':
7
```

- Model 2 is the best-fitting model (E = 1).
- Compared to Model 2, Model 3 is only slightly less likely to be the best approximating model (E = 1.12).
- Compared to Model 2, Model 1 is over seven times less likely to be the best approximating model (E = 7.87).
- Compared to Model 3, Model 1 is seven times less likely to be the best approximating model (E = 7.00).

- How large the evidence ratio needs to be to eliminate models as candidates for the best approximating model is a judgment for the researcher
- The model uncertainty should always be acknowledged
- One potential aid in making judgments is to use the parametric bootstrap to indicate values of the evidence ratio from the best approximating model that could potentially occur because of random sampling (see optional reading for this unit)

AICc and Multimodal Inference

- The effect size measures based on the AICc are used to assign a value to each model in a set of candidate models
  - Information for all of the models considered can be presented
  - This is known as *multimodal inference*.
- Once a set of LMER models is formed, the models are evaluated to:
  - Rank order the models in terms of plausibility
  - Determine relative effect size
- The effect size measures help researcher to make judgment about plausibility of models
- It is expected that after this part of the analysis, the set of candidate models will be revised
- With new data, this leads to evolution of the model set amenable to a program of research rather than *one-and-done* analyses

- Intention of multimodal inference is to assemble a portfolio of plausible models that are subject to continual change
- This goal is very different from NHST
- NHST vs. Multimodal Inference

Issues	AICc	NHST
Number of models compared	Many	2
Nested models	Not required	Required
Use of cutoff values	No	Yes
Basis for effect size measure	Yes	No
Assumption that one model is true model	No	Yes
Provides evidence for a model	Yes	No

Example of Multimodal Inference

- We will now consider some other models. These are all estimated using the `lmer()` function using the `mpls.l` data frame.
- Predata models should be formulated according to the researchers' working hypotheses
- In translating working hypotheses to LMER models, the following guidelines may be considered to ensure interpretable results
  - For all the predictors in the LMER model (time varying and static) all higher order terms should have the associated lower order terms also included in the model
  - No higher numbered random effects should appear without the inclusion of all lower numbered random effects
  - No random effect should appear in the model without its associated fixed effect (the converse is allowable)

- RQ: To what extent does risk status act as a proxy for ethnic effects, or vice versa?
- Consider LMER models that represent different plausible proxy scenarios
  - Model including both risk and ethnicity as static predictors (both predictors have non-negligible effects)
  - Model only including risk (negligible effect of ethnicity)
  - Model only including ethnicity (negligible effect of risk)
- RQ: What is the duration of the achievement gap? Does it persist? Does it close? Or widen?
  - The 'enduring gap' hypothesis is consistent with a model that has intercept only differences
  - Closing or widening of the gap is consistent with a model that has both intercept and slope differences

- The change curve and random effects for the analysis are considered to be preselected
- Here a linear change curve is selected for both the group and individual levels.
- Random intercept and slope terms will be included for every model (we will address this in future notes)
- Considering all possible combinations of static predictors (risk, ethnicity, both) and the achievement gap (constant, changing) six models are proposed for the analysis

Model	Static Predictors	Gap	Fixed-Effects Structure
1	Risk	Constant	$\beta_0 + \beta_1(\text{grade5}_{ij}) + \beta 2(\text{dadv}_i)$
2	Ethnicity	Constant	$\beta_0 + \beta_1(\text{grade5}_{ij}) + \beta 2(\text{eth2}_i)$
3	Risk, ethnicity	Constant	$\beta_0 + \beta_1(\text{grade5}_{ij}) + \beta 2(\text{dadv}_i) + \beta 3(\text{eth2}_i)$
4	Risk	Changing	$\beta_0 + \beta_1(\text{grade5}_{ij}) + \beta 2(\text{dadv}_i) + \beta 3(\text{grade5}_{ij} \cdot \text{dadv}_i)$
5	Ethnicity	Changing	$\beta_0 + \beta_1(\text{grade5}_{ij}) + \beta 2(\text{eth2}_i) + \beta 3(\text{grade5}_{ij} \cdot \text{eth2}_i)$
6	Risk, ethnicity	Changing	$\beta_0 + \beta_1(\text{grade5}_{ij}) + \beta 2(\text{dadv}_i) + \beta 3(\text{eth2}_i) + \beta 4(\text{grade5}_{ij} \cdot \text{dadv}_i) + \beta 5(\text{grade5}_{ij} \cdot \text{eth2}_i)$

- Models are consistent based on earlier constraints
  - Models differ only in their fixed effects
    - Models have same response (Y_{ij})
    - Models have same change curve (linear)
    - Models have same two random effects (b_{0i} and b_{1i})
- There are pairs of models that are nested (e.g., Model 2 within Model 3)
- There are pairs of models not nested (e.g., Model 1 and Model 2)

Estimate each of the six models

```
> model.1 <- lmer( read ~ 1 + grade5 + dadv + ( 1 + grade5 | subid ),
                   data = mpls.l, REML = FALSE )

> model.2 <- lmer( read ~ 1 + grade5 + eth2  + ( 1 + grade5 | subid ),
                   data = mpls.l, REML = FALSE)

> model.3 <- lmer( read ~ 1 + grade5 + dadv + eth2 + ( 1 + grade5 | subid ),
                   data = mpls.l, REML = FALSE)

> model.4 <- lmer( read ~ 1 + grade5 + dadv + grade5:dadv +
                   ( 1 + grade5 | subid ), data = mpls.l, REML = FALSE )

> model.5 <- lmer( read ~ 1 + grade5 + eth2 + grade5:eth2  +
                   ( 1 + grade5 | subid ), data = mpls.l, REML = FALSE )

> model.6 <- lmer( read ~ 1 + grade5 + dadv + eth2 + dadv:grade5 +
                   eth2:grade5 + ( 1 + grade5 | subid ), data = mpls.l,
                   REML = FALSE )
```

Obtain model effect sizes

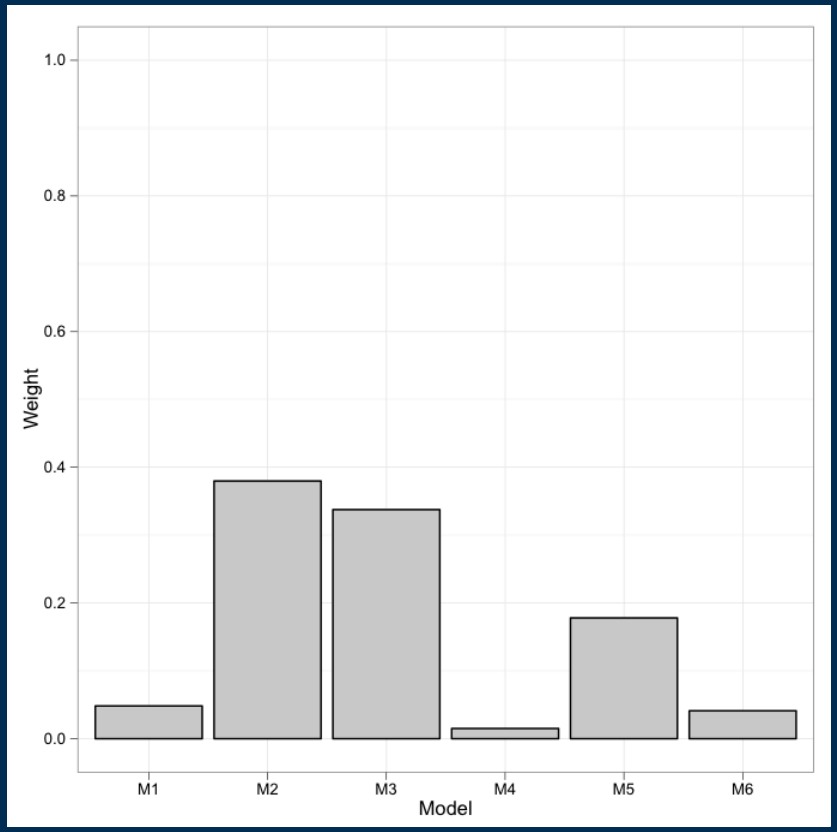
```
> myaicc <- aictab( cand.set = list( model.1, model.2, model.3, model.4,
                                    model.5, model.6 ),
                    modnames = c( "M1", "M2", "M3", "M4", "M5", "M6" ) )
> myaicc

Model selection based on AICc :

    K   AICc Delta_AICc AICcWt Cum.Wt      LL
M2  7 575.34       0.00   0.38   0.38 -279.89
M3  8 575.57       0.24   0.34   0.72 -278.77
M5  8 576.85       1.52   0.18   0.90 -279.41
M1  7 579.46       4.13   0.05   0.94 -281.95
M6 10 579.78       4.45   0.04   0.98 -278.30
M4  8 581.79       6.46   0.02   1.00 -281.88

## compute evidence ratio
> ERatio <- max(myaicc$AICcWt) / myaicc$AICcWt
> ERatio
[1]  1.000000  1.124921  2.133413  7.870225  9.232390 25.225941
```

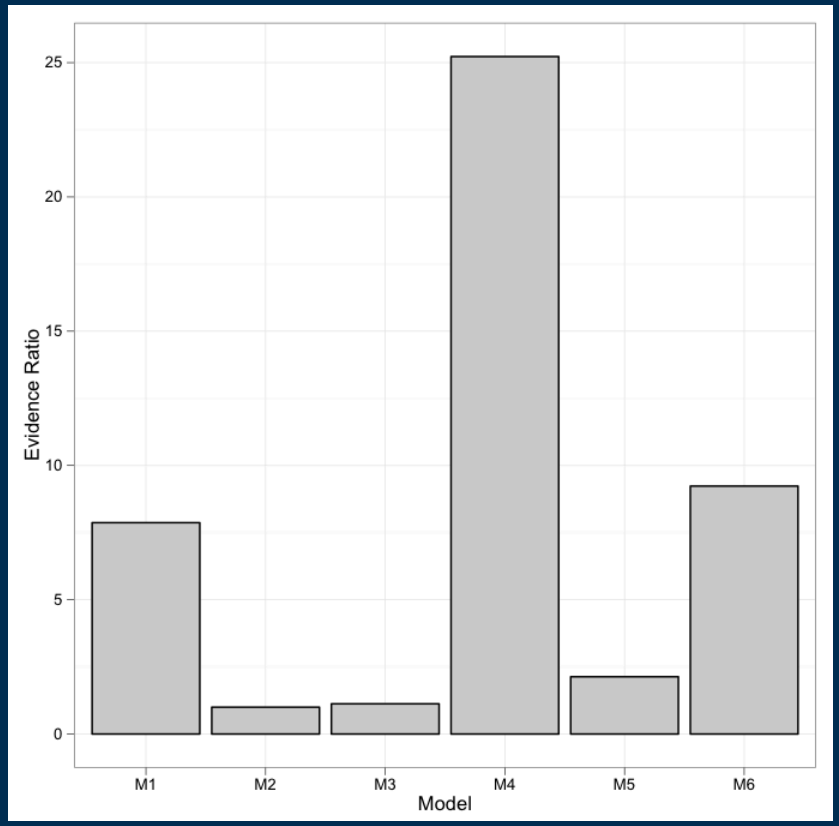
Add ERatio to the myaicc data frame

```
> myaicc <- data.frame( myaicc, ERatio )
> myaicc
```

	Modnames	K	AICc	Delta_AICc	ModelLik	AICcWt	LL	Cum.Wt	ERatio
2	M2	7	575.3352	0.0000000	1.00000000	0.37983803	-279.8898	0.3798380	1.000000
3	M3	8	575.5707	0.2354255	0.88895137	0.33765753	-278.7712	0.7174956	1.124921
5	M5	8	576.8507	1.5154459	0.46873253	0.17804244	-279.4113	0.8955380	2.133413
1	M1	7	579.4614	4.1261733	0.12706117	0.04826267	-281.9529	0.9438007	7.870225
6	M6	10	579.7807	4.4454358	0.10831432	0.04114190	-278.2961	0.9849426	9.232390
4	M4	8	581.7910	6.4557457	0.03964173	0.01505744	-281.8814	1.0000000	25.225941

```
## bar graph
> ggplot( myaicc, aes (x = Modnames, y = AICcWt ) ) +
    theme_set( theme_bw() ) +
    scale_x_discrete( name = "Model" ) +
    scale_y_continuous( limits = c(0,1), name = "Weight" ) +
    geom_bar( fill = "grey80", color = "black" )
```


1.0
0.8
0.6
0.4
0.2
0.0
Weight
M1
M2
M3
M4
M5
M6
Model


25
20
15
10
5
0
Evidence Ratio
M1
M2
M3
M4
M5
M6
Model

- Models 2, 3, and 5 account for 90% of the total probability
- This indicates there is high plausibility (given the data, the candidate models, and the unknown true model) that the best approximating model is among Models 2, 3, and 5.
- Common among these models is the ethnicity intercept effect
- In terms of the RQ, there is strong evidence of real ethnicity group differences, at least intercept differences
- The models with high evidence ratios (Models 4, 6, and 1) are all models that include risk effects
- There is strong evidence that risk is not a proxy for ethnicity

Model Details

- To this point we have examined global fit indices of the models to indicate plausible and non-plausible models
- Specific details (e.g., parameter estimates, *t*-ratios) have not yet been examined–this should be done for the best approximating model at least

Make the `summary()` output a data frame

```
> mytab <- data.frame( summary( model.2 )@coefs )
> mytab
```

```
             Estimate Std..Error   t.value
(Intercept) 195.573533  4.0392971 48.417714
grade5        4.876976  0.7484977  6.515686
eth2         24.872588  6.1181210  4.065396
```

Add confidence intervals to the mytab data frame

```
> mytab$LCI <- mytab$Estimate - 2 * mytab$Std..Error
> mytab$UCI <- mytab$Estimate + 2 * mytab$Std..Error
> round( mytab, 2 )
```

	Estimate	Std..Error	t.value	LCI	UCI
(Intercept)	195.57	4.04	48.42	187.49	203.65
grade5	4.88	0.75	6.52	3.38	6.37
eth2	24.87	6.12	4.07	12.64	37.11

$$E(y_{ij}) = \beta_0 + \beta_1(\text{grade5}_{ij}) + \beta 2(\text{eth2}_i)$$

$$E(y_{ij}) = \beta_0 + \beta_1(\text{grade5}_{ij})$$ if eth2 = 0 (non-white)

$$E(y_{ij}) = (\beta_0 + \beta_2) + \beta_1(\text{grade5}_{ij})$$ if eth2 = 1 (white)

- The β_2 estimate is 24.87, indicating the white group intercept is higher than the non-white group intercept by the stated amount
- The CI is [12.64, 37.11]. The point estimate just offered should be tempered by the poor statistical reliability shown in the CI
- The β_1 estimate suggests there is a yearly change of 4.88 reading achievement points for both groups (this should again be tempered by the CI)
- To graphically show this we need to first create a new data frame that contains the values of he predictors used in the model estimation (the `@frame` slot of the model) and the fitted values from the model (postmultiply the design matrix by the vector of fixed effects).

Create data frame to make plot

```
> fitted <- model.matrix( model.2 ) %*% fixef( model.2 )

> myplotdata <- data.frame( model.2@frame, fitted )

> head( myplotdata )

  read grade5 eth2 subid   fitted
1  172      0    0     1 195.5735
2  185      1    0     1 200.4505
3  179      2    0     1 205.3275
4  194      3    0     1 210.2045
5  200      0    0     2 195.5735
6  210      1    0     2 200.4505
```

Un-center grade5

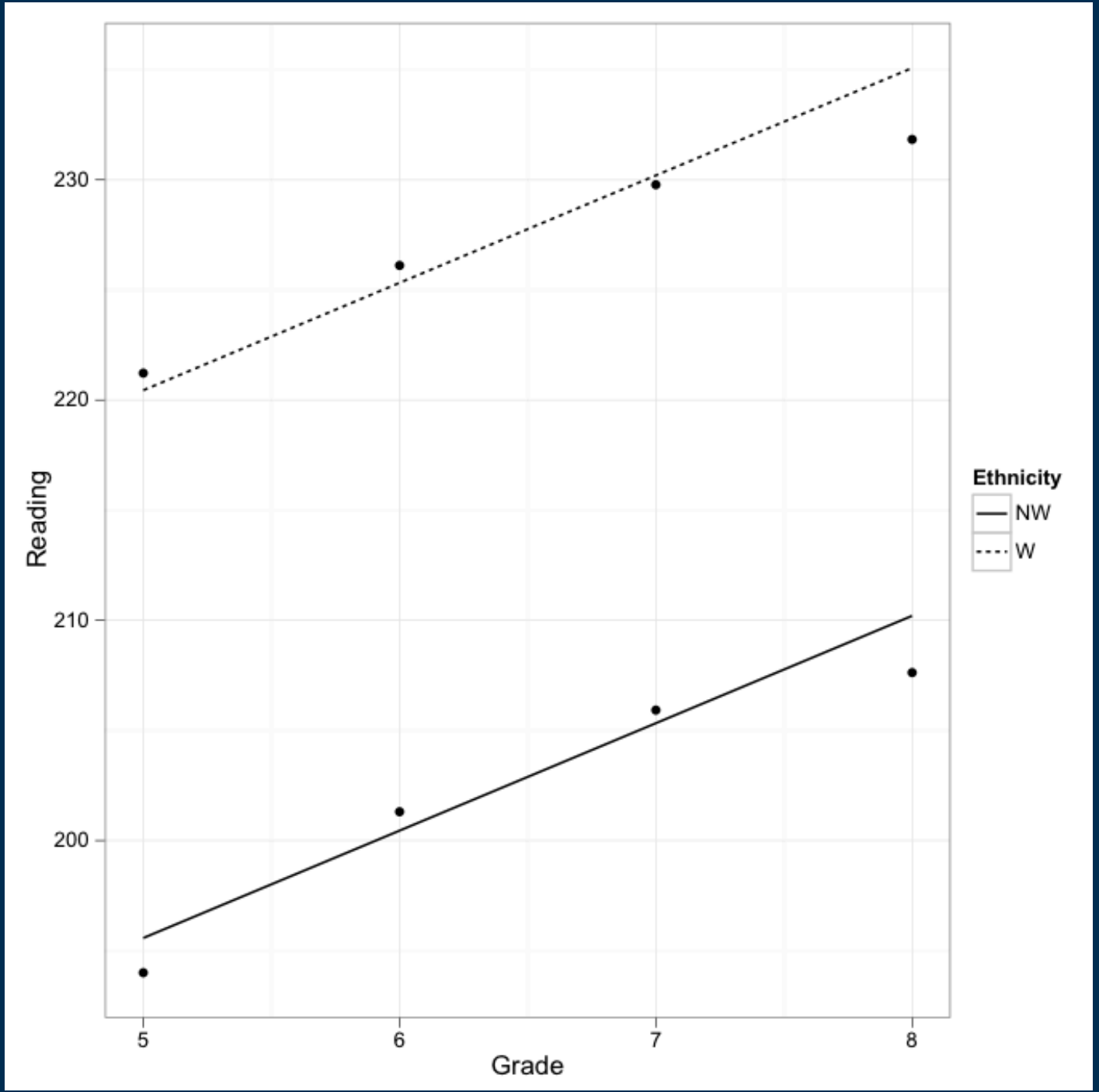
```
> myplotdata$grade <- myplotdata$grade5 + 5
```

Make eth2 a factor

```
> myplotdata$Ethnicity <- as.factor( myplotdata$eth2 )

> levels( myplotdata$Ethnicity ) <- c( "NW", "W" )
```

Plot

```
> ggplot(myplotdata, aes( x = grade, y = read, linetype = Ethnicity ) ) +
    stat_summary( fun.y = "mean", geom = "point", cex = 2 ) +
    stat_summary( aes( y = fitted ), fun.y = "mean", geom = "line" ) +
    theme_bw() +
    scale_x_continuous( name = "Grade", breaks = 5:8 ) +
    scale_y_continuous( name =  "Reading")
```


230
220
210
200
Reading
5
6
7
8
Grade
Ethnicity
NW
W

Post Hoc Models

- Create another set of models after examining the analyses
- Present results from initial analyses (w/o post hoc model set)
- Present another table of results w/post hoc model set included
- Use additional care regarding post hoc models–all indices of global fit should be taken less seriously by informally downgrading them