Multilevel Models and Random Effects

Claudio Mazzi
Department of Computer Science, University of Pisa
MeS Laboratory, Sant'Anna School for Advanced Studies, Pisa
`claudio.mazzi@santannapisa.it`

Applied Statistical Modelling 1
A.A. 2024-2025

Introduction

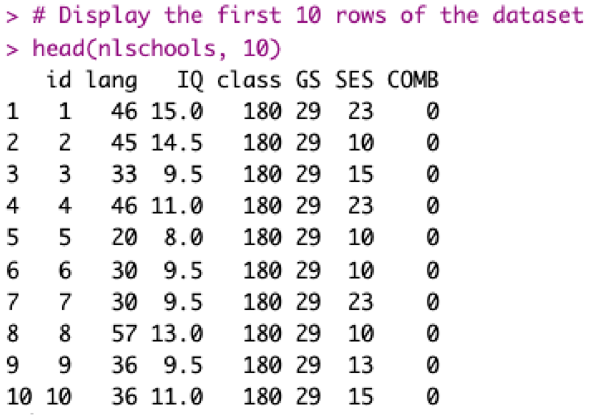
This document explores the use of multilevel models and random effects in analyzing hierarchical data. The dataset under study contains information on language test scores among eighth-grade students (13-14 years old) across 133 classes. The analysis incorporates both student-level and class-level variables to understand the variability in language test scores.

1 Dataset Description

The dataset comprises 2287 observations with the following variables:

- **lang**: Language test score (response variable).
- **IQ**: It refers to Verbal IQ (student-level independent variable), which is a measurement of the ability to understand and reason using concepts framed in words. More broadly, it is linked to problem-solving, abstract reasoning, and working memory.
- **SES**: Socio-economic status of the student's family (student-level independent variable).
- **CS**: Class size (class-level independent variable).
- **COMB**: Indicator for multi-grade teaching (class-level independent variable, 0/1). Multi-grade teaching refers to a teaching method where a single teacher educates students from multiple grades within the same classroom. This approach is common in resource-limited settings, such as rural schools, where the number of students per grade is too small to justify separate classes.

A brief preview of the dataset after preprocessing is shown in Figure 1.



```
> # Display the first 10 rows of the dataset
> head(nlschools, 10)
   id lang   IQ class GS SES COMB
1   1   46 15.0   180 29  23    0
2   2   45 14.5   180 29  10    0
3   3   33  9.5   180 29  15    0
4   4   46 11.0   180 29  23    0
5   5   20  8.0   180 29  10    0
6   6   30  9.5   180 29  10    0
7   7   30  9.5   180 29  23    0
8   8   57 13.0   180 29  10    0
9   9   36  9.5   180 29  13    0
10 10   36 11.0   180 29  15    0
```

Figure 1: Header of the pre-processed database; first 10 rows.

Listing 1: Loading and Pre-processing dataset

```
# Load necessary libraries
library(nlme)       # For mixed-effects models
library(ggplot2)    # For plots
library(dplyr)      # For data manipulation
library(corrr)      # For correlation matrix
library(readr)      # For data import
library(tidyr)      # For reshaping data
library(ggcorrplot)
library(ggplot2)

# Import the file ’nlschools.csv’
nlschools <- read.csv("/path_to/nlschools.csv")

# Add labels to the variables
colnames(nlschools) <- c("id", "lang", "IQ", "class", "CS", "SES", "COMB")
labels <- c(
  "lang" = "Language test score",
  "IQ" = "Verbal IQ",
  "class" = "Class ID",
  "CS" = "Class size",
  "SES" = "Family socio-economic status",
  "COMB" = "Dummy of multi-grade class"
)

# Display the first 10 rows of the dataset
head(nlschools, 10)

# Correlation matrix
nls_corr <- cor(nlschools, method = "kendall")
ggcorrplot(nls_corr, type = "lower", lab = TRUE, lab_size = 3)
```

2 Exploratory Data Analysis

We begin by visualizing the distribution of students and language tests across classes and analyzing the variability at both student and class levels. As usual, it is useful to perform also the correlation matrix to display possible correlations between variables, Figure 2.

Listing 2: Exploratory Data Analysis

```
# Frequency distribution of ’class’
class_distribution <- nlschools %>%
  count(class) %>%
  arrange(desc(n))
print(class_distribution)

# Descriptive statistics
summary(nlschools)
```

Now, we show the boxplot that highlights the variability in scores across classes and the potential influence of class-level factors. For practical purposes, we filtered the data binning the variable *class* with a scale sep equal to 15, as shown in the following code. The boxplot analysis, shown in Figure 3, exhibits how some classes have a narrow interquartile range, indicating consistent scores, while others show broader distributions, suggesting greater variability among students. Median scores differ noticeably across classes, reflecting disparities in overall performance. Outliers are present in many groups, particularly at the lower end of the score spectrum, suggesting individual underperformance. These results may point to differences in teaching methods, curriculum difficulty, or class composition for example.

Listing 3: Boxplot for *Language Test Scores* vs *class*

```
# Boxplot of ’lang’ by ’class’
filtered_data <- nlschools %>%
  filter(as.numeric(class) %% 15 == 0) # binning
```

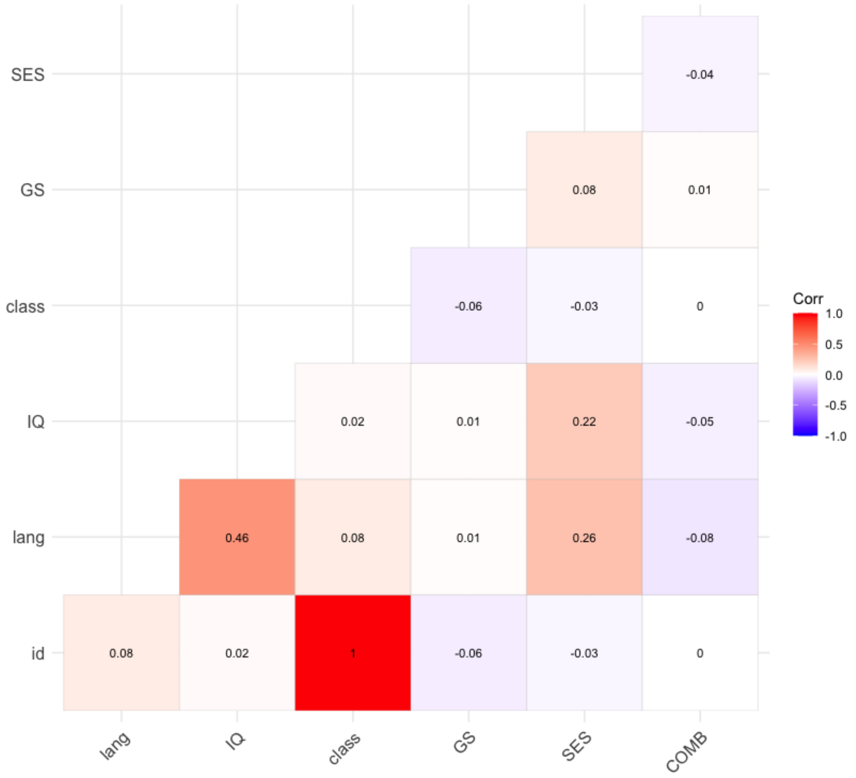


Figure 2: Correlation Matrix showing Kendall correlation across variables.

```

ggplot(filtered_data, aes(x = factor(class), y = lang, fill = factor(class))) +
  geom_boxplot(outlier.shape = 21, outlier.color = "red", alpha = 0.7) + # Outlier
  scale_fill_viridis_d(option = "C") + # Palette
  labs(
    title = "Boxplots of Language Test Scores by Class (Filtered)",
    x = "Class (Filtered)", y = "Language Test Score"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1), #
    legend.position = "none" #
  )
```

3 Comparison of Linear Mixed-Effects Models

This section presents and compares three linear mixed-effects models fitted to analyze language test scores (*lang*) across different classes (*class*) using the `nlme` package in R. The models differ in complexity, ranging from an empty model to a random-coefficient model. In the following lines it is shown the entire code for the three regression models in R, then we will comment on and compare the results from `summary` and test efficiency with different metrics.

Listing 4: Linear Mixed-Effects Models

```
# Empty model
empty_model <- lme(lang ~ 1, random = ~ 1 | class, data = nlschools, method = "REML")
summary(empty_model)

# Random-intercept model with IQ
random_intercept_model <- lme(
```

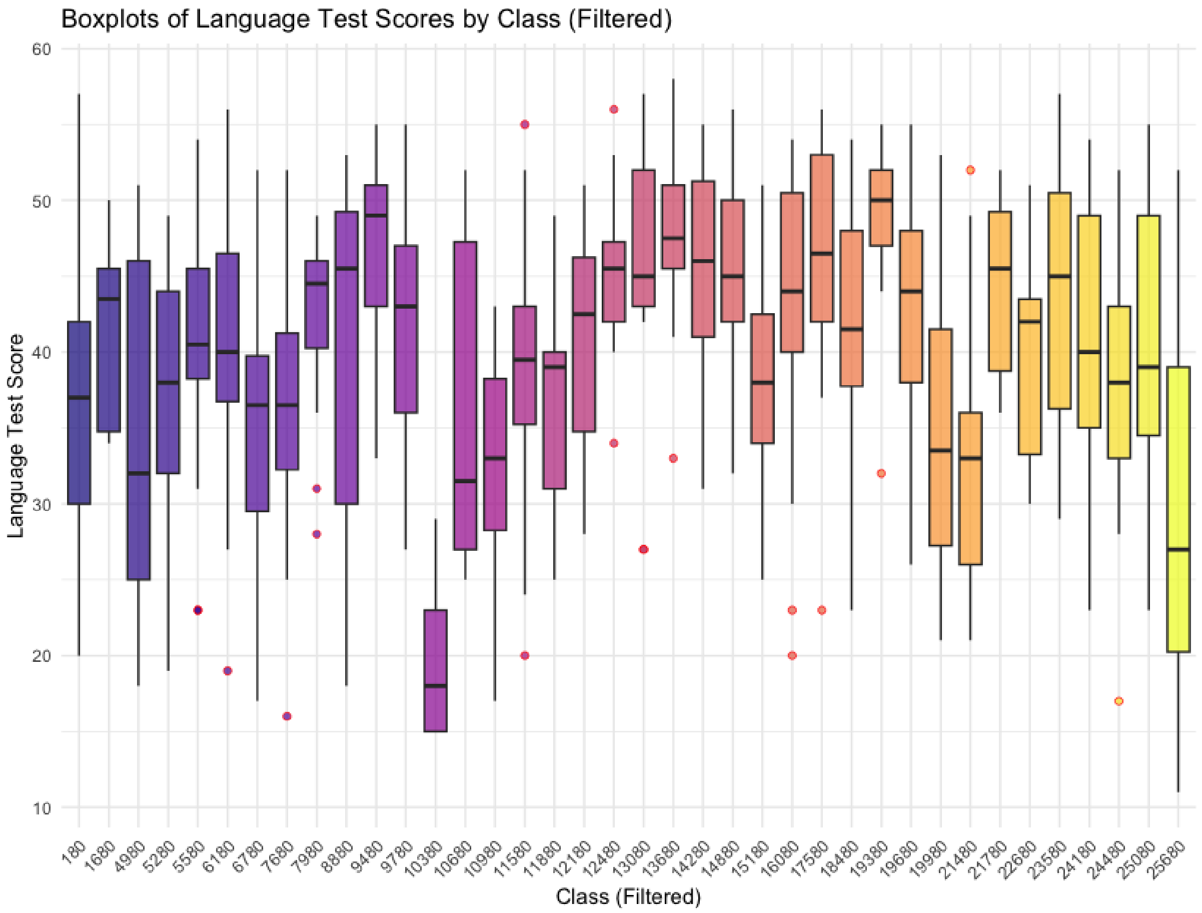


Figure 3: Boxplots of Language Test Scores by Class. The gradient color represents the different amount of observables for each class.

```
  fixed = lang ~ IQ,                    # Fixed effects
  random = ~ 1 | class,                 # Random intercept for 'class'
  data = nlschools,                     # Dataset
  method = "REML"                       # Restricted Maximum Likelihood
)
summary(random_intercept_model)

# Random-coefficient model with IQ and SES
# Random-coefficients model using nlme
random_coeff_model <- lme(
  fixed = lang ~ IQ + SES,           # Fixed effects
  random = ~ IQ | class,             # Random intercept and slope for 'IQ' by '
      class'
  data = nlschools,                  # Dataset
  method = "REML"                    # Restricted Maximum Likelihood
)

summary(random_coeff_model)
```

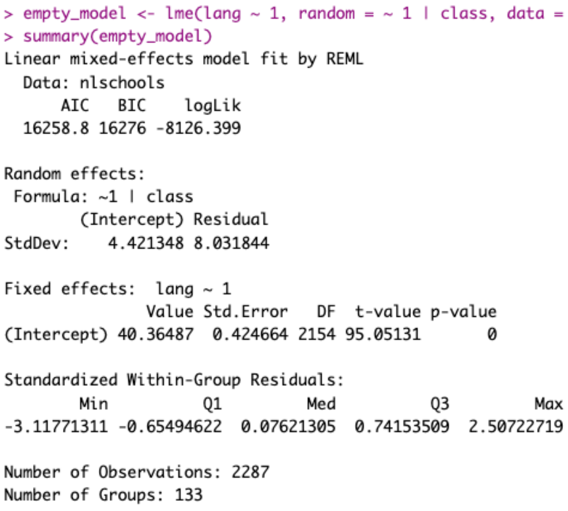
3.1 Empty Model

The empty model is specified as:

$$\texttt{lang}_{ij} = \beta_0 + u_j + \epsilon_{ij}$$

where β_0 is the fixed intercept, $u_j \sim N(0, \sigma_u^2)$ represents the random effect for class j, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is the residual error. This model assumes that all variability is partitioned into between-class and within-class components.

Summary: Figure 4 The fixed intercept is estimated as 40.36, representing the overall average language score across all classes. The random effect for *class* has a standard deviation of 4.42, indicating moderate between-class variability, while the residual standard deviation is 8.03, suggesting substantial within-class variability. This model serves as a baseline to assess the impact of adding predictors.



```
> empty_model <- lme(lang ~ 1, random = ~ 1 | class, data = nlschools, method = "REML")
> summary(empty_model)
Linear mixed-effects model fit by REML
  Data: nlschools
      AIC   BIC    logLik
  16258.8 16276 -8126.399

Random effects:
 Formula: ~1 | class
        (Intercept) Residual
StdDev:    4.421348 8.031844

Fixed effects:  lang ~ 1
               Value Std.Error   DF  t-value p-value
(Intercept) 40.36487  0.424664 2154 95.05131       0

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max 
-3.11771311 -0.65494622  0.07621305  0.74153509  2.50722719 

Number of Observations: 2287
Number of Groups: 133
```

Figure 4: Empty Model Summary

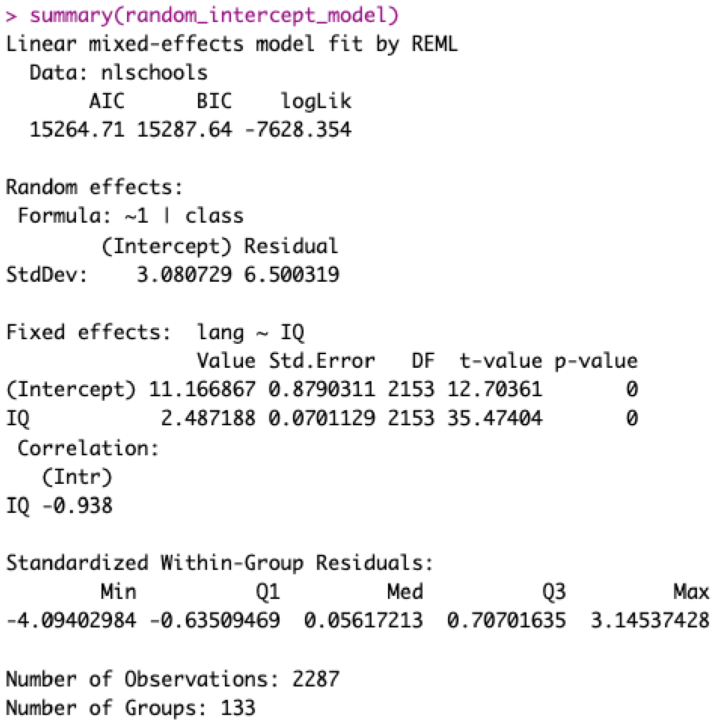
3.2 Random-Intercept Model with *IQ*

The second model incorporates `IQ` as a fixed predictor while maintaining a random intercept for `class`:

$$\texttt{lang}_{ij} = \beta_0 + \beta_1 \cdot \texttt{IQ}_{ij} + u_j + \epsilon_{ij}$$

This model assumes that the relationship between *IQ* and *lang* is constant across all classes.

Summary: Figure 5 The fixed effect of *IQ* is highly significant ($p < 0.001$), with an estimate of 2.49, indicating that each unit increase in *IQ* corresponds to an average increase of 2.49 points in language scores. The random intercept for *class* has a reduced standard deviation of 3.08, suggesting that part of the between-class variability is explained by differences in *IQ*. The residual standard deviation remains high at 6.50, highlighting substantial within-class variability.



```
> summary(random_intercept_model)
Linear mixed-effects model fit by REML
  Data: nlschools
       AIC      BIC    logLik
  15264.71 15287.64 -7628.354

Random effects:
 Formula: ~1 | class
        (Intercept) Residual
StdDev:    3.080729 6.500319

Fixed effects:  lang ~ IQ
                Value Std.Error   DF  t-value p-value
(Intercept) 11.166867 0.8790311 2153 12.70361       0
IQ           2.487188 0.0701129 2153 35.47404       0
 Correlation: 
   (Intr)
IQ -0.938

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max 
-4.09402984 -0.63509469  0.05617213  0.70701635  3.14537428 

Number of Observations: 2287
Number of Groups: 133 
```

Figure 5: Summary of Random-Intercept Model with *IQ* as *Fixed* Effect.

3.3 Random-Coefficient Model with *IQ* and *SES*

The most complex model includes both `IQ` and `SES` as fixed effects and allows the slope of `IQ` to vary across classes:

$$\texttt{lang}_{ij} = \beta_0 + \beta_1 \cdot \texttt{IQ}_{ij} + \beta_2 \cdot \texttt{SES}_{ij} + u_{j0} + u_{j1} \cdot \texttt{IQ}_{ij} + \epsilon_{ij}$$

where u_{j0} and u_{j1} represent the random intercept and slope for *IQ*, respectively.

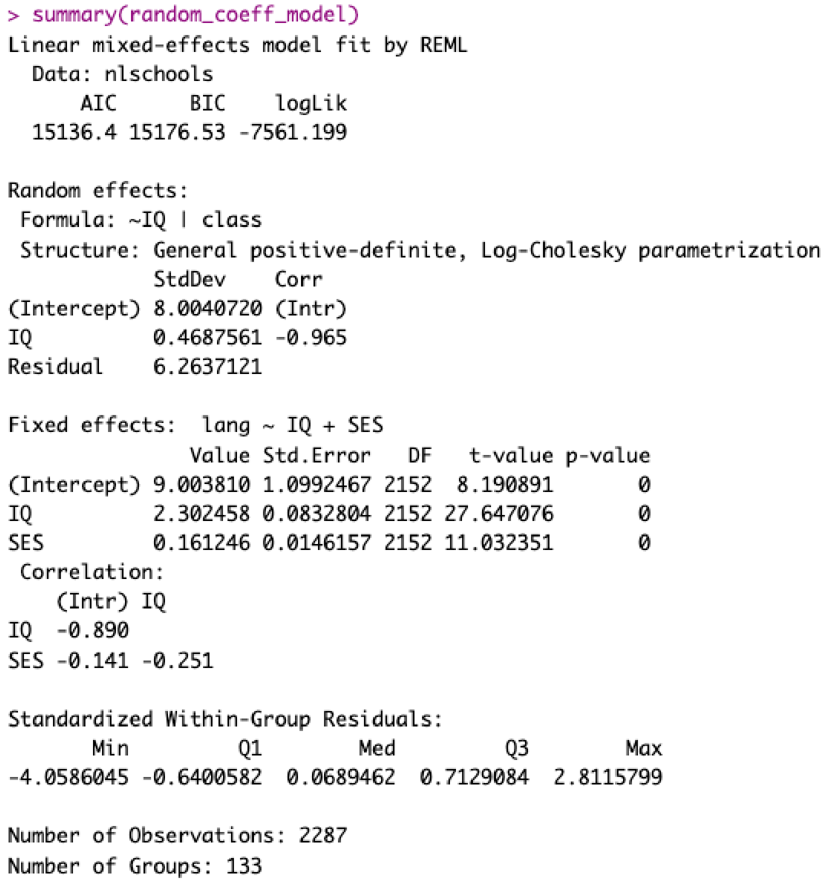
Summary: Figure 6 Both *IQ* and *SES* are significant predictors ($p < 0.001$)ì. The fixed effect for *IQ* is 2.30, and for *SES*, it is 0.16, indicating that higher *SES* is associated with a modest improvement in language scores. The random intercept for *class* has a standard deviation of 8.00, and the random slope for *IQ* has a standard deviation of 0.47, with a strong negative correlation -0.965 between intercept and slope. This suggests that classes with higher average scores tend to have a weaker relationship between *IQ* and *lang*. The residual standard deviation is 6.26, reflecting improved model fit compared to the previous models.

To better comprehend the results of the random-coefficient model we compute predictions and then plot them, as shown in Figure 7. The related R script is listed below; goal is to understand the variability in random intercepts across classes and assess their distribution relative to the overall population.

Listing 5: Predictions Plot for the Random-Coefficient Model with *IQ* and *SES*

```
# Extract random effects
random_effects <- ranef(random_coeff_model)
random_effects_df <- as.data.frame(random_effects)
colnames(random_effects_df) <- c("Intercept", "IQ")
random_effects_df$class <- rownames(random_effects)

# Dev Std
var_corr <- VarCorr(random_coeff_model)
sd_intercept <- sqrt(as.numeric(var_corr["(Intercept)", "Variance"]))
random_effects_df$sd.Intercept <- sd_intercept
```



```
> summary(random_coeff_model)
Linear mixed-effects model fit by REML
  Data: nlschools
      AIC      BIC    logLik
  15136.4 15176.53 -7561.199

Random effects:
 Formula: ~IQ | class
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 8.0040720 (Intr)
IQ          0.4687561 -0.965
Residual    6.2637121

Fixed effects:  lang ~ IQ + SES
               Value Std.Error   DF   t-value p-value
(Intercept) 9.003810 1.0992467 2152  8.190891       0
IQ          2.302458 0.0832804 2152 27.647076       0
SES         0.161246 0.0146157 2152 11.032351       0
 Correlation:
    (Intr) IQ
IQ  -0.890
SES -0.141 -0.251

Standardized Within-Group Residuals:
       Min         Q1         Med         Q3        Max
-4.0586045 -0.6400582  0.0689462  0.7129084  2.8115799

Number of Observations: 2287
Number of Groups: 133
```

Figure 6: Summary of Random-Intercept Model with *IQ* and *SES* as Fixed Effect.

```


# High-Low plot for random intercepts
random_effects_df <- random_effects_df %>%
  arrange(Intercept)
random_effects_df$id <- seq_along(random_effects_df$Intercept)

filtered_random_effects <- random_effects_df %>%
  filter(row_number() %% 5 == 0)

ggplot(filtered_random_effects, aes(x = id, y = Intercept, color = Intercept)) +
  geom_errorbar(aes(ymin = Intercept - sd.Intercept, ymax = Intercept + sd.
      Intercept), width = 0.5, alpha = 0.7) +
  geom_point(size = 3, alpha = 0.8) +
  scale_color_viridis_c(option = "C") + # Palette
  labs(
    title = "Random Effects Predictions (Filtered)",
    x = "Classes (Filtered)",
    y = "Random Intercepts"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_blank(), # Label x-axis
    axis.ticks.x = element_blank(), # No tick x-axis
    legend.position = "right" # Legenda
  )
```

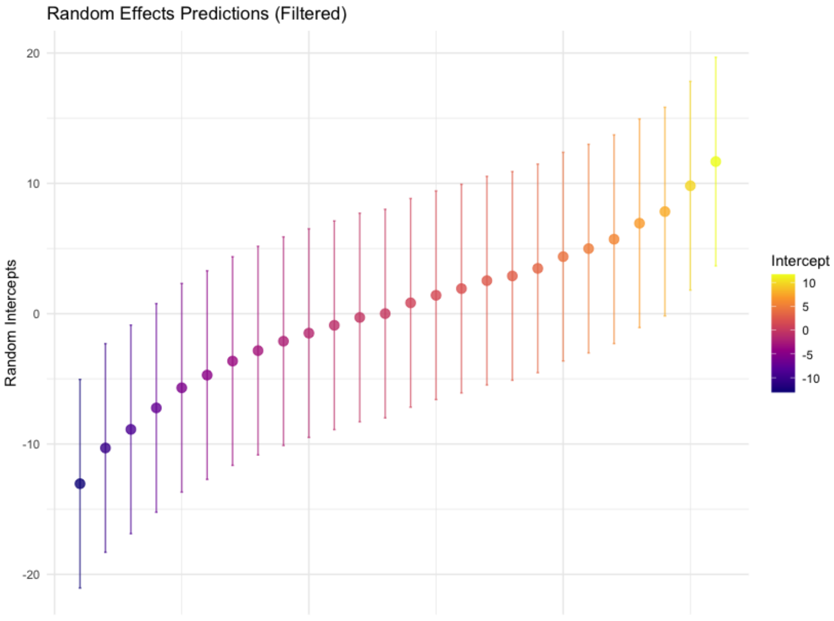


Figure 7: Predicitons plot for the Random-Coefficient Model with *IQ* and *SES*. Each point represents the random intercept for a filtered class. Error bars ($\pm\sigma_{\text{Intercept}}$) capture the variability of the random intercept estimates. A color gradient (`Viridis`) encodes the magnitude of the random intercept, with warmer colors representing higher intercepts.

The plot in Figure 7 illustrates the distribution of random intercepts across a filtered subset of classes, ranked by their values. The following observations emerge:

- The random intercepts range approximately from -20 to +20, indicating substantial variability in the baseline performance of classes after accounting for fixed effects (e.g., *IQ* and *SES*).
- Error bars show the uncertainty in estimating the random intercepts. Classes with larger absolute values of intercepts often have wider error bars, reflecting greater variability.

- The gradient color scale (from purple to yellow) highlights classes with lower and higher random intercepts, respectively. This encoding aids in identifying trends and patterns at a glance.
- Classes with random intercepts above zero correspond to higher-than-average language scores relative to the overall population, while those below zero indicate below-average performance.

This visualization underscores the role of class-level heterogeneity in explaining variability in language test scores. Classes differ markedly in their baseline performance, even after accounting for predictors such as *IQ* and *SES*. These findings justify the inclusion of random effects in the model and highlight the importance of addressing class-specific variability in hierarchical data analysis.

3.4 Model Comparison and Interpretation

The inclusion of *IQ* and *SES* as predictors, along with random slopes, progressively improves the model fit, as evidenced by decreasing AIC and BIC values:

- Empty Model: AIC = 16,258.8, BIC = 16,276
- Random-Intercept Model: AIC = 15,264.71, BIC = 15,287.64
- Random-Coefficient Model: AIC = 15,136.4, BIC = 15,176.53

The random-coefficient model provides the best fit, highlighting the importance of allowing class-specific variability in the relationship between *IQ* and language scores. This suggests that both individual-level predictors (*IQ*, *SES*) and class-level heterogeneity play critical roles in explaining language score variability.

Now we see a last advanced analysis before providing metrics comparison tests.

3.5 Cross-Level Interaction Model

The final model tested introduces a cross-level interaction between *IQ* and *CS* (a class-level predictor) in addition to the individual-level predictors *IQ* and *SES*. The model is specified as follows:

$$lang_{ij} = \beta_0 + \beta_1 \cdot IQ_{ij} + \beta_2 \cdot SES_{ij} + \beta_3 \cdot CS_j + \beta_4 \cdot (IQ_{ij} \cdot CS_j) + u_{j0} + u_{j1} \cdot IQ_{ij} + \epsilon_{ij}$$

where:

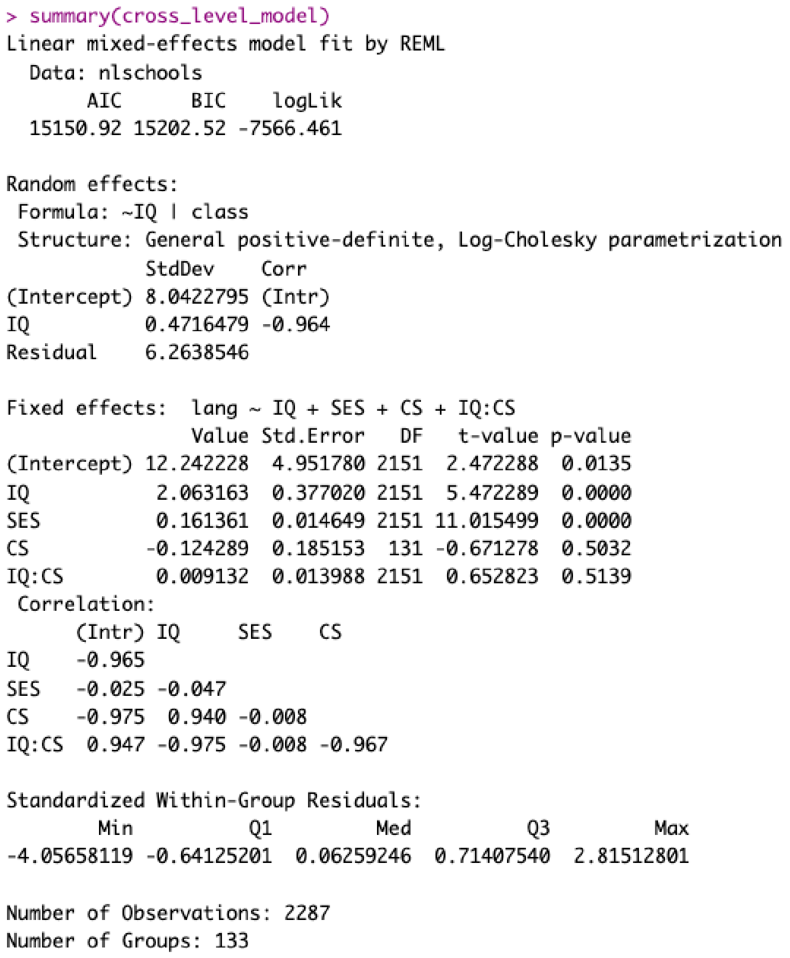
- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are the fixed effects for the intercept, *IQ*, *SES*, *CS*, and the interaction term *IQ:CS*, respectively.
- u_{j0} and u_{j1} are the random intercept and slope for *IQ*, allowing both the baseline and the effect of `IQ` to vary across classes.
- ϵ_{ij} is the residual error term.

Listing 6: Cross-Level Interaction Model

```
cross_level_model <- lme(
  fixed = lang ~ IQ + SES + CS + IQ:CS, # Fixed Effects with interaction
  random = ~ IQ | class,                # Random Effects
  data = nlschools,                     # Dataset
  method = "REML"                       # Restricted Maximum Likelihood
)
summary(cross_level_model)
```

The summary of the model is presented in Figure 8, and the key findings are as follows:

Fixed Effects: The fixed effect of *IQ* ($\beta_1 = 2.06$, $p < 0.001$) is highly significant, indicating a strong positive association between *IQ* and language scores. For each unit increase in *IQ*, the language score increases by 2.06 points on average. The effect of *SES* ($\beta_2 = 0.16$, $p < 0.001$) is also significant, showing that students from higher socio-economic status backgrounds tend to perform slightly better in language tests. Finally, the class-level predictor *CS* $\beta_3 = -0.12$ and its interaction with *IQ* $\beta_4 = 0.009$) are not statistically significant ($p > 0.05$), suggesting that *CS* and its interplay with *IQ* may not have a meaningful impact on language scores within this dataset.



```
> summary(cross_level_model)
Linear mixed-effects model fit by REML
  Data: nlschools
       AIC      BIC    logLik
  15150.92 15202.52 -7566.461

Random effects:
 Formula: ~IQ | class
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 8.0422795 (Intr)
IQ          0.4716479 -0.964
Residual    6.2638546

Fixed effects:  lang ~ IQ + SES + CS + IQ:CS
                Value Std.Error   DF   t-value p-value
(Intercept) 12.242228  4.951780 2151  2.472288  0.0135
IQ           2.063163  0.377020 2151  5.472289  0.0000
SES          0.161361  0.014649 2151 11.015499  0.0000
CS          -0.124289  0.185153  131 -0.671278  0.5032
IQ:CS        0.009132  0.013988 2151  0.652823  0.5139
 Correlation:
      (Intr) IQ     SES    CS
IQ    -0.965
SES   -0.025 -0.047
CS    -0.975  0.940 -0.008
IQ:CS  0.947 -0.975 -0.008 -0.967

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.05658119 -0.64125201  0.06259246  0.71407540  2.81512801

Number of Observations: 2287
Number of Groups: 133
```

Figure 8: Summary of the Cross - Level Interaction Model

Random Effects: The random intercept for *class* ($\sigma_{\text{Intercept}} = 8.04$) shows substantial between-class variability in baseline language scores. While, the random slope for *IQ* ($\sigma_{\text{IQ}} = 0.47$) indicates variability in the effect of *IQ* across classes, with a strong negative correlation (-0.965) between the intercept and slope. This suggests that classes with higher average baseline scores tend to show a weaker relationship between *IQ* and language scores. The residual variance ($\sigma_{\text{Residual}} = 6.26$) remains consistent with previous models, indicating that a large portion of the variability is at the individual level.

In summary, the non-significant interaction term ($p > 0.05$) suggests that the relationship between *IQ* and language scores does not depend on *CS* in this dataset. The significant fixed effects for *IQ* and *SES* reaffirm their importance in explaining language score variability.

However, the strong negative correlation between the random intercept and slope highlights an interesting phenomenon: classes with higher average scores demonstrate a weaker association between *IQ* and performance. This finding suggests that in high-performing classes, other factors beyond *IQ* may contribute to student success, warranting further investigation.

The AIC (15,150.92) and BIC (15,202.52) indicate a slightly better fit than previous models, see Figure 4-6, but with limited improvements. Overall, the cross-level interaction model provides insights into the relationship between individual and class-level predictors, but the added complexity does not substantially enhance the explanatory power of the model.

To provide a strong cross-comparison between models one can perform the following analysis:

Listing 7: Comparision of Models

```
# **************************************************************
# ---------------- COMPARISON OF MODELS ----------------------
# **************************************************************

# Compare empty model and random-coefficient model
# Fixed Effects must be the same if using REML (differently, use ML)
compare_model <- anova(random_intercept_model, random_coeff_model)
print(compare_model)
```

Note that the `ANOVA` test required the same fixed effects. To overwhelm this limitations we have to perform the Multilevel Regression using Maximum Likelihood Estimator `ML`, instead of Restricted Maximum Likelihood Estimator `RML` in the `lme` routine.

4 Conclusion

This study explored the use of multilevel models to analyze hierarchical data, focusing on language test scores among students nested within classes. Starting from an empty model to progressively more complex models, we identified key factors influencing language performance, such as individual-level predictors (*IQ* and *SES*) and class-level heterogeneity. The random-coefficient model demonstrated the best fit, highlighting the importance of allowing the effect of *IQ* to vary across classes. However, the final cross-level interaction model, which tested the interplay between *IQ* and a class-level predictor (*CS*), did not provide substantial improvements, suggesting the limited influence of *CS* within this dataset.

Overall, the results underscore the critical role of both individual and class-level factors in shaping language outcomes, while reinforcing the value of mixed-effects modeling in capturing hierarchical structures and variability across groups.