Linear Mixed Models

Advanced Numerical Data Analysis

Dr Muhammad Saufi

2024-06-08

Contents

1 Loading Packages **2**

2 Dataset **4**
2.1 Load the Dataset . . . 5
2.2 Data Wrangling . . . 5
2.3 Exploratory Data Analysis . . . 6
2.3.1 Descriptive Table . . . 6
2.3.2 Plot 1 . . . 7
2.3.3 Plot 2 . . . 8

3 Multilevel Model **9**
3.1 Null Model (Simplest Model) . . . 9
3.2 Single-Level Analysis . . . 10
3.3 Multilevel Analysis . . . 12
3.3.1 Random Intercept Models . . . 14
3.3.1.1 Add an Explanatory Variable . . . 14
3.3.1.2 Prediction . . . 17
3.3.1.2.1 Steps for Prediction . . . 18
3.3.1.2.2 Explanation . . . 19
3.3.1.2.3 Confirmation with Manual Calculation . . . 20
3.3.1.3 Plot . . . 21
3.3.1.4 Variance . . . 21
3.3.1.4.1 Between School Variance . . . 22
3.3.1.4.2 Within School Variance . . . 22
3.3.2 Random Slope Models . . . 23
3.3.2.1 The Fitted Values . . . 27
3.3.2.1.1 Interpretation of the Fitted Values . . . 27

3.3.2.2 Model Comparison: Random Intercept vs. Random Slope . . . 28
3.3.2.3 Interpretation of Random Effects Across Schools 29
3.3.2.4 Prediction from Random Slope 30
3.3.2.5 Plot of Random Effects 31
3.3.2.5.1 Scatter Plot 1 32
3.3.2.5.2 Scatter Plot 2 32
3.3.2.6 Equation for Random Slope Model 33
3.3.2.7 Plot the Fitted Values from Random Slope 34
3.3.2.8 Adding a Level-1 Variable to the Random Slope Model 35
3.3.2.8.1 Random Slope Model with Gender 36
3.3.2.8.2 Random Slope Model with Gender Having a Random Slope 39
3.3.2.8.3 Comparison Between Models 41
3.3.2.9 Adding a Level-2 Explanatory Variable to the Random Slope Model 42
3.3.2.10 Cross-Level Interaction in the Random Slope Model 45
3.3.2.11 Checking Assumptions 48

4 Acknowledgements **53**

5 References **53**

6 R Codes **53**

1 Loading Packages

```
library(haven)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(broom.mixed)
library(here)
```

```
here() starts at D:/OneDrive/My Education/R Learning Hub/DrPH Epidemiology Revision
```

```
library(gtsummary)
library(DT)
library(kableExtra)
```

```
Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```
library(lme4)
```

```
Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```
library(lmerTest)
```

```
Attaching package: 'lmerTest'

The following object is masked from 'package:lme4':

    lmer

The following object is masked from 'package:stats':

    step
```

```r
library(dplyr)
library(DT)
library(merTools)
```

```
Loading required package: arm
Loading required package: MASS

Attaching package: 'MASS'

The following object is masked from 'package:gtsummary':

    select

The following object is masked from 'package:dplyr':

    select


arm (Version 1.14-4, built: 2024-4-1)

Working directory is D:/OneDrive/My Education/R Learning Hub/DrPH Epidemiology Revision
```

```r
library(lattice)
```

2 Dataset

The dataset is from the Scottish School Leavers Survey (SSLS). It is a longitudinal dataset that captures information on several cohorts of young people over time.

Hierarchy: The dataset has a hierarchical structure with students (level 1) nested within schools (level 2).

- **Level 1 (Students)**: Individual students identified by `caseid`.
- **Level 2 (Schools)**: Schools identified by `schoolid`.

Dependent Variable:

- `score`: Total attainment score of the student, ranging from 0 to 75.

Explanatory Variables:

- `cohort90`: Year of the cohort, represented by subtracting 1990 from each value. Values range from -6 (1984) to 8 (1998), with 0 representing 1990.
- `female`: Gender of the student (1 = female, 0 = male).
- `sclass`: Social class of the student, defined as the higher class of mother or father (1 = managerial and professional, 2 = intermediate, 3 = working, 4 = unclassified).
- `schtype`: Type of school (1 = independent, 0 = state-funded).
- `schurban`: Urban-rural classification of the school (1 = urban, 0 = town or rural).
- `schdenom`: School denomination (1 = Roman Catholic, 0 = non-denominational).

2.1 Load the Dataset

```
score.sch <- read_dta("score.sch_data.dta")
glimpse(score.sch)
```

```
Rows: 33,988
Columns: 9
$ caseid   <dbl> 18, 17, 19, 20, 21, 13, 16, 14, 15, 12, 12865, 6509, 12866, 1~
$ schoolid <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ score    <dbl> 0, 10, 0, 40, 42, 4, 0, 0, 14, 27, 18, 23, 24, 0, 25, 4, 11, ~
$ cohort90 <dbl> -6, -6, -6, -6, -6, -6, -6, -6, -6, -6, -2, -4, -2, -2, -4, -~
$ female   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0~
$ sclass   <dbl> 2, 2, 4, 3, 2, 2, 3, 4, 3, 2, 2, 1, 2, 3, 2, 3, 2, 4, 3, 3, 3~
$ schtype  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ schurban <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ schdenom <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

2.2 Data Wrangling

Numerical variables are often converted to factors to facilitate easier analysis and interpretation.

```
score.sch <- score.sch %>%
  mutate(
    female2 = factor(female, labels = c('male', 'female')),
    class2 = factor(sclass, labels = c('managerial+prof', 'intermediate', 'working', 'unclass
    schtype2 = factor(schtype, labels = c('state-funded', 'independent')),
    schurban2 = factor(schurban, labels = c('town/rural', 'urban')),
    schdenom2 = factor(schdenom, labels = c('state-funded', 'independent')))
glimpse(score.sch)
```

```
Rows: 33,988
Columns: 14
$ caseid    <dbl> 18, 17, 19, 20, 21, 13, 16, 14, 15, 12, 12865, 6509, 12866, ~
$ schoolid  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ score     <dbl> 0, 10, 0, 40, 42, 4, 0, 0, 14, 27, 18, 23, 24, 0, 25, 4, 11,~
$ cohort90  <dbl> -6, -6, -6, -6, -6, -6, -6, -6, -6, -6, -2, -4, -2, -2, -4, ~
$ female    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ sclass    <dbl> 2, 2, 4, 3, 2, 2, 3, 4, 3, 2, 2, 1, 2, 3, 2, 3, 2, 4, 3, 3, ~
$ schtype   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ schurban  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ schdenom  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ female2   <fct> female, female, female, female, female, female, female, fema~
$ class2    <fct> intermediate, intermediate, unclassified, working, intermedi~
$ schtype2  <fct> state-funded, state-funded, state-funded, state-funded, stat~
$ schurban2 <fct> urban, urban, urban, urban, urban, urban, urban, urban, urba~
$ schdenom2 <fct> state-funded, state-funded, state-funded, state-funded, stat~
```

2.3 Exploratory Data Analysis

Summarize the data and create some basic plots.

2.3.1 Descriptive Table

```
score.sch %>% tbl_summary()
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

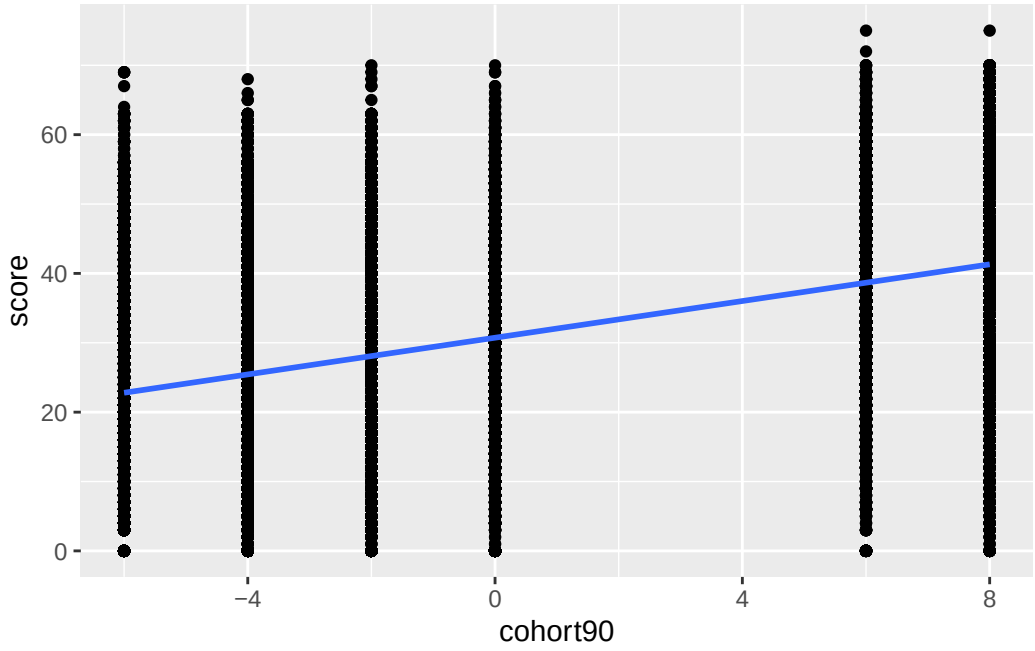
Characteristic	**N = 33,988**
Case ID	17,319 (8,532, 29,428)
School ID	256 (123, 386)
Score	33 (19, 45)
Cohort	
-6	6,478 (19%)
-4	6,325 (19%)
-2	5,245 (15%)
0	4,371 (13%)

Characteristic	N = 33,988
6	4,244 (12%)
8	7,325 (22%)
Female	17,933 (53%)
Social class	
1	11,173 (33%)
2	9,994 (29%)
3	9,486 (28%)
4	3,335 (9.8%)
School type	1,540 (4.5%)
School urban-rural classification	24,116 (71%)
School denomination	5,358 (16%)
female2	
male	16,055 (47%)
female	17,933 (53%)
class2	
managerial+prof	11,173 (33%)
intermediate	9,994 (29%)
working	9,486 (28%)
unclassified	3,335 (9.8%)
schtype2	
state-funded	32,448 (95%)
independent	1,540 (4.5%)
schurban2	
town/rural	9,872 (29%)
urban	24,116 (71%)
schdenom2	
state-funded	28,630 (84%)
independent	5,358 (16%)

2.3.2 Plot 1

```
score.sch %>%
  ggplot(aes(x = cohort90, y = score)) +
  geom_point() +
  geom_smooth(method = lm)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

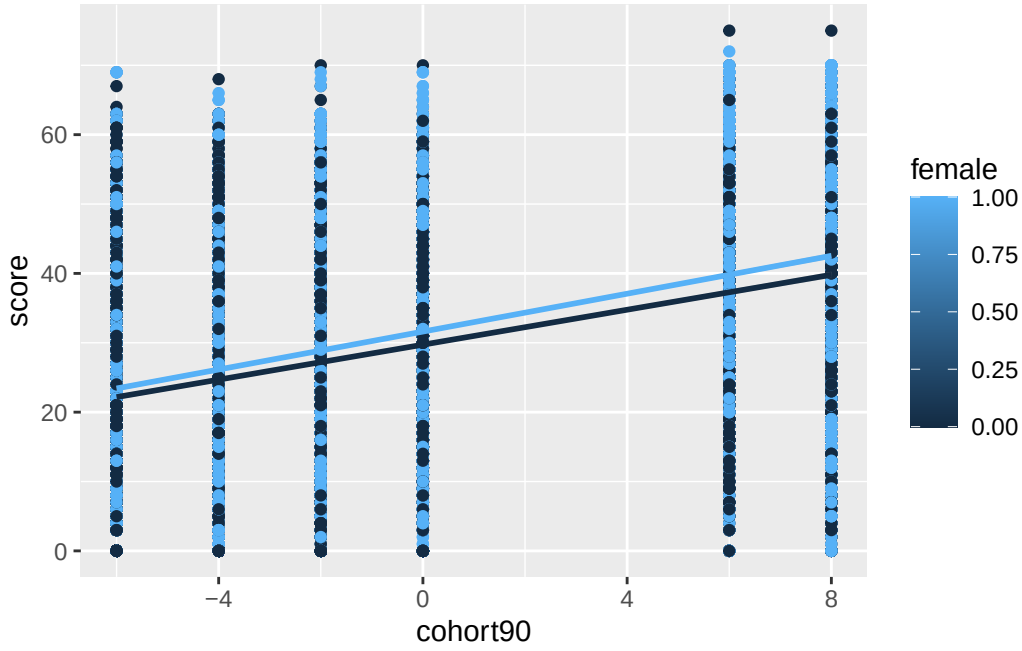


Comment: This graph displays the relationship between `cohort90` and `score`, along with a fitted linear regression line. The graph suggests that student attainment scores have generally increased over the years from 1984 to 1998. However, there is a substantial amount of variability in the scores within each cohort year. The positive slope of the regression line quantifies the trend of improvement in scores over time.

2.3.3 Plot 2

```
score.sch %>%
  ggplot(aes(x = cohort90, y = score, col = female, group = female)) +
  geom_point() +
  geom_smooth(method = lm)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Comment: This graph shows the relationship between `cohort90` and `score`, with a distinction between male and female students. The graph indicates that both male and female students have experienced improvements in attainment scores over the years from 1984 to 1998. The trends for both genders are very similar, suggesting that gender does not play a significant role in the difference in attainment scores within this dataset. The upward slope of both lines quantifies the overall improvement in scores over time, and the color differentiation helps visualize the gender distribution within each cohort.

3 Multilevel Model

Multilevel model, also known as a hierarchical linear model or mixed-effects model, accounts for the nested structure of the data. In this dataset, students are nested within schools.

3.1 Null Model (Simplest Model)

The simplest form of a multilevel model is the null model, which does not include any explanatory variables. It only includes random intercepts for the groups (schools in this case). The null model helps in understanding how much of the total variance in the outcome variable (score) can be attributed to differences between groups (schools).

The equation for the null model is:

$$score_{ij} = \beta_0 + u_{0j} + e_{ij}$$

- $score_{ij}$ is the attainment score of student i in school j.
- β_0 is the overall mean score across all schools.
- u_{0j} is the random effect for school j, capturing the deviation of school j's mean score from the overall mean.
- e_{ij} is the residual error term for student i in school j.

3.2 Single-Level Analysis

A single-level analysis uses a standard linear regression model assuming that all data points are independent and the outcome is normally distributed.

```
m.lm <- lm(score ~ 1, data = score.sch)
summary(m.lm)
```

```
Call:
lm(formula = score ~ 1, data = score.sch)

Residuals:
    Min      1Q  Median      3Q     Max
-31.095 -12.095   1.905  13.905  43.905

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.09462    0.09392   331.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.31 on 33987 degrees of freedom
```

Summary of the Output

1. **Call**: `lm(formula = score ~ 1, data = score.sch)`
    - This indicates that a linear model is being fitted with `score` as the response variable and no predictors (only an intercept).
2. **Residuals**: The residuals section provides a summary of the distribution of the residuals (differences between observed and predicted values):

- **Min**: -31.095
- **1Q (First Quartile)**: -12.095
- **Median**: 1.905
- **3Q (Third Quartile)**: 13.905
- **Max**: 43.905

3. **Coefficients**: The coefficients section provides the estimate of the model's intercept:

   - **Estimate**: 31.09462 This is the estimated mean score for all students.
   - **Std. Error**: 0.09392 This is the standard error of the intercept estimate, indicating the precision of the estimate.
   - **t value**: 331.1 This is the t-statistic for testing whether the intercept is significantly different from zero.
   - **Pr(>|t|)**: <2e-16 This is the p-value associated with the t-statistic, indicating that the intercept is highly significant ($p < 0.001$).

4. **Residual Standard Error**:

   - **Residual standard error**: 17.31 on 33987 degrees of freedom
   - This value measures the average distance that the observed scores fall from the regression line (mean score). A lower residual standard error indicates a better fit.

5. **Interpretation**:

   i) **Overall Mean Score**: The overall mean score for students is estimated to be 31.09462. This value is statistically significant, as indicated by the extremely small p-value ($p < 0.001$).

   ii) **Residuals**: The residuals indicate the spread of the differences between observed scores and the estimated mean score. The range of residuals (-31.095 to 43.905) suggests variability in scores around the mean.

   iii) **Model Fit**: The residual standard error of 17.31 suggests that there is considerable variability in student scores around the mean. This model does not account for any grouping structure (such as schools), meaning it assumes all observations are independent.

6. **Equation**:

   - $\hat{\text{score}} = \beta_0$
   - $\hat{\text{score}} = 31.09462$

7. **Conclusion**: The single-level analysis provides an estimate of the overall mean score (31.09462) for all students, which is highly significant. However, the residual standard error and the spread of residuals indicate that there is substantial variability in scores that is not explained by this simple model. This highlights the potential need for a more

complex model, such as a multilevel model, to account for the nested structure of the data and potentially explain some of the variability in scores.

3.3 Multilevel Analysis

Multilevel analysis (also known as hierarchical linear modeling or mixed-effects modeling) is used to analyze data that has a nested structure. In this dataset, students are nested within schools. This type of analysis allows us to account for variability at both the student level and the school level, providing more accurate estimates and inferences.

1. **Fixed Effects:** The overall mean score β_0 is estimated. These are the effects that are assumed to be constant across all groups. For example, the average effect of a variable like cohort90 on score across all students and schools.
2. **Random Effects:** These allow for variability between groups. For example, different schools may have different baseline scores (intercepts), and the effect of cohort90 on score may vary between schools.
    - **Between-School Variance**: Variance due to differences between schools u_{0j}.
    - **Within-School Variance**: Variance due to differences within schools e_{ij}.

```
m0 <- lmer(score ~ 1 + (1 | schoolid), data = score.sch, REML = FALSE)
summary(m0)
```

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: score ~ 1 + (1 | schoolid)
   Data: score.sch

      AIC       BIC    logLik  deviance  df.resid
 286545.1  286570.4 -143269.5  286539.1     33985

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9763 -0.7010  0.1017  0.7391  3.0817

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolid (Intercept)  61.02    7.812
 Residual             258.36   16.073
Number of obs: 33988, groups:  schoolid, 508
```

```
Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  30.6006     0.3694 451.5326   82.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary of the Output

- **Fixed Effects**: The estimated overall mean score across all schools is 30.6006. This value is statistically significant with a very small p-value (< 2e-16), indicating that the mean score is significantly different from zero.
- **Random Effects**
  - The variance of the school-level intercepts (i.e., the variability in the mean scores between schools) is 61.02, with a standard deviation of 7.812.
  - The variance of the residual errors (i.e., the variability in scores within schools) is 258.36, with a standard deviation of 16.073.
- **Model Fit Statistics**
  - **AIC**: 286545.1, **BIC**: 286570.4
  - These statistics provide measures of the model fit. Lower AIC and BIC values indicate a better-fitting model.
- **Scaled Residuals**: These are the quartiles of the residuals, which provide information about the distribution of the residuals.
- **Interpretation**

  i) **Overall Mean Score**: The overall mean score for students across all schools is 30.6006, which is statistically significant.

  ii) **Variance Components**:
  - The between-school variance (61.02) indicates that there is variability in mean scores between different schools.
  - The within-school variance (258.36) indicates that there is substantial variability in scores within schools.

  iii) **Intraclass Correlation Coefficient (ICC)**:

$$\text{ICC} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}} = \frac{61.02}{61.02 + 258.36} \approx 0.191$$

  - The ICC can be calculated to understand the proportion of total variance that is attributable to the grouping structure (schools in this case).

  - Approximately 19.1% of the total variance in scores is attributable to differences between schools, while the remaining 80.9% is due to differences within schools.

- **Equation** $score_{ij} = 30.6006 + u_{0j} + e_{ij}$
  - $\beta_0 = 30.6006$ is the fixed effect (overall mean score).
  - $u_{0j} \sim N(0, 61.02)$ is the random effect for school (j) with a variance of 61.02.
  - $e_{ij} \sim N(0, 258.36)$ is the residual error term for student (i) in school (j) with a variance of 258.36.

- **Conclusion**: The multilevel analysis shows that there is significant variability in student scores both within and between schools. The mean score across all schools is estimated to be 30.6006, with substantial variability observed within schools and some variability observed between schools. The intraclass correlation coefficient (ICC) indicates that a meaningful proportion (19.1%) of the total variance in scores can be attributed to differences between schools. This justifies the use of a multilevel model, as it captures the hierarchical structure of the data and provides a more nuanced understanding of the variability in student scores.

3.3.1 Random Intercept Models

Random intercept models are a type of multilevel model where each group (e.g., school) has its own intercept but shares the same slope for explanatory variables. This allows the model to account for the fact that different groups may have different baseline levels of the outcome variable.

3.3.1.1 Add an Explanatory Variable

Adding an explanatory variable (predictor) to the random intercept model helps to explain some of the variability in the outcome variable. This allows us to understand how much of the variability is due to the predictor, in addition to the variability captured by the random intercepts.

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + u_{0j} + e_{ij}$$

This equation represents a random intercept model with one explanatory variable (`cohort90`):

- score_{ij}: The score for student i in school j.
- β_0: The overall intercept, representing the average score when all predictors are zero. This is the fixed effect for the intercept.

- $\beta_1 \text{cohort90}_{ij}$: The fixed effect of the explanatory variable `cohort90` on the score. β_1 is the coefficient that quantifies how much the score changes with each unit change in `cohort90`.
- u_{0j}: The random effect for school j. This term captures the deviation of school j's intercept from the overall intercept β_0. It accounts for the fact that different schools may have different baseline scores.
- e_{ij}: The residual error term for student i in school j. This term captures the deviation of the individual student's score from their school's mean score, after accounting for the effects of `cohort90` and the school-level random intercept.

```r
# Adding an explanatory variable (cohort90)
ri_model <- lmer(score ~ cohort90 + (1 | schoolid), data = score.sch, REML = FALSE)
summary(ri_model)
```

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: score ~ cohort90 + (1 | schoolid)
   Data: score.sch

      AIC       BIC    logLik  deviance  df.resid
 280921.6  280955.3 -140456.8  280913.6     33984

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1487 -0.7242  0.0363  0.7339  3.7097

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolid (Intercept)  45.99    6.781
 Residual             219.29   14.808
Number of obs: 33988, groups:  schoolid, 508

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept) 3.056e+01  3.225e-01 4.326e+02   94.74   <2e-16 ***
cohort90    1.215e+00  1.553e-02 3.392e+04   78.24   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
         (Intr)
cohort90 -0.002
```

effect	group	term	estimate	std.error	statistic	df	p.value	conf.l
fixed	NA	(Intercept)	30.559151	0.3225430	94.74442	432.5598	0	29.9252
fixed	NA	cohort90	1.214954	0.0155293	78.23636	33924.8362	0	1.1845
ran_pars	schoolid	sd__(Intercept)	6.781462	NA	NA	NA	NA	N
ran_pars	Residual	sd__Observation	14.808372	NA	NA	NA	NA	N

```
# For a nicer output:
tidy(ri_model, conf.int = TRUE) %>% kbl %>% kable_styling()
```

Summary of the Output

1. **Fixed Effects**:
    - **(Intercept)**: The estimated overall mean score when `cohort90` is zero is 30.56. This value is statistically significant with a very small p-value (< 2e-16), indicating that the intercept is significantly different from zero.
    - **cohort90**: The coefficient for `cohort90` is 1.215. This means that for each unit increase in `cohort90`, the score increases by 1.215 points on average. This effect is also highly significant (p < 2e-16).
2. **Random Effects**:
    - **schoolid (Intercept)**: **Variance**: 45.99, **Std.Dev.**: 6.781. This indicates that the variance of the school-level intercepts (i.e., the variability in the mean scores between schools) is 45.99, with a standard deviation of 6.781.
    - **Residual**: **Variance**: 219.29, **Std.Dev.**: 14.808. This indicates the variance of the residual errors (i.e., the variability in scores within schools) is 219.29, with a standard deviation of 14.808.
    - **Number of Observations**: 33,988 students
    - **Number of Groups (schools)**: 508 schools
3. **Model Fit Statistics**: These statistics provide measures of the model fit. Lower AIC and BIC values indicate a better-fitting model.
    - **AIC**: 280921.6, **BIC**: 280955.3
4. **Scaled Residuals**: These are the quartiles of the residuals, which provide information about the distribution of the residuals.
    - **Min**: -3.1487, **1Q**: -0.7242, **Median**: 0.0363, **3Q**: 0.7339, **Max**: 3.7097
5. **Interpretation**:

i. **Overall Mean Score**: The overall mean score for students, when `cohort90` is zero, is estimated to be 30.56.

ii. **Effect of cohort90**: The score increases by 1.215 points for each unit increase in `cohort90`. This indicates a positive relationship between `cohort90` and `score`.

iii. **Variance Components**: The variance of the school-level random intercepts is 45.99, indicating variability in mean scores between schools. The residual variance is 219.29, indicating variability in scores within schools.

iv. **Intraclass Correlation Coefficient (ICC)**: Approximately 17.3% of the total variance in scores is attributable to differences between schools, while the remaining 82.7% is due to differences within schools.

$$\text{ICC} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}} = \frac{45.99}{45.99 + 219.29} \approx 0.173$$

v. **Model Fit**: The model fit statistics (AIC, BIC, logLik, deviance) suggest that the model fits the data reasonably well. Lower AIC and BIC values compared to other models would indicate a better fit.

6. **Conclusion**: The random intercept model with `cohort90` as an explanatory variable shows that there is significant variability in student scores both within and between schools. The average score increases with higher values of `cohort90`, suggesting a positive trend over time or across cohorts. The random intercepts capture the variability between schools, and the residual variance captures the variability within schools. The intraclass correlation coefficient (ICC) indicates that a meaningful proportion of the variance in scores is due to differences between schools, justifying the use of a multilevel model.

3.3.1.2 Prediction

In multilevel modeling, the goal of prediction is to estimate the outcome variable (in this case, `score`) for each individual observation by considering both the fixed effects and the random effects. This allows for more accurate predictions by taking into account the hierarchical structure of the data. The general form of the prediction equation in a random intercept model with one explanatory variable (`cohort90`) is:

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + u_{0j} + e_{ij}$$

3.3.1.2.1 Steps for Prediction

1. Fitting the model - use the random intercept model using the `lmer` function earlier.
2. Generating Predicted Values To generate predicted values for the outcome variable, use the `fitted` function. This function provides the fitted values (predictions) for each observation in the dataset based on the fixed and random effects:

```
pred_score <- fitted(ri_model)
head(pred_score, 10)  # Display the first 10 predicted scores
```

```
       1        2        3        4        5        6        7        8
16.54111 16.54111 16.54111 16.54111 16.54111 16.54111 16.54111 16.54111
       9       10
16.54111 16.54111
```

3. Extracting Random Effects The random effects (school-specific intercepts) can be extracted using the `ranef` function. This provides the deviations of each school's intercept from the overall intercept:

```
rand_ef <- ranef(ri_model)
datatable(head(rand_ef$schoolid, 12))  # Display the random intercepts for the first 12 schoo
```

4. Using `broom.mixed::augment` The `broom.mixed` package's `augment` function creates a data frame that includes the original data, fitted values, and residuals. This can be useful for further analysis and visualization:

```
ri_fitted <- augment(ri_model)
ri_fitted %>% slice(1:12)  # Display the first 12 rows of the augmented data frame
```

```
# A tibble: 12 x 14
   score cohort90 schoolid .fitted .resid   .hat  .cooksd .fixed   .mu .offset
   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <dbl>    <dbl>  <dbl> <dbl>   <dbl>
 1     0       -6        1    16.5 -16.5  0.0219 0.0143     23.3  16.5       0
 2    10       -6        1    16.5  -6.54 0.0219 0.00223    23.3  16.5       0
 3     0       -6        1    16.5 -16.5  0.0219 0.0143     23.3  16.5       0
 4    40       -6        1    16.5  23.5  0.0219 0.0287     23.3  16.5       0
 5    42       -6        1    16.5  25.5  0.0219 0.0338     23.3  16.5       0
 6     4       -6        1    16.5 -12.5  0.0219 0.00819    23.3  16.5       0
 7     0       -6        1    16.5 -16.5  0.0219 0.0143     23.3  16.5       0
 8     0       -6        1    16.5 -16.5  0.0219 0.0143     23.3  16.5       0
 9    14       -6        1    16.5  -2.54 0.0219 0.000336   23.3  16.5       0
10    27       -6        1    16.5  10.5  0.0219 0.00570    23.3  16.5       0
11    18       -2        1    21.4  -3.40 0.0219 0.000603   28.1  21.4       0
12    23       -4        1    19.0   4.03 0.0219 0.000845   25.7  19.0       0
# i 4 more variables: .sqrtXwt <dbl>, .sqrtrwt <dbl>, .weights <dbl>,
#   .wtres <dbl>
```

3.3.1.2.2 Explanation

1. **Fixed Effects**:
    - The fixed effects (β_0 and β_1) provide the overall intercept and the effect of `cohort90` on `score`.
    - These effects are assumed to be the same across all schools.
2. **Random Effects**:
    - The random effects (u_{0j}) represent the school-specific deviations from the overall intercept.
    - Each school has its own intercept, which is the sum of the overall intercept (β_0) and the school's random intercept (u_{0j}).
3. **Residuals**:

- The residuals (e_{ij}) represent the individual-specific deviations from the school's predicted score.
- These capture the within-school variability that is not explained by the model.

4. **Combined Prediction**:

   - The predicted score for each student is the sum of the overall intercept, the effect of `cohort90`, the school-specific random intercept, and the individual-specific residual.
   - This combined approach ensures that the prediction accounts for both the fixed effects (common to all students) and the random effects (specific to each school).

3.3.1.2.3 Confirmation with Manual Calculation

Manually calculate the fitted values (predicted scores) to confirm the results from the model.

1. **Intercept (β_0)**: 30.55915. This is the overall average score when `cohort90` is zero.
2. **Level-2 Residual (school level residual, u_{0j})**: For `schoolid = 1`, this is -6.73. This represents the deviation of this school's intercept from the overall intercept.
3. **Coefficient for `cohort90` (β_1)**: 1.214955. This indicates the average change in score for each unit change in `cohort90`.
4. **Example**: For the first observation = **cohort90**: -6, **schoolid**: 1

   - The fitted value (predicted score) can be calculated as:

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + u_{0j} + e_{ij}$$

   - Since e_{ij} (the individual-level residual) is what we are trying to predict, we simplify to:

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + u_{0j}$$

   - Using the provided values:

$$\text{score}_{ij} = 30.55915 - 6.728309492 + 1.214955 \times (-6)$$

$$\text{score}_{ij} = 30.55915 - 6.728309492 - 7.28973$$

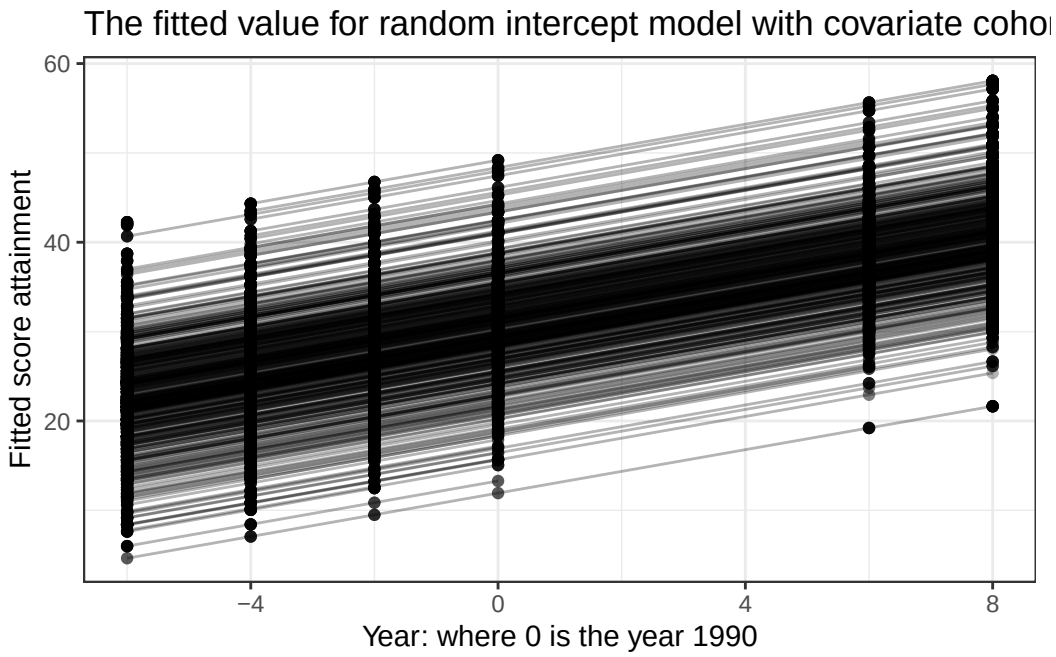
$$\text{score}_{ij} = 16.54111$$

   - This matches the calculated fitted value.

3.3.1.3 Plot

Plotting the model helps to visualize the fitted values from a random intercept model with an explanatory variable, such as `cohort90`. Visualization helps in understanding the model fit and the relationship between the variables.

```
ggplot(ri_fitted, aes(cohort90, .fitted, group = schoolid)) +
  geom_point(alpha = 0.3) +
  geom_line(alpha = 0.3) +
  ylab('Fitted score attainment') +
  xlab('Year: where 0 is the year 1990') +
  ggtitle('The fitted value for random intercept model with covariate cohort90') +
  theme_bw()
```



Comment: This plot visualizes the fitted values from the random intercept model with cohort90 as the explanatory variable. It provides a visual confirmation of the model's results, showing how the fitted scores vary across different cohort years and schools. The general upward trend indicates improving scores over time, while the variation between schools highlights the importance of accounting for school-specific effects in the model. This visualization helps in understanding the combined impact of fixed and random effects on the predicted scores.

3.3.1.4 Variance

3.3.1.4.1 Between School Variance

Between-school variance refers to the variation in the outcome variable (score) that is attributed to differences between schools. This variance component is crucial in multilevel modeling as it helps to understand how much of the total variance in scores can be explained by the differences between schools.

1. **Constant Only Model**:
    - In the constant only (null) model, the variance due to differences between schools (random intercept) is 61.02.
    - This model does not include any explanatory variables and serves as a baseline to compare other models.
2. **Model with Explanatory Variable (cohort90)**:
    - When `cohort90` is added as an explanatory variable, the between-school variance reduces to 45.99.
    - This reduction indicates that some of the variance between schools is explained by the cohort year.
3. **Proportion of Unexplained Variance**:
    - After accounting for the cohort effects, the proportion of unexplained variance due to differences between schools is calculated as:

$$\frac{45.99}{45.99 + 219.29} = 17\%$$

    - This means that 17% of the total unexplained variance in scores is attributable to differences between schools after accounting for the cohort effect.

3.3.1.4.2 Within School Variance

Within-school variance refers to the variation in the outcome variable (score) within schools. This component captures the differences in scores among students within the same school. Understanding this variance is essential for identifying how much of the total variance is due to individual differences within schools.

1. **Constant Only Model**:
    - In the constant only (null) model, the within-school variance (residual variance) is 258.36.
    - This model does not include any explanatory variables and serves as a baseline to compare other models.
2. **Model with Explanatory Variable (cohort90)**:

- When `cohort90` is added as an explanatory variable, the within-school variance reduces to 219.29.
- This reduction indicates that some of the variance within schools is explained by the cohort year.

3. **Reduction in Variance**:

   - The addition of `cohort90` reduces both the between-school and within-school variances.
   - The between-school variance reduces from 61.02 to 45.99.
   - The within-school variance reduces from 258.36 to 219.29.
   - The decrease in within-school variance is expected because `cohort90` is a student-level variable, meaning it explains part of the individual differences within schools.

3.3.2 Random Slope Models

Random slope models allow both the intercept and the slope of the regression line to vary across groups (e.g., schools). This flexibility accounts for the possibility that different schools may not only have different starting points (intercepts) but also different rates of change (slopes) in the outcome variable. The random slope model is an extension of the random intercept model. It can be expressed as:

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + u_{0j} + u_{1j}\text{cohort90}_{ij} + e_{ij}$$

- β_0: Overall intercept.
- β_1: Overall slope for `cohort90`.
- u_{0j}: Random intercept for school j.
- u_{1j}: Random slope for `cohort90` for school j.
- cohort90_{ij}: Value of `cohort90` for student i in school j.
- e_{ij}: Residual error term for student i in school j.

```
# Random Slope Model
rs_model1 <- lmer(score ~ cohort90 + (1 + cohort90 | schoolid), data = score.sch, REML = FALS
```

```
Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
Model failed to converge with max|grad| = 0.00542168 (tol = 0.002, component 1)
```

```
summary(rs_model1)
```

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: score ~ cohort90 + (1 + cohort90 | schoolid)
   Data: score.sch

      AIC       BIC    logLik  deviance  df.resid
 280698.2  280748.8 -140343.1  280686.2     33982

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1008 -0.7202  0.0387  0.7264  3.5220

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schoolid (Intercept)  42.8573  6.5465
          cohort90      0.1606  0.4008  -0.39
 Residual             215.7393 14.6881
Number of obs: 33988, groups:  schoolid, 508

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  30.60967    0.31344 426.77095   97.66   <2e-16 ***
cohort90      1.23391    0.02532 316.40138   48.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
         (Intr)
cohort90 -0.266
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.00542168 (tol = 0.002, component 1)
```

```r
# Model failed to converge, switch to `bobyqa` optimizer
rs_model2 <- lmer(score ~ cohort90 + (1 + cohort90 | schoolid), data = score.sch, control = l
summary(rs_model2)
```

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: score ~ cohort90 + (1 + cohort90 | schoolid)
   Data: score.sch
Control: lmerControl(optimizer = "bobyqa")
```

effect	group	term	estimate	std.error	statistic	df	p.value
fixed	NA	(Intercept)	30.6096334	0.3134467	97.65499	426.7634	0
fixed	NA	cohort90	1.2339026	0.0253138	48.74420	316.4513	0
ran_pars	schoolid	sd__(Intercept)	6.5466118	NA	NA	NA	NA
ran_pars	schoolid	cor__(Intercept).cohort90	-0.3903901	NA	NA	NA	NA
ran_pars	schoolid	sd__cohort90	0.4007387	NA	NA	NA	NA
ran_pars	Residual	sd__Observation	14.6880667	NA	NA	NA	NA

```
     AIC       BIC    logLik  deviance  df.resid
 280698.2  280748.8 -140343.1  280686.2     33982

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1008 -0.7202  0.0387  0.7264  3.5220

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schoolid (Intercept)  42.8581  6.5466
          cohort90      0.1606  0.4007  -0.39
 Residual             215.7393 14.6881
Number of obs: 33988, groups:  schoolid, 508

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  30.60963    0.31345 426.76336   97.66   <2e-16 ***
cohort90      1.23390    0.02531 316.45130   48.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
         (Intr)
cohort90 -0.266
```

```
# For a nicer output
tidy(rs_model2, conf.int = TRUE) %>% kbl %>% kable_styling()
```

Summary of Output

1. **Model Fit Statistics**

   - **AIC**: 280698.2, **BIC**: 280748.8
   - Lower AIC and BIC values indicate a better fit of the model to the data.

2. **Scaled Residuals**

   - **Min**: -3.1008, **1Q (First Quartile)**: -0.7202, **Median**: 0.0387, **3Q (Third Quartile)**: 0.7264, **Max**: 3.5220
   - The residuals have a median close to zero, indicating a good fit of the model. The range of residuals (-3.1008 to 3.5220) suggests that there are no extreme outliers.

3. **Random Effects**

   - **Variance of Intercepts (schoolid)**: The variance of 42.8581 (standard deviation of 6.5466) indicates significant variability in the baseline scores between schools.
   - **Variance of Slopes (cohort90)**: The variance of 0.1606 (standard deviation of 0.4007) indicates that the effect of `cohort90` on scores varies across schools.
   - **Correlation between Intercepts and Slopes**: The correlation of -0.39 suggests that schools with higher baseline scores tend to have a less steep increase in scores with respect to `cohort90`.
   - **Residual Variance**: 215.7393, Standard Deviation: The variance of 215.7393 (standard deviation of 14.6881) indicates variability in scores within schools.

4. **Fixed Effects**

   - **Intercept (β_0)**: The average score when `cohort90` is zero is 30.60963, which is statistically significant ($p < 2e\text{-}16$).
   - **cohort90 (β_1)**: The average effect of `cohort90` on scores is 1.23390, indicating that for each unit increase in `cohort90`, the score increases by 1.23390 points, which is also statistically significant ($p < 2e\text{-}16$).

5. **Correlation of Fixed Effects**

   - **Intercept and cohort90**: The negative correlation (-0.266) between the intercept and `cohort90` indicates that higher initial scores are associated with a slightly lower effect of `cohort90` on scores.

6. **Conclusion**: The random slope model indicates that student scores increase with `cohort90`, with an average increase of 1.2339 points per unit increase in `cohort90`, which is statistically significant. The intercept, representing the average score when `cohort90` is zero, is 30.60963. Significant variability exists both between schools (intercept variance of 42.8581) and in how `cohort90` affects scores across schools (slope variance of 0.1606). The negative correlation (-0.39) between the intercept and slope suggests that schools with higher initial scores tend to have a slower increase in scores over time. Residual variance of 215.7393 indicates considerable within-school variability in scores. Overall, the model effectively captures the hierarchical data structure, revealing important insights into the factors influencing student scores.

3.3.2.1 The Fitted Values

Fitted values are the predicted scores obtained from the model, incorporating both fixed effects and random effects. These values help in understanding how well the model fits the data and in interpreting the effects of the explanatory variables at different levels.

```
rs_res <- augment(rs_model2) # Fitted random slope model
datatable(head(rs_res, 20)) # Display the first 20 fitted values
```

Show 10 entries   Search:

	score	cohort90	schoolid	.fitted	.resid	.hat	.cooksd	.fixed	.mu	.offset	.sqrtXwt	.sqrtrwt	.weights	.wtres
1	0	-6	1	16.11345608304478	-16.11345608304478	0.02521543893746914	0.01596861972578989	23.20621800792647	16.11345608304478	0	1	1	1	-16.11345608304478
2	10	-6	1	16.11345608304478	-6.113456083044785	0.02521543893746914	0.002298600901840579	23.20621800792647	16.11345608304478	0	1	1	1	-6.113456083044785
3	0	-6	1	16.11345608304478	-16.11345608304478	0.02521543893746914	0.01596861972578989	23.20621800792647	16.11345608304478	0	1	1	1	-16.11345608304478
4	40	-6	1	16.11345608304478	23.88654391695522	0.02521543893746914	0.03509107024266908	23.20621800792647	16.11345608304478	0	1	1	1	23.88654391695522
5	42	-6	1	16.11345608304478	25.88654391695522	0.02521543893746914	0.04121336957540039	23.20621800792647	16.11345608304478	0	1	1	1	25.88654391695522
6	4	-6	1	16.11345608304478	-12.11345608304478	0.02521543893746914	0.009024561679956638	23.20621800792647	16.11345608304478	0	1	1	1	-12.11345608304478
7	0	-6	1	16.11345608304478	-16.11345608304478	0.02521543893746914	0.01596861972578989	23.20621800792647	16.11345608304478	0	1	1	1	-16.11345608304478
8	0	-6	1	16.11345608304478	-16.11345608304478	0.02521543893746914	0.01596861972578989	23.20621800792647	16.11345608304478	0	1	1	1	-16.11345608304478
9	14	-6	1	16.11345608304478	-2.113456083044785	0.02521543893746914	0.0002747112435190886	23.20621800792647	16.11345608304478	0	1	1	1	-2.113456083044785
10	27	-6	1	16.11345608304478	10.88654391695522	0.02521543893746914	0.007289035024475487	23.20621800792647	16.11345608304478	0	1	1	1	10.88654391695522

Showing 1 to 10 of 20 entries   Previous 1 2 Next

3.3.2.1.1 Interpretation of the Fitted Values

1. **Cohort Effect**:

   - The cohort effect for school j is estimated as $1.234 + \hat{u}_{1j}$.
   - The between-school variance in these slopes is estimated as 0.160.
   - This means that for the average school, we predict an increase of 1.234 points in the attainment score for each successive cohort year.

2. **Intercept Variance**:

   - The intercept variance of 42.858 is interpreted as the between-school variance when `cohort90` $= 0$, i.e., for the 1990 cohort.
   - This indicates the variability in baseline scores (intercepts) between schools.

3. **Intercept-Slope Correlation**:

    - The intercept-slope correlation is estimated as -0.39.
    - This suggests that schools with higher intercepts (above-average attainment in 1990) tend to have a flatter-than-average slope.
    - In other words, schools that started with higher scores in 1990 show less increase in scores over successive cohorts.

4. Example for the first observation from the table.

    - The fitted value (16.1) is calculated by adding the fixed effect (23.2) and the random effect (16.1).
    - The residual (-16.1) is the difference between the observed score and the fitted value.

3.3.2.2 Model Comparison: Random Intercept vs. Random Slope

Comparing models using the `anova` function helps to statistically test whether the addition of random slopes significantly improves the fit of the model compared to a simpler random intercept model. This is done through a likelihood ratio test.

```
anova(ri_model, rs_model2)
```

```
Data: score.sch
Models:
ri_model: score ~ cohort90 + (1 | schoolid)
rs_model2: score ~ cohort90 + (1 + cohort90 | schoolid)
          npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
ri_model     4 280922 280955 -140457   280914
rs_model2    6 280698 280749 -140343   280686 227.4  2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Output Summary

1. **Number of Parameters (npar)**:

    - **ri_model**: 4 parameters
    - **rs_model2**: 6 parameters
    - The random slope model has more parameters due to the additional random slope component.

2. **AIC (Akaike Information Criterion)**:

    - **ri_model**: 280922

- **rs_model2**: 280698
- The lower AIC for the random slope model indicates a better fit, despite the increased complexity.

3. **Log-Likelihood (logLik)**:

   - **ri_model**: -140457
   - **rs_model2**: -140343
   - The higher log-likelihood for the random slope model indicates a better fit.

4. **Chi-Square Statistic (Chisq)**: 227.4, which represents the improvement in fit by adding the random slope.

5. **p-Value (Pr(>Chisq))**:

   - The p-value is less than 2.2e-16, which is highly significant.
   - This indicates that the random slope model provides a significantly better fit to the data than the random intercept model.

ANOVA Interpretation

- **Model Fit**: The random slope model (`rs_model2`) fits the data significantly better than the random intercept model (`ri_model`). This is evidenced by the lower AIC, higher log-likelihood, and significant Chi-square test result.
- **Statistical Significance**: The p-value from the likelihood ratio test is less than 2.2e-16, which is highly significant. This means that the improvement in model fit when adding random slopes is not due to chance, and the additional complexity of the random slope model is justified.

3.3.2.3 Interpretation of Random Effects Across Schools

Random effects capture the variability in the data at different levels, such as between schools. Understanding these effects is crucial for making meaningful inferences about how different schools deviate from the overall population average.

1. **Cohort Effect for School j**:

   - The cohort effect for school j is estimated as $1.234 + \hat{u}\_1j$.
   - Here, 1.234 is the overall average effect of `cohort90` across all schools, and $\hat{u}_{1j}$ is the random slope for school $ j $, representing the deviation of school $ j $ from this overall average.

2. **Between-School Variance in Slopes**:

   - The between-school variance in the slopes is estimated as 0.160.

- This indicates the variability in the effect of `cohort90` on scores across different schools.

3. **Average School Prediction**: For the average school, the predicted increase in attainment score for each successive cohort is 1.234 points.
4. **95% Coverage Interval for School Slopes**:

$$1.234 \pm 1.96\sqrt{0.160} = 1.234 \pm 1.96 \times 0.400 = 1.234 \pm 0.784 = 0.45 \text{ to } 2.018$$

   - This interval suggests that, assuming a normal distribution, we would expect the middle 95% of schools to have a slope between 0.45 and 2.018.

3.3.2.4 Prediction from Random Slope

This section discusses how to obtain and interpret the predictions from a random slope model, specifically focusing on the random effects for each school. The random intercepts indicate how much each school's average score deviates from the overall average, while the random slopes indicate how much the effect of `cohort90` on scores deviates for each school. The random effects for the random slope model can be extracted using the `ranef` function. The `condVar = TRUE` argument includes the conditional variances of the random effects in the output.

```
ra.eff.rs <- ranef(rs_model2, condVar = TRUE)
datatable(ra.eff.rs$schoolid)
```

Show 10 entries Search:

	(Intercept)	cohort90
1	-5.904643002629726	0.1980198203753271
2	2.683567550250376	0.04868371276024705
3	1.727236829279539	0.1499586586340995
4	-7.802839083190386	0.3047288998198396
5	3.295991406917849	-0.4539884081985074
6	12.11617012636377	-0.5801553022264188
7	-1.64037368522911	0.008941235818721596
8	18.04779422354533	-0.2193145178773735
9	-7.708387317728408	0.184207343887608
10	2.764884270839438	0.2024456928972398

Showing 1 to 10 of 508 entries   Previous 1 2 3 4 5 … 51 Next

- The table shows the random intercepts and random slopes for each school.
- Each row corresponds to a school, identified by an ID (e.g., 1, 2, 3, etc.).
- **(Intercept)**: The random intercept for the school, indicating how much the school's average score deviates from the overall average score.
- **cohort90**: The random slope for the school, indicating how much the effect of `cohort90` (the cohort year) on the score deviates from the overall effect.

3.3.2.5 Plot of Random Effects

This section discusses how to visualize random effects from a random slope model. Visualizing these effects helps in understanding the variations across different groups—in this case, schools. In a random slope model, random effects can be categorized into:

1. **Random Intercepts**: Differences in starting points (intercepts) for each school.
2. **Random Slopes**: Differences in the rate of change (slopes) for each school.

The objective is to create a scatter plot where:

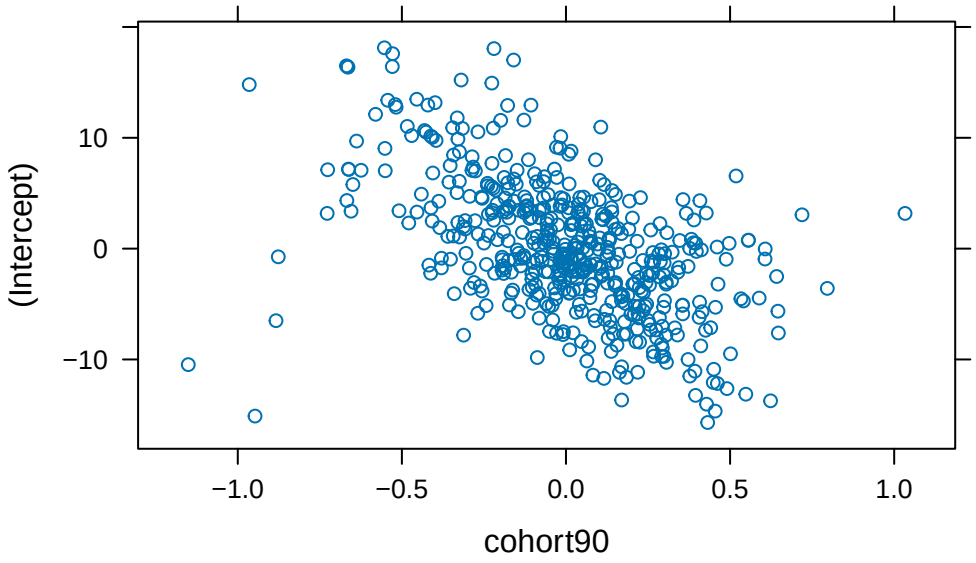
- The x-axis represents the random intercepts.
- The y-axis represents the random slopes.

3.3.2.5.1 Scatter Plot 1

To quickly visualize the random effects extracted from a mixed-effects model using base R graphics.

```
plot(ra.eff.rs)
```

```
$schoolid
```



Interpretation: The scatter plot presents the relationship between random intercepts and random slopes for different schools. There appears to be a downward trend in the scatter plot, indicating a negative correlation between random intercepts and random slopes. Schools with higher intercepts tend to have lower slopes, suggesting that schools with higher starting scores improve at a slower rate.

3.3.2.5.2 Scatter Plot 2

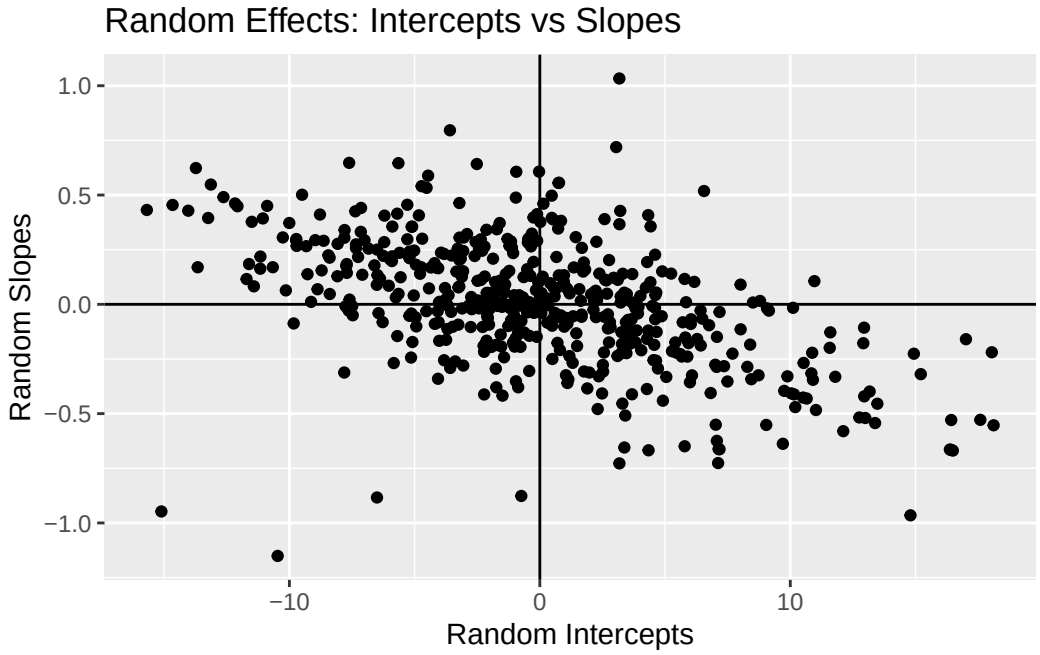
To create a customized scatter plot of random intercepts vs. random slopes using `ggplot2`.

```
# Extract random effects for `schoolid`
ra.eff.rs.sc <- ra.eff.rs$schoolid

# Rename columns for clarity
```

```r
ra.eff.rs.sc <- ra.eff.rs.sc %>% rename(rs_slope = cohort90, rs_int = "(Intercept)")

# Plot random effects
ggplot(ra.eff.rs.sc, aes(x = rs_int, y = rs_slope)) +
  geom_point() +
  geom_vline(xintercept = 0) +
  geom_hline(yintercept = 0) +
  xlab("Random Intercepts") +
  ylab("Random Slopes") +
  ggtitle("Random Effects: Intercepts vs Slopes")
```



Interpretation: The scatter plot visualizes the relationship between random intercepts and random slopes for different schools. There is a slight downward trend in the scatter plot, indicating a negative correlation between random intercepts and random slopes. Schools with higher intercepts tend to have lower slopes, suggesting that schools starting with higher scores improve at a slower rate.

3.3.2.6 Equation for Random Slope Model

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + u_{0j} + u_{1j}\text{cohort90}_{ij} + e_{ij}$$
$$\hat{score}_{ij} = (30.610 + \hat{u}_{0j}) + (1.234 + \hat{u}_{1j})\text{cohort90}_{ij}$$

- $\hat{score}_{ij}$: The predicted score attainment for student i in school j.
- 30.610: The fixed part of the intercept, representing the average score across all schools when `cohort90` is zero (the year 1990).
- $\hat{u}_{0j}$: The random intercept for school j, which captures the deviation of school j's intercept from the overall intercept (30.610).
- 1.234: The fixed part of the slope, representing the average change in score attainment per unit change in `cohort90`.
- $\hat{u}_{1j}$: The random slope for school j, which captures the deviation of school j's slope from the overall slope (1.234).

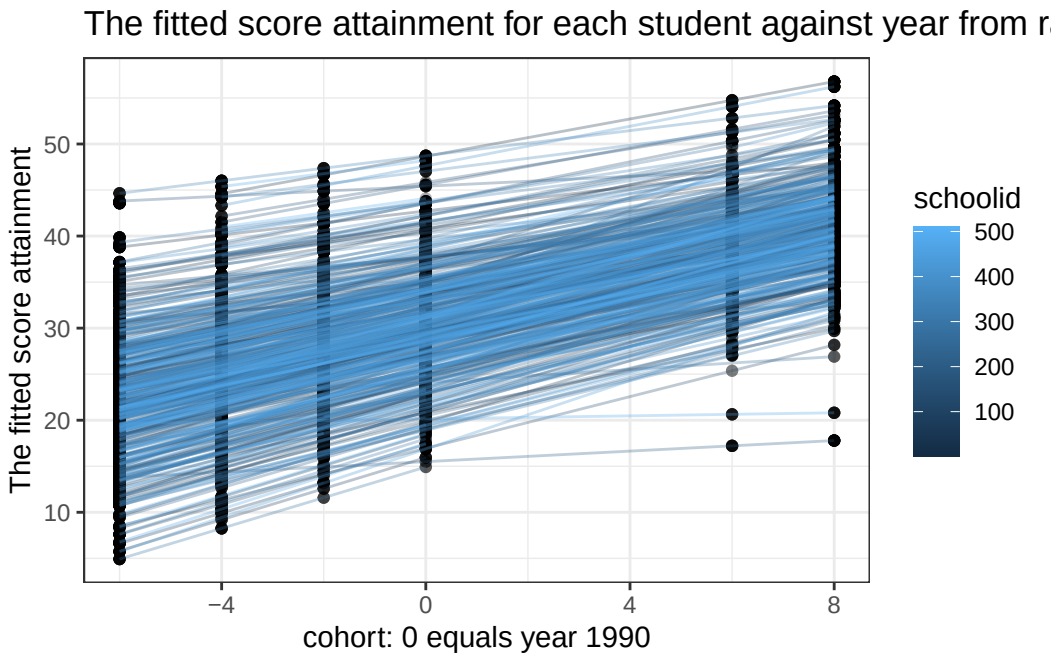
Interpretation: The model indicates that on average, students' scores increase by 1.234 points for each unit increase in `cohort90`. Each school has its own intercept $(30.610 + \hat{u}_{0j})$ and slope $(1.234 + \hat{u}_{1j})$, allowing for variability in how different schools start and how the scores change over time.

3.3.2.7 Plot the Fitted Values from Random Slope

The random slope model allows both the intercept and the slope to vary across different groups (schools). By plotting the fitted values from the random slope model, we can effectively visualize the differences in trends and baselines across the groups (schools) in the dataset.

Use the extracted fitted values (`rs_res`) which was generated in The Fitted Values.

```
rs_res %>%
  ggplot(aes(cohort90, .fitted, group = schoolid)) +
  geom_point(alpha = 0.3) +
  geom_line(aes(colour = schoolid), alpha = 0.3) +
  ylab('The fitted score attainment') +
  xlab('cohort: 0 equals year 1990') +
  theme_bw() +
  ggtitle('The fitted score attainment for each student against year from random slope model
```



Interpretation : The graph presented shows the fitted score attainment for each student against the cohort year from a random slope model. The overall trend shows that scores generally increase over time (as cohort year increases), indicating that later cohorts tend to have higher scores. The positive slope in most lines indicates improvement in scores over time for most schools. There is considerable variability in the slopes and intercepts of the lines, reflecting differences in how scores change over time across different schools. Some schools start at a higher baseline score (higher intercept) and show steady improvement, while others start lower and improve more gradually.

3.3.2.8 Adding a Level-1 Variable to the Random Slope Model

Random slope model can be extended by incorporating an additional level-1 explanatory variable (for example, gender). By doing this, the effect of gender on score attainment across schools can be analyzed. The model equation is extended to:

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + \beta_2 \text{female}_{ij} + u_{0j} + u_{1j} \text{cohort90}_{ij} + e_{ij}$$

- β_0 is the overall intercept.
- β_1 is the coefficient for the cohort90 variable.
- β_2 is the coefficient for the gender variable (female).
- u_{0j} is the random intercept for school j.
- u_{1j} is the random slope for the cohort90 variable within school j. e_{ij} is the residual error term.

3.3.2.8.1 Random Slope Model with Gender

```
# Model fitting
rs_gend1 <- lmer(score ~ cohort90 + female2 + (1 + cohort90 | schoolid), data = score.sch, RE
```

```
Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
Model failed to converge with max|grad| = 0.00245224 (tol = 0.002, component 1)
```

```
summary(rs_gend1)
```

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: score ~ cohort90 + female2 + (1 + cohort90 | schoolid)
   Data: score.sch

      AIC       BIC    logLik  deviance  df.resid
 280558.1  280617.2 -140272.1  280544.1     33981

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1617 -0.7185  0.0395  0.7282  3.5326

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schoolid (Intercept)  42.5750  6.5250
          cohort90      0.1613  0.4016  -0.39
 Residual             214.8370 14.6573
Number of obs: 33988, groups:  schoolid, 508

Fixed effects:
               Estimate Std. Error        df t value Pr(>|t|)
(Intercept)   2.958e+01  3.241e-01 4.944e+02   91.30   <2e-16 ***
cohort90      1.227e+00  2.533e-02 3.168e+02   48.46   <2e-16 ***
female2female 1.945e+00  1.630e-01 3.364e+04   11.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) chrt90
cohort90    -0.253
female2feml -0.265 -0.022
```

```
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.00245224 (tol = 0.002, component 1)
```

> **Note**
>
> Upon running the model, a warning indicates that the model did not converge due to the gradient being too high. This means the default optimizer was unable to find a suitable solution within the specified tolerance.

```
# model is re-fitted using the `bobyqa` optimizer
rs_gend2 <- lmer(score ~ cohort90 + female2 + (1 + cohort90 | schoolid), data = score.sch, RE
summary(rs_gend2)
```

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: score ~ cohort90 + female2 + (1 + cohort90 | schoolid)
   Data: score.sch
Control: lmerControl(optimizer = "bobyqa")

      AIC       BIC    logLik  deviance  df.resid
 280558.1  280617.2 -140272.1  280544.1     33981

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1617 -0.7185  0.0395  0.7282  3.5326

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schoolid (Intercept)  42.5750  6.5249
          cohort90      0.1613  0.4016  -0.39
 Residual             214.8374 14.6573
Number of obs: 33988, groups:  schoolid, 508

Fixed effects:
               Estimate Std. Error        df t value Pr(>|t|)
(Intercept)   2.958e+01  3.241e-01 4.944e+02   91.30   <2e-16 ***
cohort90      1.227e+00  2.533e-02 3.169e+02   48.46   <2e-16 ***
female2female 1.945e+00  1.630e-01 3.364e+04   11.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

effect	group	term	estimate	std.error	statistic	df	p.value
fixed	NA	(Intercept)	29.5848725	0.3240553	91.29575	494.3961	0
fixed	NA	cohort90	1.2273265	0.0253268	48.45966	316.9007	0
fixed	NA	female2female	1.9445257	0.1629805	11.93104	33639.7891	0
ran_pars	schoolid	sd___(Intercept)	6.5249483	NA	NA	NA	NA
ran_pars	schoolid	cor___(Intercept).cohort90	-0.3933045	NA	NA	NA	NA
ran_pars	schoolid	sd___cohort90	0.4015806	NA	NA	NA	NA
ran_pars	Residual	sd___Observation	14.6573317	NA	NA	NA	NA

```
Correlation of Fixed Effects:
            (Intr) chrt90
cohort90    -0.253
female2feml -0.265 -0.022
```

```
# For a nicer output:
tidy(rs_gend2, conf.int = TRUE) %>% kbl %>% kable_styling()
```

Interpretation:

1. **Random Effects**

   - **Intercept (schoolid)**: The variance in intercepts (42.5750) indicates substantial variability in starting scores between schools.
   - **cohort90 (schoolid)**: The variance in slopes (0.1613) indicates variability in how scores change over time across different schools.
   - **Correlation (-0.39)**: A moderate negative correlation between intercepts and slopes for cohort90 suggests that schools with higher starting scores tend to have slower rates of score increase.

2. **Fixed Effects**

   - **(Intercept)**: The average starting score for a student in 1990 is approximately 29.58.
   - **cohort90**: On average, each additional year (cohort90) increases the score by 1.227 points, which is highly significant.
   - **female2female**: Female students, on average, score 1.945 points higher than male students, which is also highly significant.

3. **Conclusion**: The model reveals significant effects of both `cohort90` and `female` on scores. The random effects show variability across schools in both starting scores and the rate of change over time. The negative correlation between intercepts and slopes indicates that schools with higher initial scores tend to improve more slowly over time.

3.3.2.8.2 Random Slope Model with Gender Having a Random Slope

The model is further extended to allow the gender variable (female) to have a random slope. This means we are considering the possibility that the effect of gender on score attainment might vary across different schools. The model equation becomes:

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + \beta_2 \text{female}_{ij} + u_{0j} + u_{1j} \text{cohort90}_{ij} + u_{2j} \text{female}_{ij} + e_{ij}$$

> **Note**
>
> u_{2j} is the random slope for the gender variable within school j.

```
# Model fitting
rs_gend2_slope <- lmer(score ~ cohort90 + female2 + (1 + cohort90 + female2 | schoolid), data
summary(rs_gend2_slope)
```

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: score ~ cohort90 + female2 + (1 + cohort90 + female2 | schoolid)
   Data: score.sch
Control: lmerControl(optimizer = "bobyqa")

      AIC       BIC    logLik  deviance  df.resid
 280558.9  280643.2 -140269.4  280538.9     33978

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1572 -0.7182  0.0388  0.7267  3.5316

Random effects:
 Groups   Name          Variance Std.Dev. Corr
 schoolid (Intercept)    40.5580  6.3685
          cohort90        0.1617  0.4021  -0.39
          female2female   1.3711  1.1710   0.21 -0.11
 Residual                214.5158 14.6464
Number of obs: 33988, groups:  schoolid, 508

Fixed effects:
               Estimate Std. Error        df t value Pr(>|t|)
(Intercept)    29.58909    0.31766 406.88434   93.15   <2e-16 ***
cohort90        1.22777    0.02535 316.04102   48.44   <2e-16 ***
```

```
female2female   1.93145    0.17390 345.08083   11.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) chrt90
cohort90    -0.253
female2feml -0.201 -0.046
```

Interpretation:

1. **Random Effects**

   - **Intercept (schoolid)**: The variance in intercepts (40.5580) indicates substantial variability in starting scores between schools.
   - **cohort90 (schoolid)**: The variance in slopes (0.1617) indicates variability in how scores change over time across different schools.
   - **female2female (schoolid)**: The variance in slopes (1.3711) indicates variability in the effect of gender (female) on scores across different schools.
   - **Correlation**:
     - A moderate negative correlation (-0.39) between intercepts and slopes for cohort90 suggests that schools with higher starting scores tend to have slower rates of score increase.
     - A positive correlation (0.21) between intercepts and slopes for female2female suggests that schools with higher starting scores tend to have a stronger gender effect.
     - A slight negative correlation (-0.11) between the slopes for cohort90 and female2female suggests a small inverse relationship between these effects.

2. **Fixed Effects**

   - **(Intercept)**: The average starting score for a student in 1990 is approximately 29.58909.
   - **cohort90**: On average, each additional year (cohort90) increases the score by 1.22777 points, which is highly significant.
   - **female2female**: Female students, on average, score 1.93145 points higher than male students, which is also highly significant.

3. **Conclusion**: The model reveals significant effects of both `cohort90` and `female2` on scores. The random effects show variability across schools in both starting scores, the rate of change over time, and the effect of gender. The correlations between random effects suggest relationships between the intercepts and slopes, as well as between the slopes for different variables.

3.3.2.8.3 Comparison Between Models

To compare the random slope model for `cohort90` only with the model including random slopes for both `cohort90` and gender, use the `anova` function.

```
anova(rs_gend2, rs_gend2_slope)
```

```
Data: score.sch
Models:
rs_gend2: score ~ cohort90 + female2 + (1 + cohort90 | schoolid)
rs_gend2_slope: score ~ cohort90 + female2 + (1 + cohort90 + female2 | schoolid)
               npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
rs_gend2          7 280558 280617 -140272   280544
rs_gend2_slope   10 280559 280643 -140269   280539 5.2362  3     0.1553
```

The `anova` function performs a likelihood ratio test to compare the two models. The output will include statistics such as AIC, BIC, log-likelihood, deviance, and the chi-square test for the difference in deviance.

Interpretation:

1. **Model Fit Metrics**: The AIC values are very close for both models, indicating that the addition of random slopes for `female2` does not significantly improve the model fit.
2. **Chi-Square Test**: A p-value of 0.1553 indicates that the difference in deviance between the two models is not statistically significant at the typical alpha level of 0.05. Therefore, the additional complexity of allowing `female2` to have a random slope is not justified based on this test.
3. **Conclusion**:
    - The addition of random slopes for `female2` (in `rs_gend2_slope`) does not significantly improve the model fit compared to having only random slopes for `cohort90` (in `rs_gend2`).
    - Given the p-value of 0.1553, the simpler model (`rs_gend2`) is preferred because it provides a similar fit to the data with fewer parameters.

- This suggests that the effect of gender on score attainment does not vary significantly across schools, or at least not enough to justify the added complexity in the model.
- The results indicate that while there is some variability in the slopes for `cohort90` across schools, allowing the gender effect (`female2`) to vary across schools does not provide a statistically significant improvement in the model fit. Therefore, the simpler model (`rs_gend2`) is more appropriate for this dataset.

3.3.2.9 Adding a Level-2 Explanatory Variable to the Random Slope Model

A level-2 variable is one that varies between groups rather than within groups (students within a school). In the previous models, the level-1 variables are `cohort90` and `female2`, which vary within schools.

Now, the model is extended by including a level-2 variable that varies between schools. Here, `class2` (social class) is used as the level-2 explanatory variable. The extended model can be written as:

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + \beta_2 \text{female}_{ij} + \beta_3 \text{class2}_j + u_{0j} + u_{1j} \text{cohort90}_{ij} + u_{2j} \text{female}_{ij} + e_{ij}$$

> **Note**
>
> $\beta_3 \text{class2}_j$: The fixed effect of the level-2 variable `class2`, which is a characteristic of the school.

```
# Add the level-2 explanatory variable `class2` to the model.
rs_gend2_class <- lmer(score ~ cohort90 + female2 + class2 + (1 + cohort90 | schoolid), data
summary(rs_gend2_class)
```

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: score ~ cohort90 + female2 + class2 + (1 + cohort90 | schoolid)
   Data: score.sch
Control: lmerControl(optimizer = "bobyqa")

      AIC       BIC    logLik  deviance  df.resid
 276712.3  276796.6 -138346.1  276692.3     33978

Scaled residuals:
    Min      1Q  Median      3Q     Max
```

```
-3.4875 -0.6997  0.0304  0.7071  3.8091

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schoolid (Intercept)  22.5133  4.7448
          cohort90      0.1508  0.3884  -0.32
 Residual             192.9457 13.8905
Number of obs: 33988, groups:  schoolid, 508

Fixed effects:
                     Estimate Std. Error         df t value Pr(>|t|)
(Intercept)         3.570e+01  2.742e-01  7.089e+02  130.18   <2e-16 ***
cohort90            1.183e+00  2.431e-02  3.218e+02   48.65   <2e-16 ***
female2female       1.961e+00  1.543e-01  3.371e+04   12.71   <2e-16 ***
class2intermediate -5.210e+00  1.970e-01  3.372e+04  -26.45   <2e-16 ***
class2working      -1.109e+01  2.064e-01  3.382e+04  -53.71   <2e-16 ***
class2unclassified -1.482e+01  2.856e-01  3.384e+04  -51.90   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) chrt90 fml2fm clss2nt clss2w
cohort90    -0.194
female2feml -0.296 -0.023
clss2ntrmdt -0.354  0.036  0.001
class2wrkng -0.350  0.054 -0.008  0.489
clss2nclssf -0.260  0.003  0.007  0.348   0.366
```

```r
# For a nicer output:
tidy(rs_gend2_class, conf.int = TRUE) %>% kbl %>% kable_styling()
```

Interpretation:

1. **Random Effects**

   - **Intercept (schoolid)**: The variance in intercepts (22.5133) indicates substantial variability in starting scores between schools.
   - **cohort90 (schoolid)**: The variance in slopes (0.1508) indicates variability in how scores change over time across different schools.
   - **Residual Variance**: The residual variance (192.9457) represents the within-school variability not explained by the model.

effect	group	term	estimate	std.error	statistic	df	p.val
fixed	NA	(Intercept)	35.6955454	0.2742111	130.17543	708.8876	
fixed	NA	cohort90	1.1828314	0.0243149	48.64636	321.7982	
fixed	NA	female2female	1.9613418	0.1542812	12.71277	33710.3149	
fixed	NA	class2intermediate	-5.2104781	0.1969767	-26.45225	33719.7899	
fixed	NA	class2working	-11.0856769	0.2063932	-53.71146	33816.4649	
fixed	NA	class2unclassified	-14.8234167	0.2855945	-51.90371	33837.7167	
ran_pars	schoolid	sd___(Intercept)	4.7448211	NA	NA	NA	N
ran_pars	schoolid	cor___(Intercept).cohort90	-0.3169885	NA	NA	NA	N
ran_pars	schoolid	sd___cohort90	0.3883867	NA	NA	NA	N
ran_pars	Residual	sd___Observation	13.8904899	NA	NA	NA	N

- **Correlation**:A moderate negative correlation (-0.32) between intercepts and slopes for cohort90 suggests that schools with higher starting scores tend to have slower rates of score increase.

2. **Fixed Effects**

   - **(Intercept)**: The average starting score for a student is approximately 35.70 when all predictors are zero.
   - **cohort90**: On average, each additional year (cohort90) increases the score by 1.183 points, which is highly significant.
   - **female2female**: Female students, on average, score 1.961 points higher than male students, which is also highly significant.
   - **class2**:
     - Students from intermediate social class score 5.210 points lower on average compared to the baseline class.
     - Students from working social class score 11.09 points lower on average compared to the baseline class.
     - Students from unclassified social class score 14.82 points lower on average compared to the baseline class.

3. **Conclusion**: The model reveals significant effects of `cohort90`, `female2`, and `class2` on scores. The random effects show variability across schools in both starting scores and the rate of change over time. The negative correlation between intercepts and slopes indicates that schools with higher initial scores tend to improve more slowly over time. The fixed effects indicate that social class has a substantial impact on student scores, with lower social classes associated with lower scores.

3.3.2.10 Cross-Level Interaction in the Random Slope Model

A cross-level interaction involves an interaction between a level-1 variable (within-group) and a level-2 variable (between-group). This type of interaction allows the effect of a level-1 variable to vary depending on the value of a level-2 variable.

```
# cross-level interactions
m.int <- lmer(score ~ cohort90 + female2 + class2 + schtype + schurban + cohort90:schtype +
summary(m.int)
```

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula:
score ~ cohort90 + female2 + class2 + schtype + schurban + cohort90:schtype +
    (1 + cohort90 | schoolid)
   Data: score.sch
Control: lmerControl(optimizer = "bobyqa")

      AIC       BIC    logLik  deviance  df.resid
 276651.0  276760.7 -138312.5  276625.0     33975

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4880 -0.7014  0.0296  0.7079  3.6237

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schoolid (Intercept)  20.414   4.5182
          cohort90      0.138   0.3715  -0.23
 Residual             192.851  13.8871
Number of obs: 33988, groups:  schoolid, 508

Fixed effects:
                     Estimate Std. Error         df t value Pr(>|t|)
(Intercept)         3.621e+01  4.301e-01  4.637e+02  84.182  < 2e-16 ***
cohort90            1.214e+00  2.442e-02  3.223e+02  49.687  < 2e-16 ***
female2female       1.970e+00  1.542e-01  3.373e+04  12.779  < 2e-16 ***
class2intermediate -5.189e+00  1.971e-01  3.366e+04 -26.323  < 2e-16 ***
class2working      -1.102e+01  2.069e-01  3.384e+04 -53.268  < 2e-16 ***
class2unclassified -1.476e+01  2.858e-01  3.384e+04 -51.643  < 2e-16 ***
schtype             5.291e+00  8.307e-01  5.242e+02   6.369 4.16e-10 ***
schurban           -1.404e+00  4.829e-01  3.523e+02  -2.907  0.00388 **
cohort90:schtype   -5.994e-01  1.038e-01  7.138e+02  -5.775 1.15e-08 ***
```

effect	group	term	estimate	std.error	statistic	df	p
fixed	NA	(Intercept)	36.2064860	0.4301002	84.181518	463.6748	0.00
fixed	NA	cohort90	1.2135295	0.0244234	49.687149	322.2759	0.00
fixed	NA	female2female	1.9702553	0.1541846	12.778548	33730.0261	0.00
fixed	NA	class2intermediate	-5.1886524	0.1971121	-26.323357	33658.1879	0.00
fixed	NA	class2working	-11.0194081	0.2068684	-53.267723	33836.7967	0.00
fixed	NA	class2unclassified	-14.7622456	0.2858496	-51.643407	33843.3703	0.00
fixed	NA	schtype	5.2907883	0.8307005	6.369068	524.1785	0.00
fixed	NA	schurban	-1.4037690	0.4828684	-2.907146	352.2789	0.00
fixed	NA	cohort90:schtype	-0.5994210	0.1037948	-5.775060	713.8247	0.00
ran_pars	schoolid	sd___(Intercept)	4.5181675	NA	NA	NA	
ran_pars	schoolid	cor___(Intercept).cohort90	-0.2327578	NA	NA	NA	
ran_pars	schoolid	sd___cohort90	0.3715041	NA	NA	NA	
ran_pars	Residual	sd___Observation	13.8870913	NA	NA	NA	

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

Correlation of Fixed Effects:
            (Intr) chrt90 fml2fm clss2nt clss2w clss2nc schtyp schrbn
cohort90    -0.118
female2feml -0.186 -0.022
clss2ntrmdt -0.244  0.035  0.001
class2wrkng -0.231  0.056 -0.008  0.490
clss2nclssf -0.160  0.004  0.008  0.349   0.369
schtype     -0.139  0.043  0.002  0.045   0.080  0.053
schurban    -0.765  0.038 -0.004  0.015  -0.007 -0.019  -0.050
chrt90:scht  0.032 -0.235 -0.005  0.004  -0.009 -0.004  -0.175 -0.014
```

```
# For a nicer output:
tidy(m.int, conf.int = TRUE) %>% kbl %>% kable_styling()
```

Components of Model:

1. **Fixed Effects**:
    - `cohort90`: A level-1 predictor variable representing the cohort year.
    - `female2`: A level-1 predictor variable representing gender.
    - `factor(sclass)`: A level-2 predictor variable representing social class.
    - `schtype`: A level-2 predictor variable representing school type.
    - `schurban`: A level-2 predictor variable representing whether the school is urban.

- `cohort90:schtype`: The cross-level interaction term between cohort year and school type.

2. **Random Effects** `(1 + cohort90 | schoolid)`: Random intercepts and random slopes for cohort90 within each school (`schoolid`).

Interpretation:

1. **Random Effects**

   - **Intercept (schoolid)**: The variance in intercepts (20.414) indicates substantial variability in starting scores between schools.
   - **cohort90 (schoolid)**: The variance in slopes (0.138) indicates variability in how scores change over time across different schools.
   - **Residual Variance**: The residual variance (192.851) represents the within-school variability not explained by the model.
   - **Correlation**: A moderate negative correlation (-0.23) between intercepts and slopes for cohort90 suggests that schools with higher starting scores tend to have slower rates of score increase.

2. **Fixed Effects**

   - **(Intercept)**: The average starting score for a student is approximately 36.21 when all predictors are zero.
   - **cohort90**: On average, each additional year (cohort90) increases the score by 1.214 points, which is highly significant.
   - **female2female**: Female students, on average, score 1.970 points higher than male students, which is also highly significant.
   - **class2**:
     - Students from intermediate social class score 5.189 points lower on average compared to the baseline class.
     - Students from working social class score 11.02 points lower on average compared to the baseline class.
     - Students from unclassified social class score 14.76 points lower on average compared to the baseline class.
   - **schtype**: Schools of a particular type have scores that are 5.291 points higher on average, which is highly significant.
   - **schurban**: Students in urban schools score 1.404 points lower on average, which is significant.

- **cohort90:schtype**: The interaction term indicates that the effect of cohort year on score is modified by the school type. Specifically, for each additional year, the score increases by 1.214 points, but this increase is reduced by 0.5994 points in certain school types, which is highly significant.

3. **Conclusion**: The model reveals significant effects of `cohort90`, `female2`, `class2`, `schtype`, and `schurban` on scores. The random effects show variability across schools in both starting scores and the rate of change over time. The interaction between `cohort90` and `schtype` indicates that the effect of cohort year on scores varies depending on the school type. The fixed effects indicate that social class, school type, and whether the school is urban significantly impact student scores.

3.3.2.11 Checking Assumptions

Checking assumptions is a critical step in validating the results of any statistical model. The objective is to validate that the model fits the data well and that the assumptions underlying the model are not violated.

1. **Residual Analysis**: Augment the model to get residuals and fitted values for further analysis.

```
res.rs.gend2.slope <- augment(rs_gend2_slope)
datatable(res.rs.gend2.slope)
```

```
Warning in instance$preRenderHook(instance): It seems your data is too big for
client-side DataTables. You may consider server-side processing:
https://rstudio.github.io/DT/server.html
```

Show 10 entries

Search:

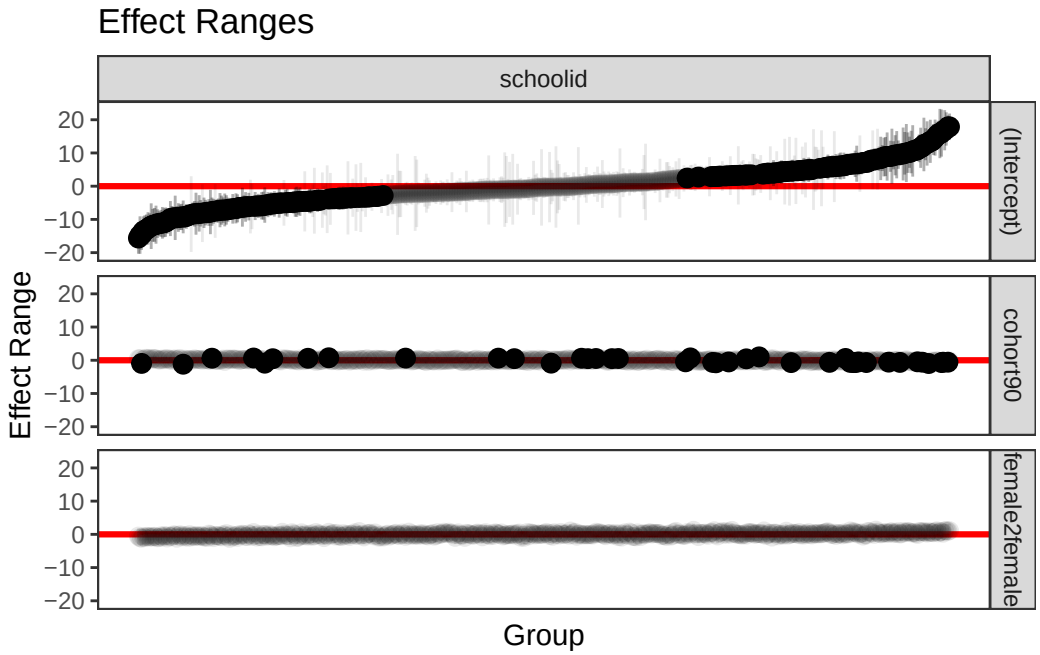
	score	cohort90	female2	schoolid	.fitted	.resid	.hat	.cooksd	.fixed	.mu	.offset	.sqrtXwt	.sqrtrwt	.weights	.wtres
1	0	-6	female	1	16.62070581680562	-16.62070581680562	0.02756855877852723	0.01251453020647271	24.1539094646354	16.62070581680562	0	1	1	1	-16.62070581680562
2	10	-6	female	1	16.62070581680562	-6.620705816805621	0.02756855877852723	0.001985748337055262	24.1539094646354	16.62070581680562	0	1	1	1	-6.620705816805621
3	0	-6	female	1	16.62070581680562	-16.62070581680562	0.02756855877852723	0.01251453020647271	24.1539094646354	16.62070581680562	0	1	1	1	-16.62070581680562
4	40	-6	female	1	16.62070581680562	23.37929418319438	0.02756855877852723	0.02476158366031369	24.1539094646354	16.62070581680562	0	1	1	1	23.37929418319438
5	42	-6	female	1	16.62070581680562	25.37929418319438	0.02756855877852723	0.02917928899821149	24.1539094646354	16.62070581680562	0	1	1	1	25.37929418319438
6	4	-6	female	1	16.62070581680562	-12.62070581680562	0.02756855877852723	0.007215773840075516	24.1539094646354	16.62070581680562	0	1	1	1	-12.62070581680562
7	0	-6	female	1	16.62070581680562	-16.62070581680562	0.02756855877852723	0.01251453020647271	24.1539094646354	16.62070581680562	0	1	1	1	-16.62070581680562
8	0	-6	female	1	16.62070581680562	-16.62070581680562	0.02756855877852723	0.01251453020647271	24.1539094646354	16.62070581680562	0	1	1	1	-16.62070581680562
9	14	-6	female	1	16.62070581680562	-2.620705816805621	0.02756855877852723	0.0003111373660921191	24.1539094646354	16.62070581680562	0	1	1	1	-2.620705816805621
10	27	-6	female	1	16.62070581680562	10.37929418319438	0.02756855877852723	0.004880353365348472	24.1539094646354	16.62070581680562	0	1	1	1	10.37929418319438

Showing 1 to 10 of 33,988 entries

Previous 1 2 3 4 5 … 3,399 Next

2. **Plotting Random Effects Plot 1**: `REsim` function generates simulated random effects, providing a distribution of these effects. **`plotREsim`** visualizes these distributions to check if they are approximately normal and to assess the variability between groups.

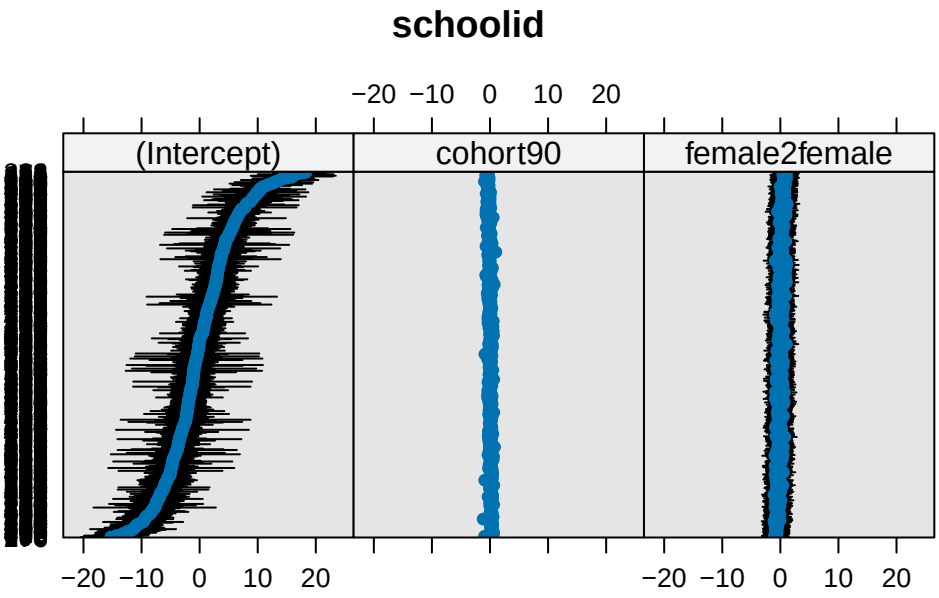
```
# Using the `merTools` package to visualize the random effects
re.rs.gend2.slope <- REsim(rs_gend2_slope)
plotREsim(re.rs.gend2.slope)
```



Plot 2: The `dotplot` of random effects helps visualize the distribution and variability of the random effects. Consistency in the spread of dots across the plot indicates that the random effects are well-modeled.

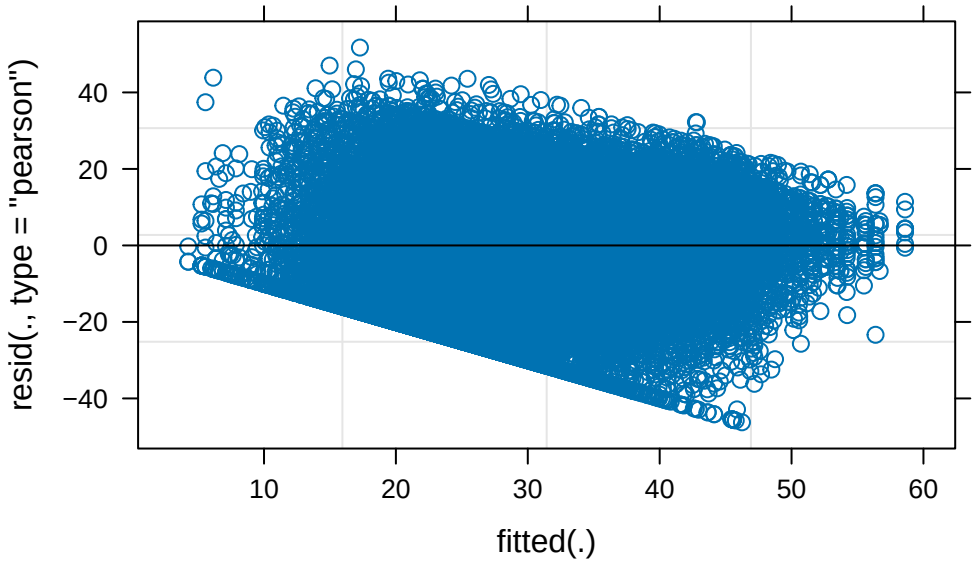
```
# Using the `lattice` package to create a dot plot of the random effects
randoms <- ranef(rs_gend2_slope, condVar = TRUE)
dotplot(randoms)
```

```
$schoolid
```



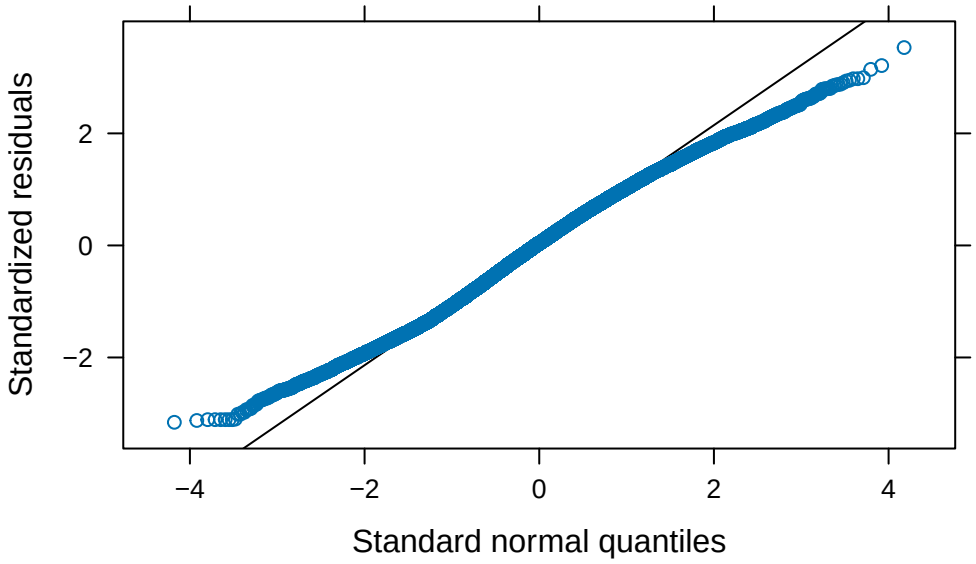
3. **Fitted vs Residuals Plot**: This plot helps check the linearity assumption and the homogeneity of variances. A random scatter of residuals around zero without a clear pattern indicates that these assumptions are likely met.

```
# Plot fitted values against residuals to check for patterns
plot(rs_gend2_slope)
```



4. **Normality of Residuals**: The Q-Q plot checks if the residuals are normally distributed. Points that closely follow the reference line in the plot suggest that the residuals are normally distributed.

```
# Create a Q-Q plot for checking the normality of residuals
qqmath(rs_gend2_slope)
```



4 Acknowledgements

I would like to extend my sincere gratitude to Professor Dr. Kamarul Imran Musa, Medical Epidemiologist and Statistician, and Professor in Epidemiology and Biostatistics at the School of Medical Sciences, Universiti Sains Malaysia. His guidance and teaching in statistical analysis and R software have been invaluable in the development of this analysis.

5 References

- Introduction to Multilevel Modelling
- UCLA: Repeated Measures Analysis with R
- More Advanced Guide: Longitudinal Data Analysis using lme4
- Bolker's Guide to Mixed Models

6 R Codes

The R codes used in this analysis are available at the following GitHub repository: DrPH-Epidemiology-Revision. This repository includes all scripts and data necessary to replicate the analyses presented in this work.