Assignment 3: Multilevel Analysis of School Effectiveness

Model answer

Preliminary setup

In this assignment, we will use the R packages `ggplot2`, `dplyr`, `emmeans`, `lmerTest`. You can use the script below to automatically install and load them using the `pacman` package.

```
# Check whether pacman is available and install if needed
options(repos = c(CRAN = "https://cloud.r-project.org"))
if (!requireNamespace("pacman", quietly = TRUE)) install.packages("pacman")

# Use pacman to install (if needed) and load the required packages
pacman::p_load(dplyr, ggplot2, emmeans, lmerTest)
```

ILEA School Effectiveness Data

The Inner London Education Authority (ILEA) dataset contains examination records of 15,362 students from 140 secondary schools over the years 1985, 1986, and 1987. This dataset was sourced from the data library of the Centre for Multilevel Modelling at the University of Bristol and is used to examine school effectiveness and the factors that influence student exam scores.

The dataset contains the following variables:

- School: A numeric variable representing the school identifier
- ExamScore: A numeric variable representing the exam score of each student
- PercentFSM: The percentage of students in the school eligible for free school meals (an indicator of socioeconomic status)
- Gender: A categorical variable representing the gender of the student
- VRBand: The verbal reasoning band of the student (VR1, VR2, or VR3)

- SchoolDenomination: The denomination of the school (Maintained, Church of England, Roman Catholic)

Exploratory Data Analysis

Let's load the data and use the `str()` function to inspect the structure of the dataset.

```
# Load the ILEA data
ilea_data <- read.csv("downloads/ilea_data.csv")

# Convert all character variables into factors
ilea_data <- ilea_data %>%
  mutate(across(where(is.character), as.factor)) %>%
  mutate(School = as.factor(School))

# Inspect the structure of the dataset
str(ilea_data)
```

```
'data.frame':   15362 obs. of  6 variables:
 $ School            : Factor w/ 139 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ ExamScore         : int  17 5 16 12 7 20 15 16 26 5 ...
 $ PercentFSM        : int  24 24 24 24 24 24 24 24 24 24 ...
 $ Gender            : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 1 1 1 1 1 ...
 $ VRBand            : Factor w/ 3 levels "VR1","VR2","VR3": 2 2 2 2 2 2 2 3 2 2 ...
 $ SchoolDenomination: Factor w/ 3 levels "Church of England",..: 2 2 2 2 2 2 2 2 2 2 ...
```

This initial inspection makes clear that the ILEA dataset has a hierarchical structure: 15,362 individual students are nested within 140 schools. This is a two-level hierarchical structure with students (level 1) nested within schools (level 2).

Grouping structure

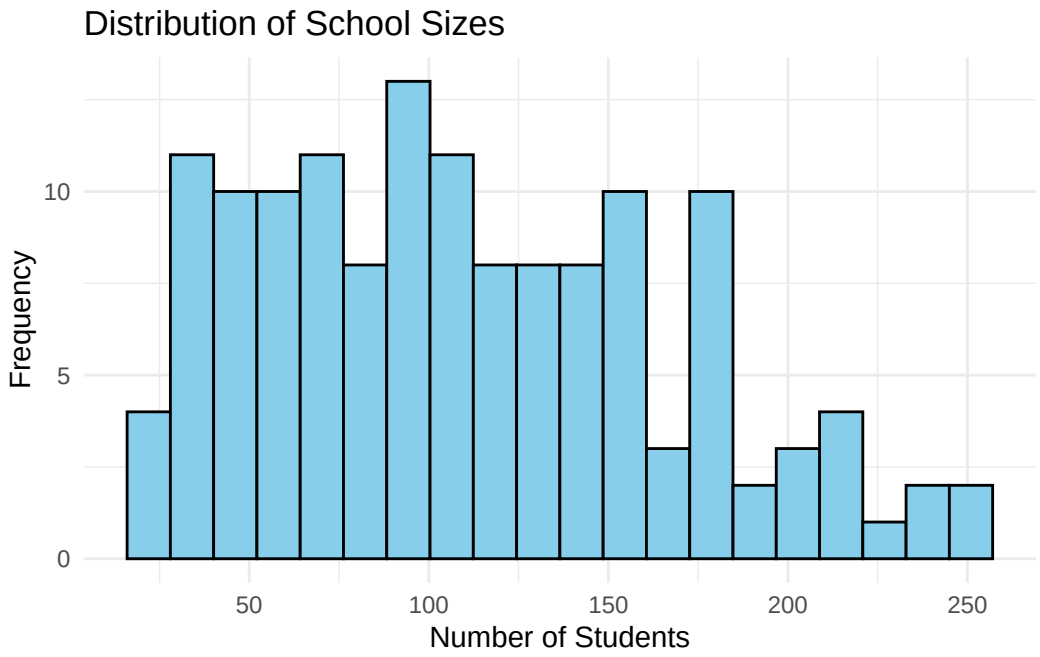
To get a better feeling for the grouping structure, we create a summary table showing the number of students within each school and visualize it with a bar chart:

```
# Create a summary table showing students per school
school_summary <- ilea_data %>%
  group_by(School) %>%
  summarise(num_students = n(), .groups = "drop") %>%
  arrange(desc(num_students))
```

```r
# Display summary statistics for the number of students per school
summary(school_summary$num_students)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   22.0    64.5   103.0   110.5   150.5   251.0
```

```r
# Create a histogram showing the distribution of school sizes
ggplot(school_summary, aes(x = num_students)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(title = "Distribution of School Sizes",
       x = "Number of Students",
       y = "Frequency") +
  theme_minimal()
```

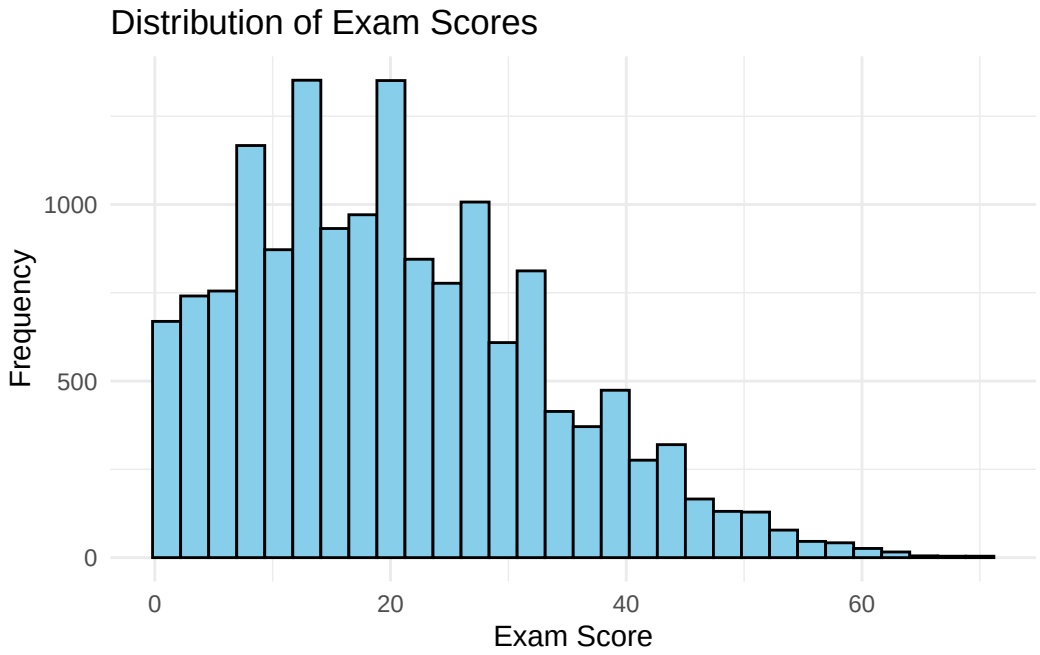


The distribution shows substantial variation in school sizes, with most schools having between 50 and 150 students.

Outcome variable

Next, we create a histogram to visualize the distribution of the outcome variable `ExamScore`:

```r
# Create a histogram of ExamScore
ggplot(ilea_data, aes(x = ExamScore)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Exam Scores",
       x = "Exam Score",
       y = "Frequency") +
  theme_minimal()
```

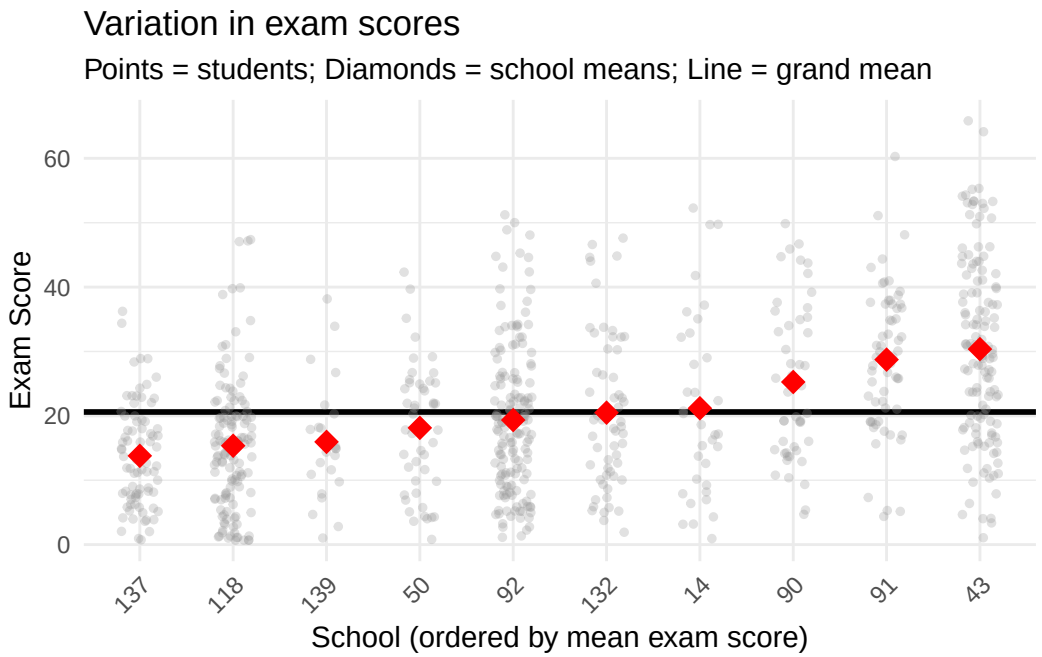


The distribution of exam scores is right-skewed with a notable floor effect: a substantial number of students scored zero or near-zero on the exam. This floor effect occurs because exam scores are bounded at zero - students cannot score below this minimum. Such distributional characteristics may lead to heteroscedasticity in the residuals, which we will examine in the model diagnostics section.

Variance components

We also want to get a sense of the variability in exam scores within and between schools. For this, we are going to randomly select 10 schools and create a scatter plot with the exam score on the y-axis and the school on the x-axis:

```r
# Set seed for reproducibility
set.seed(123)

# Randomly select 10 unique schools
random_schools <- sample(unique(ilea_data$School), size = 10)

# Filter the data to include only the randomly selected schools
random_school_data <- ilea_data %>%
  filter(School %in% random_schools)

# Calculate the average ExamScore for each school
school_avg_score <- random_school_data %>%
  group_by(School) %>%
  summarise(avg_exam_score = mean(ExamScore, na.rm = TRUE), .groups = "drop")

# Calculate the overall grand mean across all schools
grand_mean <- mean(ilea_data$ExamScore, na.rm = TRUE)

# Create scatter plot with jittered observations
ggplot(random_school_data,
       aes(x = reorder(School, ExamScore, FUN = mean),
           y = ExamScore)) +
  geom_hline(yintercept = grand_mean,
             linetype = "solid", color = "black", linewidth = 1) +
  geom_jitter(alpha = 0.3, width = 0.2, size = 1, color = "gray60") +
  geom_point(data = school_avg_score,
             aes(x = School, y = avg_exam_score),
             size = 4, shape = 18, color = "red") +
  labs(title = "Variation in exam scores",
       subtitle = "Points = students; Diamonds = school means; Line = grand mean",
       x = "School (ordered by mean exam score)",
       y = "Exam Score") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



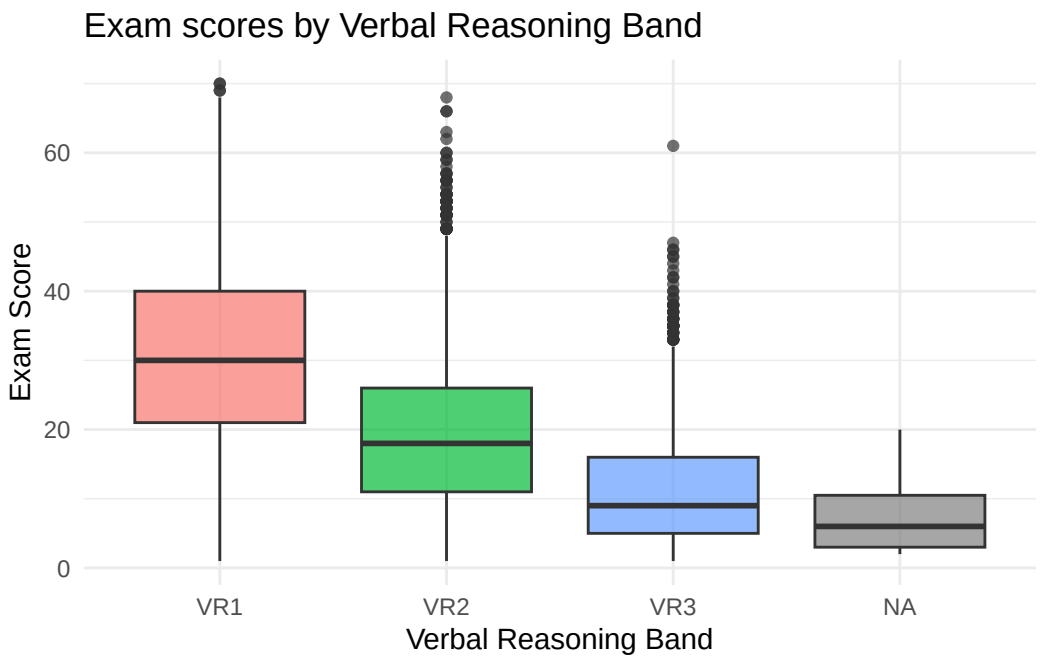
This plot reveals both within-school variability (the spread of individual student scores around their school mean) and between-school variability (the differences in school means from the grand mean). The red diamonds show that schools differ in their average exam scores, suggesting that school-level factors may influence student performance.

Covariate effects

Before fitting models with covariates, it is useful to explore the relationships between potential predictors and the outcome variable.

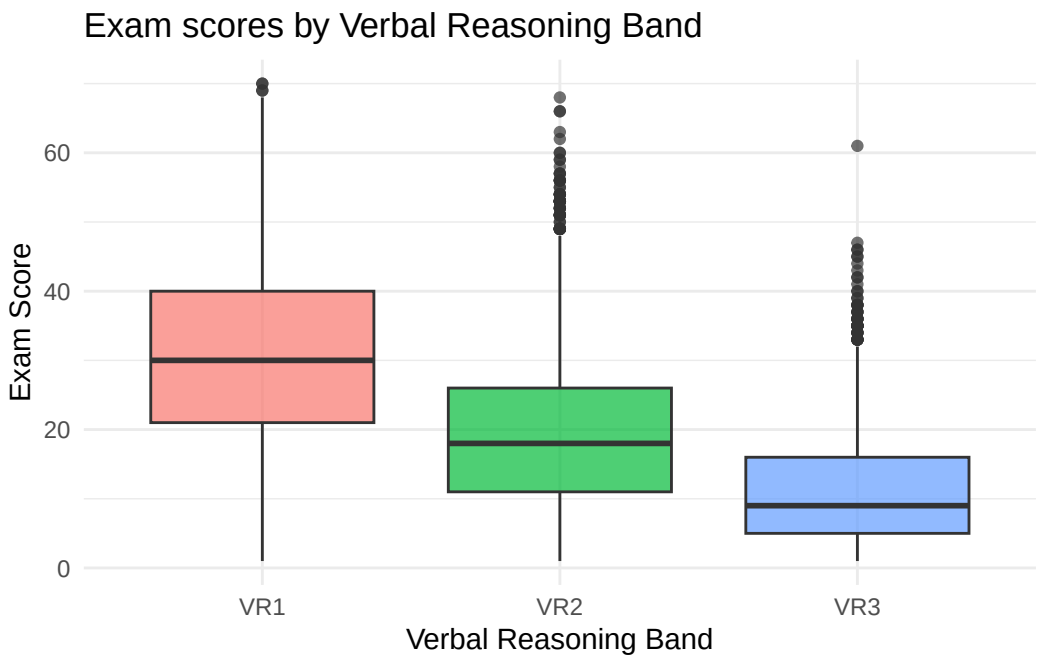
First, let's examine how exam scores vary across different verbal reasoning bands:

```
# Boxplot for VRBand vs ExamScore
ggplot(ilea_data, aes(x = VRBand, y = ExamScore, fill = VRBand)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Exam scores by Verbal Reasoning Band",
       x = "Verbal Reasoning Band",
       y = "Exam Score") +
  theme_minimal() +
  theme(legend.position = "none")
```



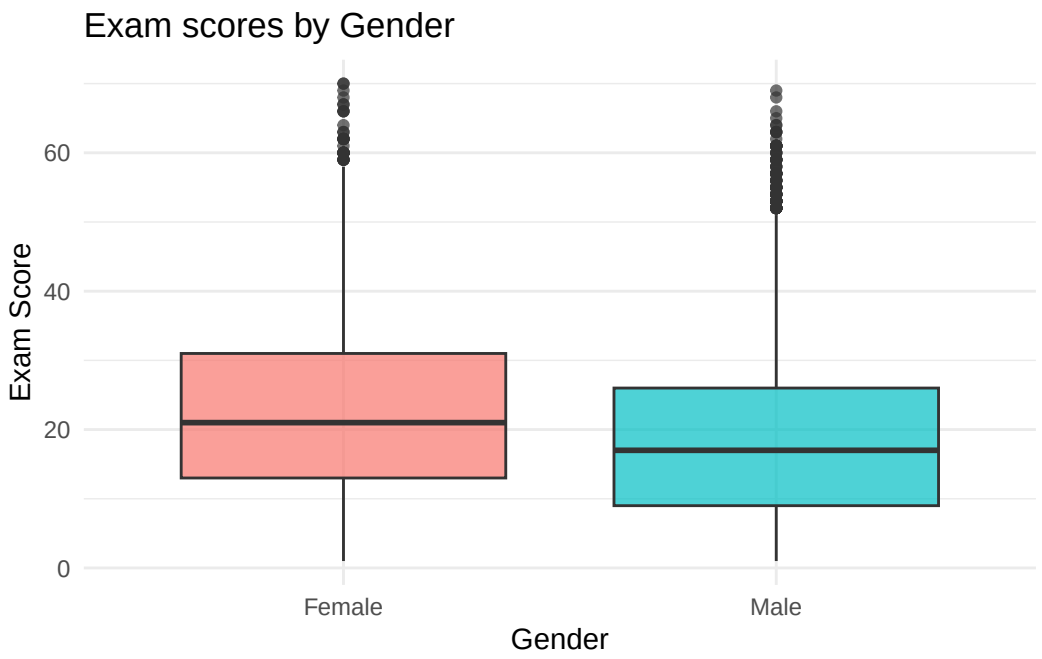
Note that the plot above includes an `NA` category for students with missing VRBand values. This is default `ggplot2` behavior. To exclude missing values from the plot, we can filter them out before plotting:

```
# Boxplot for VRBand vs ExamScore (excluding missing values)
ilea_data %>%
  filter(!is.na(VRBand)) %>%
  ggplot(aes(x = VRBand, y = ExamScore, fill = VRBand)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Exam scores by Verbal Reasoning Band",
       x = "Verbal Reasoning Band",
       y = "Exam Score") +
  theme_minimal() +
  theme(legend.position = "none")
```



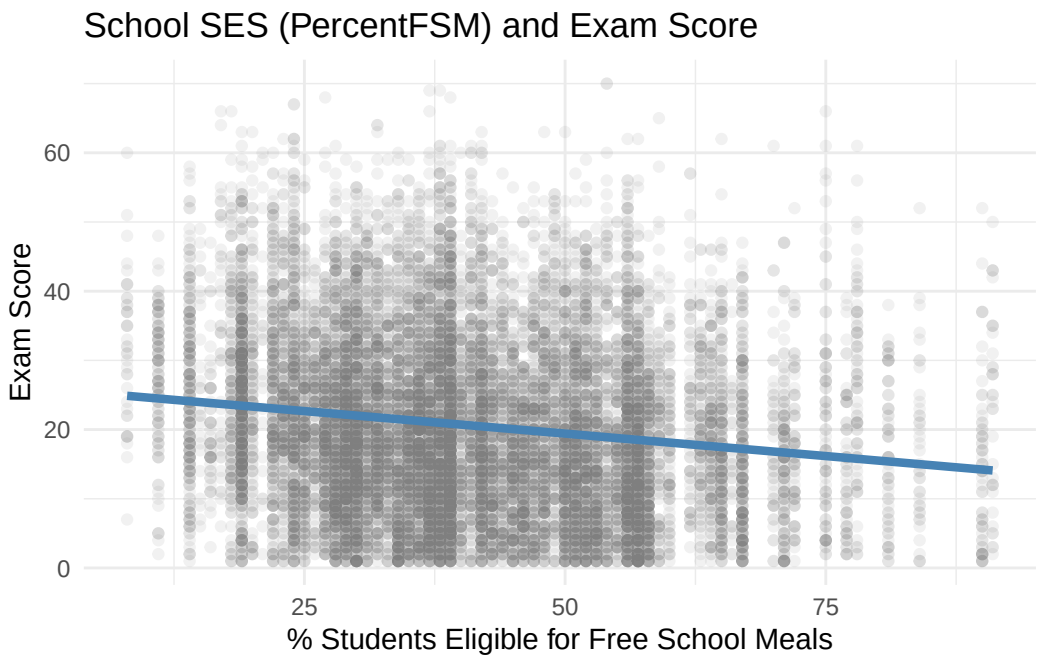
Next, let's examine how exam scores differ by gender:

```
# Boxplot for Gender vs ExamScore
ggplot(ilea_data, aes(x = Gender, y = ExamScore, fill = Gender)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Exam scores by Gender",
       x = "Gender",
       y = "Exam Score") +
  theme_minimal() +
  theme(legend.position = "none")
```



Now let's explore the relationship between PercentFSM (school-level socioeconomic indicator) and exam scores

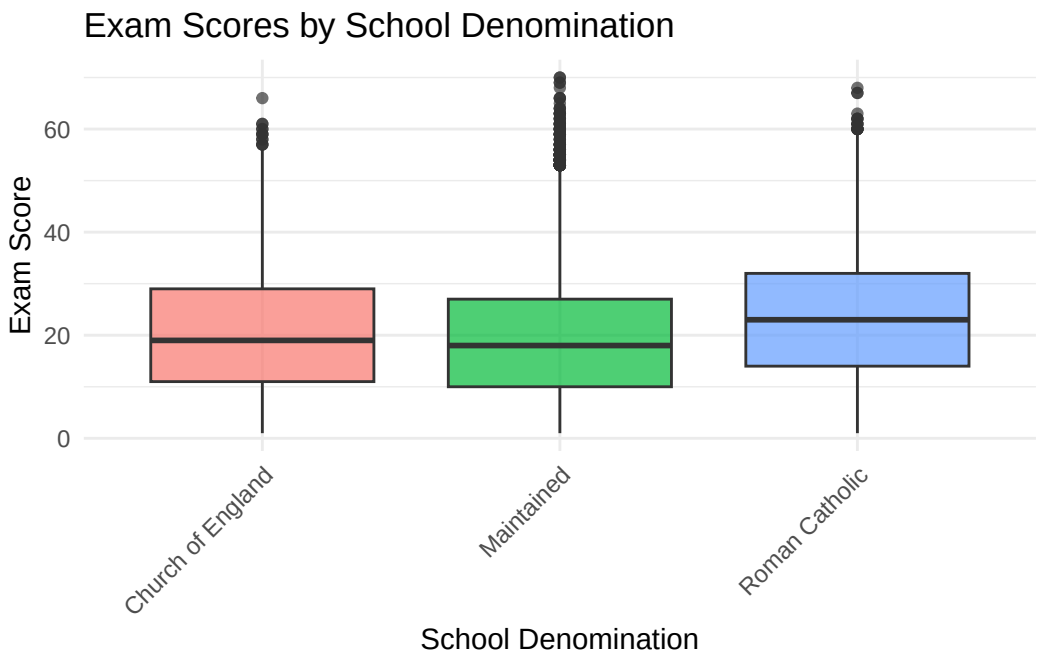
```
# Scatterplot of individual exam scores vs PercentFSM
ggplot(ilea_data, aes(x = PercentFSM, y = ExamScore)) +
  geom_point(alpha = 0.1, color = "gray50") +
  geom_smooth(method = "lm", se = FALSE,
              color = "steelblue", linewidth = 1.5) +
  labs(title = "School SES (PercentFSM) and Exam Score",
       x = "% Students Eligible for Free School Meals",
       y = "Exam Score") +
  theme_minimal()
```



The plot shows a weak negative relationship between school SES and individual exam scores. The large vertical spread of points at each %FSM value indicates that most of the variability in exam scores is at the individual level, not the school level. This is consistent with the variance components plot.

Finally, let's examine exam scores by school denomination:

```
# Boxplot for SchoolDenomination vs ExamScore (individual level)
ggplot(ilea_data,
       aes(x = SchoolDenomination, y = ExamScore,
           fill = SchoolDenomination)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Exam Scores by School Denomination",
       x = "School Denomination",
       y = "Exam Score") +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1))
```



Random intercept model (null model)

To partition the variance in exam scores between students and schools, we start by fitting a null model (random intercept model without any predictors):

```
# Fit the null model (random intercept only)
null_model <- lmer(ExamScore ~ 1 + (1 | School), data = ilea_data)
summary(null_model)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: ExamScore ~ 1 + (1 | School)
   Data: ilea_data

REML criterion at convergence: 120364.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4962 -0.7550 -0.1073  0.6290  4.0683

Random effects:
 Groups   Name        Variance Std.Dev.
```

```
 School   (Intercept)  19.52    4.419
 Residual             144.53   12.022
Number of obs: 15362, groups:  School, 139

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  20.4527     0.3913 135.8938   52.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the variance components, we can calculate the Intraclass Correlation Coefficient (ICC):

```r
# Extract variance components
var_between <- as.data.frame(VarCorr(null_model))$vcov[1]
var_within <- as.data.frame(VarCorr(null_model))$vcov[2]

# Calculate ICC
ICC <- var_between / (var_between + var_within)
ICC
```

```
[1] 0.1190031
```

The ICC indicates that approximately 11.9% of the total variability in exam scores can be attributed to differences between schools, while the remaining 88.1% is due to differences between students within schools. This suggests that while most of the variability in exam scores is at the student level, there is still meaningful variation between schools that warrants a multilevel approach.

Extending the random intercept model with individual-level variables

As a second step, we extend the random intercept model with individual-level variables (`Gender` and `VRBand`). To facilitate the interpretation of the model coefficients, we use effects coding for categorical variables:

```r
# Use effects coding for the categorical variables
options(contrasts = c("contr.sum", "contr.poly"))

# Fit the random intercept model with individual-level variables
model_L1 <- lmer(ExamScore ~ Gender + VRBand + (1 | School), data = ilea_data)
summary(model_L1)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: ExamScore ~ Gender + VRBand + (1 | School)
   Data: ilea_data

REML criterion at convergence: 115230.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2419 -0.6751 -0.0707  0.6204  4.6713

Random effects:
 Groups   Name        Variance Std.Dev.
 School   (Intercept)  12.1     3.479
 Residual             104.3    10.214
Number of obs: 15347, groups:  School, 139

Fixed effects:
             Estimate Std. Error         df t value Pr(>|t|)
(Intercept)    20.1630     0.3123   138.0603  64.572   <2e-16 ***
Gender1         1.4029     0.1166  9890.3703  12.027   <2e-16 ***
VRBand1         9.7789     0.1309 15279.0725  74.689   <2e-16 ***
VRBand2        -1.0408     0.1116 15233.0240  -9.326   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
        (Intr) Gendr1 VRBnd1
Gender1 -0.014
VRBand1  0.013  0.003
VRBand2 -0.114 -0.022 -0.316
```

```
# Display the coding scheme for the categorical variables
cat("\nContrasts for Gender:\n")
```

```
Contrasts for Gender:
```

```
contrasts(ilea_data$Gender)
```

```
       [,1]
```

```
Female    1
Male     -1
```

```
cat("\nContrasts for VRBand:\n")
```

```
Contrasts for VRBand:
```

```
contrasts(ilea_data$VRBand)
```

```
    [,1] [,2]
VR1    1    0
VR2    0    1
VR3   -1   -1
```

```
# Obtain the ANOVA table for the individual-level variables
anova(model_L1)
```

```
Type III Analysis of Variance Table with Satterthwaite's method
       Sum Sq Mean Sq NumDF   DenDF F value    Pr(>F)
Gender  15090   15090     1  9890.4  144.64 < 2.2e-16 ***
VRBand 605573  302786     2 15259.3 2902.24 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show that both `Gender` and `VRBand` are significant predictors of exam scores. With effects coding:

- The intercept represents the grand mean exam score across all groups
- The coefficient for Gender1 (Female = 1) indicates how much the mean for females differs from the grand mean; males differ by the same amount in the opposite direction
- The coefficients for VRBand indicate how much each verbal reasoning band differs from the grand mean

Including context-level variables

As a third step, we extend the model with context-level variables (`PercentFSM` and `SchoolDenomination`). We start by centering the continuous context-level variable:

```r
# Center the context-level variable
ilea_data <- ilea_data %>%
  mutate(PercentFSM_c = scale(PercentFSM, scale = FALSE))

# Fit the random intercept model with individual-level and context-level variables
model_L1L2 <- lmer(ExamScore ~ Gender + VRBand + PercentFSM_c +
                     SchoolDenomination + (1 | School),
                   data = ilea_data)
summary(model_L1L2)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: ExamScore ~ Gender + VRBand + PercentFSM_c + SchoolDenomination +
    (1 | School)
   Data: ilea_data

REML criterion at convergence: 115220.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2472 -0.6746 -0.0683  0.6213  4.6762

Random effects:
 Groups   Name        Variance Std.Dev.
 School   (Intercept)  10.83    3.291
 Residual             104.33   10.214
Number of obs: 15347, groups:  School, 139

Fixed effects:
                     Estimate Std. Error         df t value Pr(>|t|)
(Intercept)          2.070e+01  3.676e-01  1.252e+02  56.293   <2e-16 ***
Gender1              1.398e+00  1.163e-01  8.937e+03  12.026   <2e-16 ***
VRBand1              9.772e+00  1.310e-01  1.526e+04  74.607   <2e-16 ***
VRBand2             -1.042e+00  1.116e-01  1.523e+04  -9.334   <2e-16 ***
PercentFSM_c         1.131e-03  1.082e-02  1.753e+03   0.105   0.9167
SchoolDenomination1 -1.215e-01  6.076e-01  1.215e+02  -0.200   0.8418
SchoolDenomination2 -1.389e+00  4.279e-01  1.260e+02  -3.246   0.0015 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) Gendr1 VRBnd1 VRBnd2 PrFSM_ SchlD1
```

```
Gender1     -0.031
VRBand1      0.005  0.004
VRBand2     -0.096 -0.023 -0.316
PercntFSM_c  0.042  0.022  0.024 -0.001
SchlDnmntn1  0.456 -0.022  0.002  0.002  0.050
SchlDnmntn2 -0.565  0.033  0.013  0.000 -0.142 -0.573
```

```r
# Obtain the ANOVA table
anova(model_L1L2)
```

```
Type III Analysis of Variance Table with Satterthwaite's method
                   Sum Sq Mean Sq NumDF   DenDF   F value    Pr(>F)
Gender              15088   15088     1  8936.5  144.6163 < 2.2e-16 ***
VRBand             604156  302078     2 15244.0 2895.3715 < 2.2e-16 ***
PercentFSM_c            1       1     1  1752.6    0.0109 0.9167420
SchoolDenomination   1759     879     2   124.7    8.4288 0.0003686 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To obtain a more parsimonious model, we can use the `step()` function from the `lmerTest` package to perform backward elimination of fixed-effect terms:

```r
# Perform stepwise selection to identify the most important predictors
# Set reduce.random = FALSE to only eliminate fixed effects (not random effects)
step(model_L1L2, reduce.random = FALSE)
```

```
Backward reduced random-effect table:

             Eliminated npar logLik    AIC    LRT Df Pr(>Chisq)
<none>                    9 -57610 115239
(1 | School)          0   8 -58050 116116 879.47  1  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Backward reduced fixed-effect table:
Degrees of freedom method: Satterthwaite

                   Eliminated Sum Sq Mean Sq NumDF   DenDF   F value    Pr(>F)
PercentFSM_c                1      1       1     1  1752.6    0.0109 0.9167420
Gender                      0  15105   15105     1  9141.4  144.7771 < 2.2e-
16
```

```
VRBand                      0 604523  302261     2 15260.5 2897.1818 < 2.2e-
16
SchoolDenomination          0   1789     895     2   132.0    8.5744 0.0003154

PercentFSM_c
Gender              ***
VRBand              ***
SchoolDenomination ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model found:
ExamScore ~ Gender + VRBand + SchoolDenomination + (1 | School)
```

Understanding the non-significant PercentFSM effect

An interesting observation from the model output is that `PercentFSM_c` is **not statistically significant** ($p = 0.92$). This might seem surprising given that school SES is often considered an important predictor of educational outcomes. However, this result illustrates an important principle in multi-level modeling: the effect of a context-level variable can be substantially reduced (or disappear entirely) when individual-level confounders are controlled for.

> **!** Compositional Effects in Multilevel Data
>
> In our multi-level model, `PercentFSM_c` is estimated while controlling for individual-level predictors like `Gender` and `VRBand`. The key insight is that `VRBand` is doing most of the explanatory work. Students in higher verbal reasoning bands score much higher (VRBand1 effect +10 points), and these high-ability students are likely concentrated in schools with lower %FSM. Once we adjust for individual ability, the contextual effect of school SES largely disappears.

To demonstrate this, let's fit a model **without** the `VRBand` predictor:

```r
# Model without VRBand - to demonstrate the effect
model_no_vr <- lmer(ExamScore ~ Gender + PercentFSM_c +
                      SchoolDenomination + (1 | School),
                    data = ilea_data)
summary(model_no_vr)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: ExamScore ~ Gender + PercentFSM_c + SchoolDenomination + (1 |
```

```
    School)
   Data: ilea_data

REML criterion at convergence: 120239.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5121 -0.7623 -0.1188  0.6257  4.0733

Random effects:
 Groups   Name        Variance Std.Dev.
 School   (Intercept)  12.96    3.60
 Residual             143.79   11.99
Number of obs: 15362, groups:  School, 139

Fixed effects:
                     Estimate Std. Error         df t value Pr(>|t|)
(Intercept)          21.09779    0.40317  117.91642  52.330  < 2e-16 ***
Gender1               1.41165    0.13559 7841.25505  10.411  < 2e-16 ***
PercentFSM_c         -0.02566    0.01246 1425.93849  -2.060 0.039580 *
SchoolDenomination1  -0.26043    0.66942  116.58109  -0.389 0.697954
SchoolDenomination2  -1.77459    0.47191  121.12771  -3.760 0.000263 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) Gendr1 PrFSM_ SchlD1
Gender1     -0.035
PercntFSM_c  0.044  0.024
SchlDnmntn1  0.458 -0.023  0.051
SchlDnmntn2 -0.568  0.034 -0.148 -0.572
```

Without adjusting for individual verbal reasoning ability, `PercentFSM_c` shows a significant negative effect (= -0.026, p = 0.04): for every 1 percentage point increase in the proportion of students eligible for free school meals, exam scores decrease by approximately 0.026 points.

Exploring cross-level interactions

Finally, we explore potential cross-level interactions. These are interactions between individual-level variables (`Gender`, `VRBand`) and context-level variables (`SchoolDenomination`). We fit a full model that includes both cross-level interactions, then use backward elimination to identify significant terms.

```r
# Fit full model with cross-level interactions
# (excluding PercentFSM which was not significant)
model_full <- lmer(ExamScore ~ Gender * SchoolDenomination +
                     VRBand * SchoolDenomination + (1 | School),
                   data = ilea_data)

# Use step() to perform backward elimination
step(model_full, reduce.random = FALSE)
```

```
Backward reduced random-effect table:

             Eliminated npar logLik    AIC    LRT Df Pr(>Chisq)
<none>                    14 -57601 115231
(1 | School)          0   13 -58072 116169 940.29  1  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Backward reduced fixed-effect table:
Degrees of freedom method: Satterthwaite

                          Eliminated  Sum Sq Mean Sq NumDF   DenDF  F value
Gender:SchoolDenomination          1   388.8   194.4     2  7785.9   1.8648
Gender                             0 14876.8 14876.8     1  9138.8 142.7095
SchoolDenomination:VRBand          0  1701.0   425.2     4 15248.1   4.0793
                             Pr(>F)
Gender:SchoolDenomination  0.154997
Gender                    < 2.2e-16 ***
SchoolDenomination:VRBand  0.002631 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model found:
ExamScore ~ Gender + SchoolDenomination + VRBand + (1 | School) + SchoolDenomination:VRBand
```

The stepwise selection eliminates the `Gender × SchoolDenomination` interaction (not significant) but retains the `VRBand × SchoolDenomination` interaction. This indicates that the effect of verbal reasoning band on exam scores differs across school denominations. Let's fit and examine the final model:

```r
# Fit the final model with the significant interaction
model_final <- lmer(ExamScore ~ Gender + VRBand * SchoolDenomination +
```

```
                      (1 | School),
                    data = ilea_data)
summary(model_final)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: ExamScore ~ Gender + VRBand * SchoolDenomination + (1 | School)
   Data: ilea_data

REML criterion at convergence: 115203.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2397 -0.6802 -0.0696  0.6248  4.6536

Random effects:
 Groups   Name        Variance Std.Dev.
 School   (Intercept)  10.76    3.28
 Residual             104.25   10.21
Number of obs: 15347, groups:  School, 139

Fixed effects:
                             Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                   20.6431     0.3675   136.1497  56.173  < 2e-16
Gender1                        1.3883     0.1162  9138.8255  11.946  < 2e-16
VRBand1                       10.0433     0.1591 15261.4816  63.140  < 2e-16
VRBand2                       -0.9569     0.1385 15224.1388  -6.907 5.14e-12
SchoolDenomination1           -0.2546     0.6096   134.3471  -0.418  0.67688
SchoolDenomination2           -1.3076     0.4257   135.9655  -3.072  0.00257
VRBand1:SchoolDenomination1    0.8449     0.2591 15252.6411   3.262  0.00111
VRBand2:SchoolDenomination1    0.1962     0.2243 15218.1053   0.875  0.38165
VRBand1:SchoolDenomination2   -0.4698     0.1853 15265.1202  -2.535  0.01126
VRBand2:SchoolDenomination2   -0.1436     0.1597 15227.6114  -0.899  0.36873

(Intercept)                  ***
Gender1                      ***
VRBand1                      ***
VRBand2                      ***
SchoolDenomination1
SchoolDenomination2          **
VRBand1:SchoolDenomination1 **
VRBand2:SchoolDenomination1
```

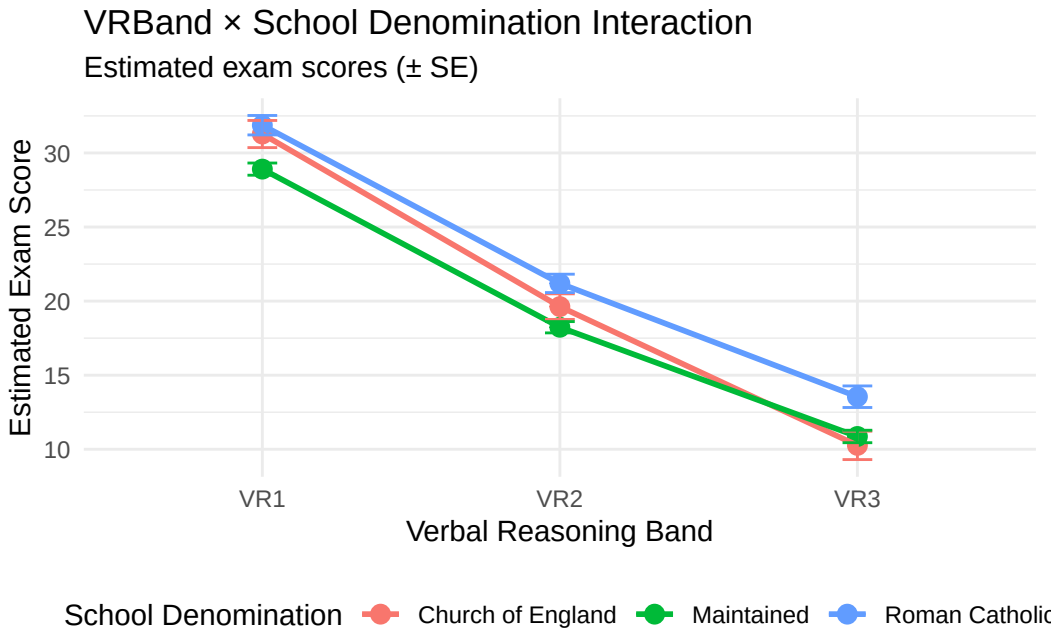
```
VRBand1:SchoolDenomination2 *
VRBand2:SchoolDenomination2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) Gendr1 VRBnd1 VRBnd2 SchlD1 SchlD2 VRB1:SD1 VRB2:SD1
Gender1     -0.031
VRBand1     -0.004 -0.010
VRBand2     -0.122 -0.021 -0.275
SchlDnmntn1  0.454 -0.022  0.008 -0.041
SchlDnmntn2 -0.569  0.035  0.013  0.072 -0.570
VRBnd1:ScD1  0.009 -0.013  0.401 -0.140  0.002 -0.014
VRBnd2:ScD1 -0.042 -0.006 -0.142  0.384 -0.117  0.057 -0.296
VRBnd1:ScD2  0.013  0.023 -0.552  0.132 -0.013  0.006 -0.532    0.186
VRBnd2:ScD2  0.073  0.002  0.134 -0.581  0.056 -0.122  0.187   -0.510
            VRB1:SD2
Gender1
VRBand1
VRBand2
SchlDnmntn1
SchlDnmntn2
VRBnd1:ScD1
VRBnd2:ScD1
VRBnd1:ScD2
VRBnd2:ScD2 -0.296
```

Visualizing the VRBand × SchoolDenomination interaction

We use the `emmeans` package to visualize how the effect of verbal reasoning band varies across school denominations:

```r
# Compute estimated marginal means for VRBand × SchoolDenomination
emm_vr_denom <- emmeans(model_final,
                        specs = ~ VRBand | SchoolDenomination,
                        lmer.df = "satterthwaite")

# Convert to data frame for plotting
pred_vr_denom <- as.data.frame(summary(emm_vr_denom))

# Create interaction plot for VRBand × SchoolDenomination
ggplot(pred_vr_denom,
```

```
      aes(x = VRBand, y = emmean,
          color = SchoolDenomination,
          group = SchoolDenomination)) +
  geom_point(size = 3) +
  geom_line(linewidth = 1) +
  geom_errorbar(aes(ymin = emmean - SE, ymax = emmean + SE),
                width = 0.1) +
  labs(title = "VRBand × School Denomination Interaction",
       subtitle = "Estimated exam scores (± SE)",
       x = "Verbal Reasoning Band",
       y = "Estimated Exam Score",
       color = "School Denomination") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



The interaction plot shows how the ability-achievement relationship differs across school types. Non-parallel lines indicate that the gap between high-ability (VR1) and low-ability (VR3) students varies depending on the school denomination.

The plot reveals several key findings:

- **Roman Catholic** and **Maintained** schools show relatively parallel lines, indicating a similar ability-achievement relationship in these school types

- **Church of England schools** show a steeper slope, indicating a stronger effect of verbal reasoning band on exam scores
- The **VR1-VR3 gap is largest in Church of England schools** (approximately 21 points: from ~31 at VR1 to ~10 at VR3)
- In contrast, **Roman Catholic schools show a smaller ability gap** (approximately 18 points: from ~32 at VR1 to ~14 at VR3)

This pattern suggests that Church of England schools show greater differentiation in exam performance between high- and low-ability students, while Roman Catholic schools show more similar outcomes across ability levels (with generally higher scores for lower-ability students compared to other denominations).

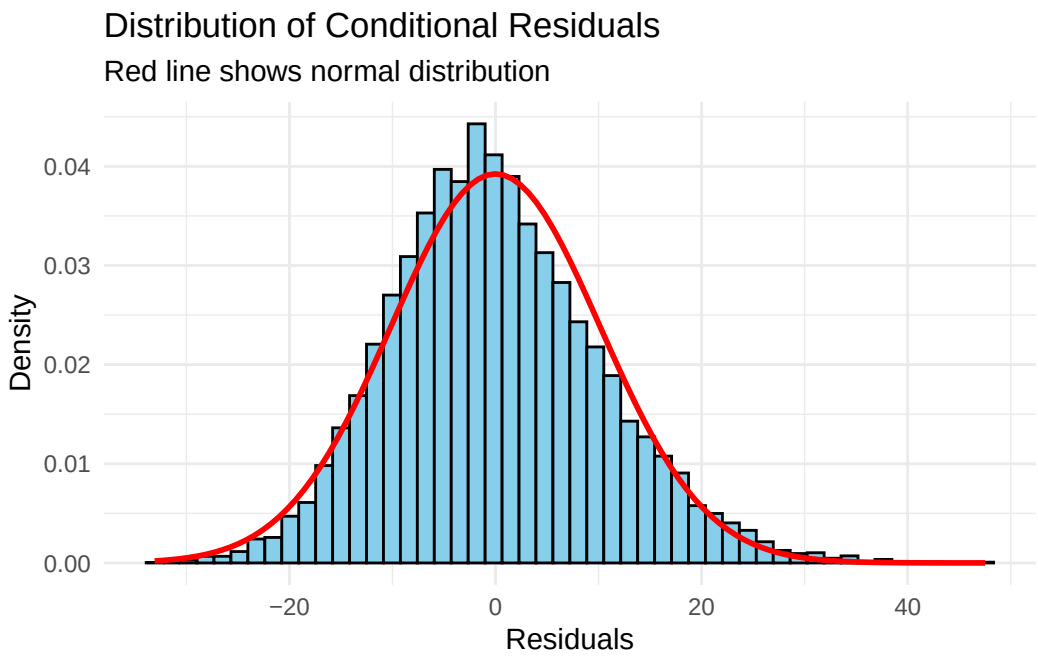
Model diagnostics

To assess the assumptions of the multilevel model, we examine the conditional residuals (the difference between observed values and fitted values). Key assumptions include:

1. Residuals are normally distributed
2. Residuals have constant variance (homoscedasticity)
3. Residuals are independent

Histogram of conditional residuals

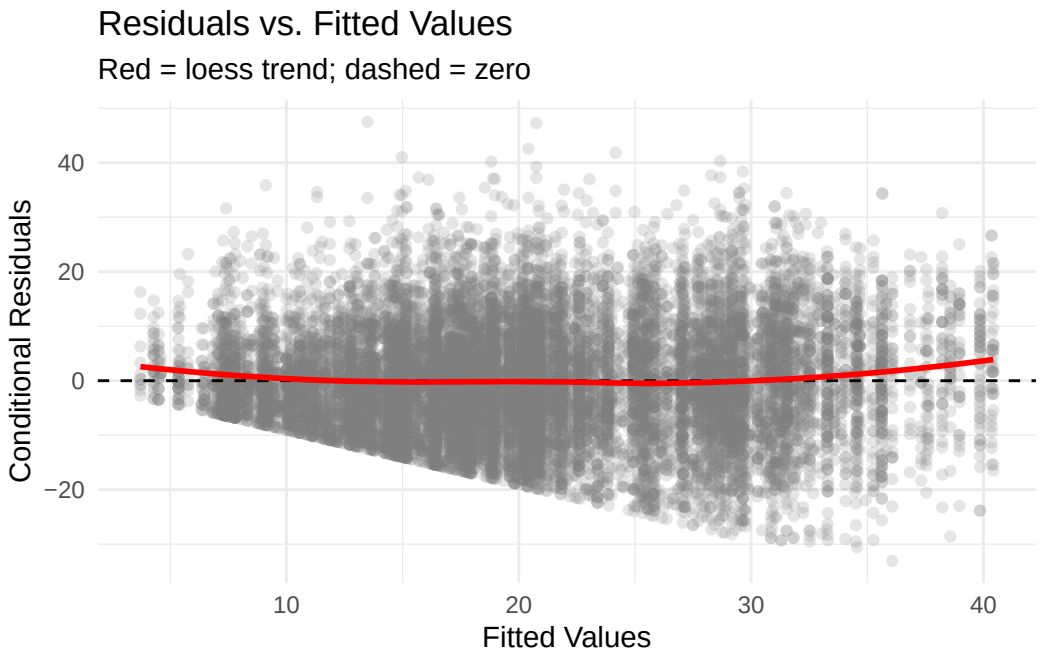
```
# Extract conditional residuals from the final model
residuals_cond <- residuals(model_final)

# Create histogram of conditional residuals
ggplot(data.frame(residuals = residuals_cond), aes(x = residuals)) +
  geom_histogram(bins = 50, fill = "skyblue", color = "black",
                 aes(y = after_stat(density))) +
  stat_function(fun = dnorm,
                args = list(mean = mean(residuals_cond),
                            sd = sd(residuals_cond)),
                color = "red", linewidth = 1) +
  labs(title = "Distribution of Conditional Residuals",
       subtitle = "Red line shows normal distribution",
       x = "Residuals",
       y = "Density") +
  theme_minimal()
```



The histogram shows that the conditional residuals are approximately normally distributed, with a shape that closely follows the theoretical normal distribution (red line).

Residuals vs. fitted values

```
# Extract fitted values
fitted_values <- fitted(model_final)

# Create residuals vs fitted values plot
ggplot(data.frame(fitted = fitted_values,
                  residuals = residuals_cond),
       aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.2, color = "gray50") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") +
  geom_smooth(method = "loess", se = FALSE,
              color = "red", linewidth = 1) +
  labs(title = "Residuals vs. Fitted Values",
       subtitle = "Red = loess trend; dashed = zero",
       x = "Fitted Values",
       y = "Conditional Residuals") +
  theme_minimal()
```



The residuals vs. fitted values plot reveals some violation of the homoscedasticity assumption. The loess trend line curves upward at both the low and high ends of the fitted values, and the spread of residuals appears more constrained at lower fitted values. This pattern is likely caused by a **floor effect** in the exam scores: as seen in the histogram of the outcome variable, a substantial number of students scored zero or near-zero on the exam. Since students cannot score below zero, residuals for students with low predicted scores are truncated on the negative side, leading to:

1. A "fan-shaped" pattern where residual variance increases with fitted values
2. A curved loess line that deviates from the horizontal zero line

This floor effect, combined with the right-skewed distribution of exam scores, suggests that the assumption of homogeneous residual variance is not fully met. In practice, mixed models are reasonably robust to moderate violations of this assumption, especially with large sample sizes. However, for a more rigorous analysis, one could consider:

- Using a generalized linear mixed model with a distribution that better accommodates bounded outcomes
- Applying robust standard errors
- Using bootstrapping to estimate standard errors