# Multilevel Model for General mathematics

Brendi Ang

17/10/2021

## Contents

# General Mathematics

## Exploring the dataset with basic linear model

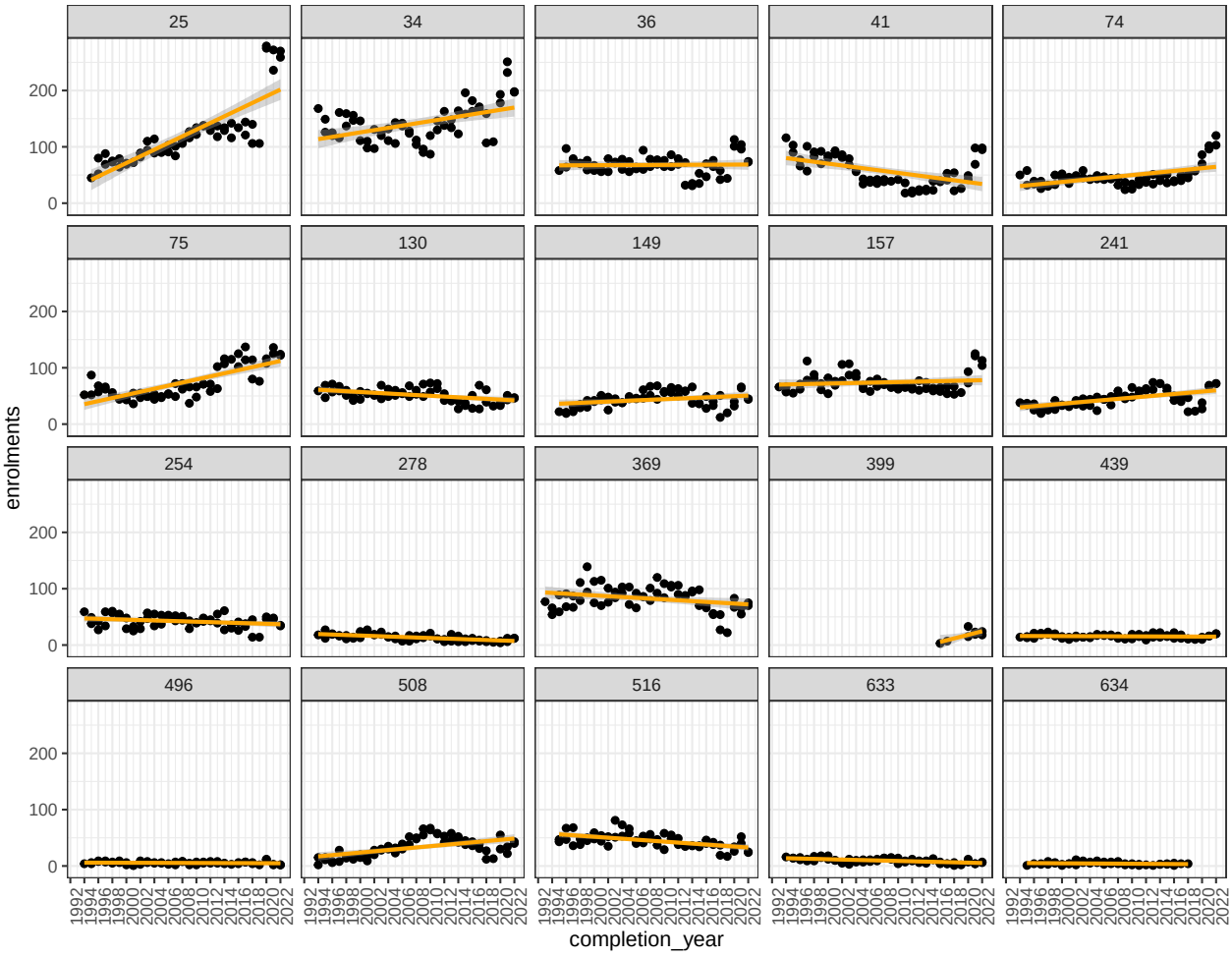


Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for General Mathematics subject

With reference to the first step, Figure 1 fits a linear model for enrolments for a random sample of 20 schools. Various school sizes can be seen, school 633 and 634 (bottom-right) appears to have approximately 10 enrolments per cohort, while school 25 and 34 have enrolments above 200 per cohort. Some schools showed relatively large increase in enrolments over the years, while some showed a decrease (*e.g.* school 41).

## Getting the data ready for modelling

### Removing zero enrolments

All zero enrolments in a given year will be removed for modelling. As aforementioned, most of the zero enrolments in year 11 (refer to Figure **??**) were attributed to the 2007 prep year cohort while zero enrolments in year 12 relates to the first year in which a school introduces the subject. Other zero enrolments mostly relates to smaller schools with little to no enrolments in the subject for a given year. These zero enrolments will be removed for modelling purposes.

### Linearise response variable using log transformation

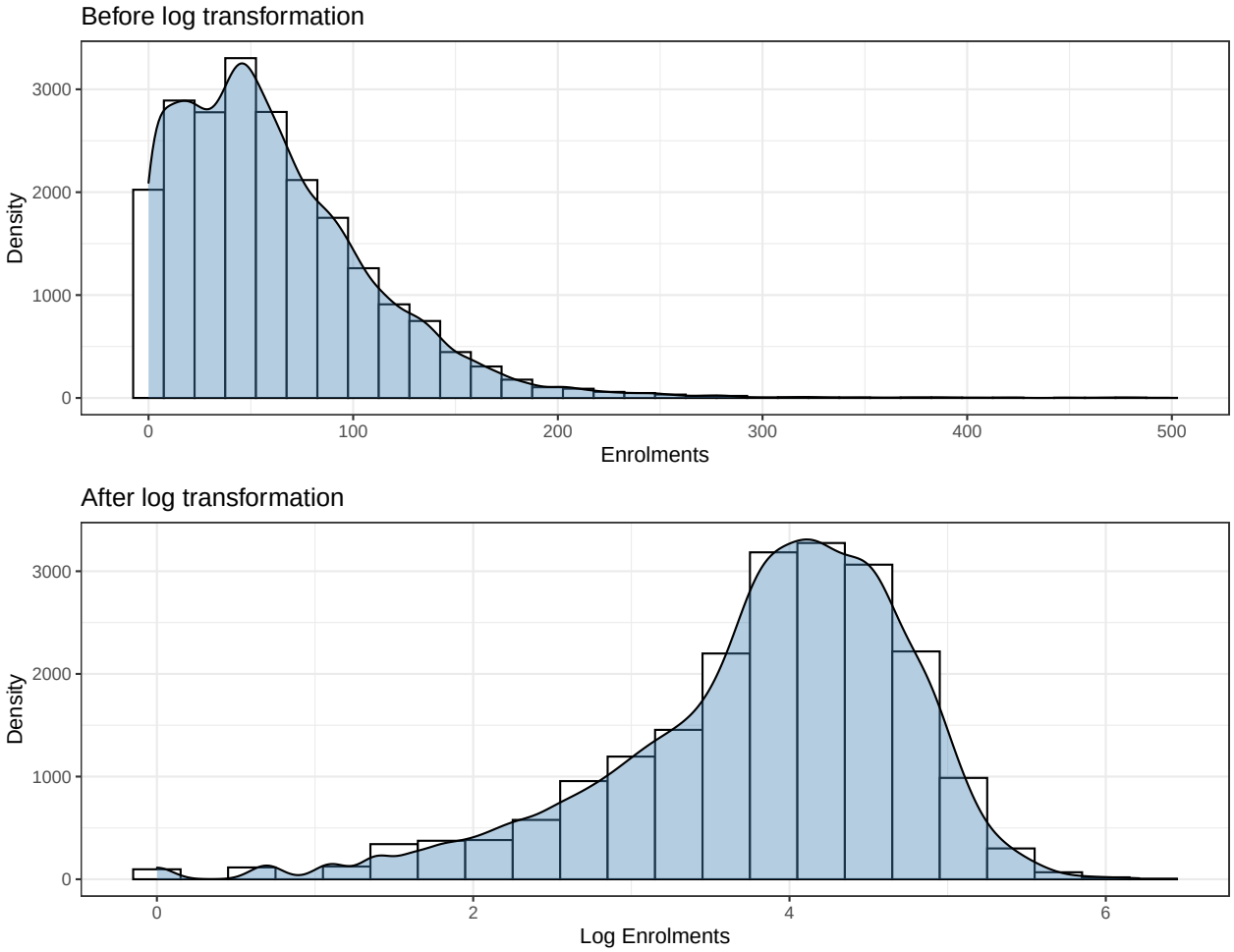


Figure 2: Effects of log transformation for response variable (enrolments) in General Mathematics

The enrolments were right skewed, which is likely to be attributed to the various school sizes (as seen in Figure 2). A log transformation was implemented to the response variable (*i.e.* `enrolments`) to allow the the multilevel model to better capture the enrolment patterns.

## Unconditional means model

Table 1: AIC values for all candidate models for General Mathematics

| | df | AIC |
|---|---|---|
| Model0.2: Schools nested within districts | 4 | 24964.05 |
| Model0.0: Within schools | 3 | 24979.49 |
| Model0.1: Schools nested within postcodes | 4 | 24981.49 |

As per Step 3, the three potential models are fitted, with the AIC shown in Table 1. Based on the AIC, `model0.2`, corresponding the schools nested within districts is the best model and will be used in the subsequent analysis.

### Intraclass correlation ($ICC$)

```
## Random effects:

##  Groups                         Name        Variance Std.Dev.
##  qcaa_school_id:qcaa_district (Intercept) 0.891424 0.94415
##  qcaa_district                (Intercept) 0.069765 0.26413
##  Residual                                 0.174806 0.41810

##
##  Fixed effects:

##             Estimate Std. Error  t value
## (Intercept)  3.68386 0.08530786 43.18313

##
##  Number of schools (level-two group) = 481
##  Number of district (level-three group) = 13
```

This model will takes into account 481 schools nested in 13 districts. In a three-level multilevel model, two intraclass correlations can be obtained using the model summary output above:

The **level-two ICC** relates to the correlation between school $i$ from district $k$ in time $t$ and in time $t^* \neq t$:

$$ICC(school) = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.89142}{(0.89142 + 0.06976 + 0.17481)} = 0.7847$$

This can be conceptualised as the correlation between enrolments of two random draws from the same school at two different years. In other words, 78.47% of the total variability is attributable to the changes overtime within schools.

The **level-three ICC** refers to the correlation between different schools $i$ and $i^*$ from a specific district $j$ in time $t$ and time $t^* \neq t$.

$$ICC(school) = \frac{\phi_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.1077}{(0.5888 + 0.1077 + 0.2094)} = 0.0614$$

Similarly, this can be conceptualised as the correlation between enrolments of two randomly selected schools from the same district – *i.e.* 6.14% of the total variability is due to the difference between districts.

## Unconditional Growth model

The unconditional growth model introduces the time predictor at level one, the model specification can be found in step 4. This allows for assessing within-school variability which can be attributed to linear changes over time. Furthermore, variability in intercepts and slopes can be obtained to compare schools within the same districts, and schools from different districts.

```
##  Groups                        Name        Variance   Std.Dev.  Corr
##  qcaa_district:qcaa_school_id (Intercept) 3.7608e+00 1.9392756
##                               year92      3.2360e-03 0.0568856 -0.869
##  qcaa_district                (Intercept) 1.7958e-02 0.1340077
##                               year92      5.8655e-05 0.0076587 1.000
##  Residual                                 9.8464e-02 0.3137898
```

```
##               Estimate  Std. Error  t value
## (Intercept) 2.91024486 0.098867525 29.43580
## year92      0.03476855 0.003475268 10.00457
```

```
##  Number of Level Two groups =  481
##  Number of Level Three groups =  13
```

- $\pi_{0ij} = 2.9102$: Initial status for school $i$ in district $j$ (*i.e.* expected log enrolments when time = 0)
- $\pi_{1ij} = 0.0348$: Growth rate for school $i$ in district $j$
- $\epsilon_{tij} = 0.0984$: Variance in within-school residuals after accounting for linear growth overtime

When the subject was first introduced in 1992, schools were expected to have 18.3613 ($e^{2.9102449}$) enrolments, on average. On average, enrolments were expected to increase by 3.5380% ($(e^{0.0347686} - 1) \times 100$) per year. The estimated within-schools variance decreased by 90.16% (0.17481 to 0.0984), implying that 90.16% of within-school variability can be explained by the linear growth over time.

## Dealing with boundary issues

A singular fit is observed in the model as the correlation between the intercept and slope between districts are perfectly correlation (*i.e.* $\phi_{01} = 1$). This may suggest that the model is overfitted – *i.e.* the random

effects structure is too complex to be supported by the data and may require some re-parameterisation. Naturally, the higher-order random effects (*e.g.* random slope of the third level (between district)) can be removed, especially where the variance and correlation terms are estimated on the boundaries (*add bookdown reference*).

```
##  Groups                          Name        Variance  Std.Dev. Corr
##  qcaa_district:qcaa_school_id (Intercept) 3.8040956 1.950409
##                                  year92      0.0033141 0.057568 -0.870
##  qcaa_district                   (Intercept) 0.1242578 0.352502
##  Residual                                    0.0984432 0.313757
```

```
##              Estimate  Std. Error  t value
## (Intercept) 2.8959781 0.134506478 21.53040
## year92      0.0353323 0.002779433 12.71205
```

```
##  Number of Level Two groups =  481
##  Number of Level Three groups =  13
```

To elaborate, two parameters were removed by setting variance components $\phi^2_{10} = \phi_{01}$ equal to zero Which indirectly assumes that the growth rate for district $j$ to be fixed. As shown in the model output above, this produced a more stable model and is free from any boundary constraints. As compared to the unconditional growth model (`model1.0`), the fixed effects remained rather similar.

Level one and level two will be identical to the unconditional growth model (`model1.0`), however, the random slope for level 3 will be removed. This implies that the error assumption at level three now follows a univariate normal distribution where $r_{00j} \sim N(0, \phi^2_{00})$.

The new Level three (districts):

$$\beta_{00j} = \gamma_{000} + r_{00j}\beta_{10j} = \gamma_{100}$$

And therefore composite model:

$$\begin{aligned} Y_{tij} &= \pi_{0ij} + \pi_{1ij} year92_{tij} + \epsilon_{tij} \\ &= (\beta_{00j} + u_{0ij}) + (beta_{10j} + u_{1ij}) year92_{tij} + \epsilon_{tij} \\ &= (\gamma_{000} + r_{00j} + u_{0ij}) + (\gamma_{100} + u_{1ij}) year92_{tij} + \epsilon_{tij} \\ &= [\gamma_{000} + \gamma_{100} year92_{tij}] + [r_{00j} + u_{0ij} + u_{1ij} year92_{tij} + \epsilon_{tij}] \end{aligned}$$

## Testing fixed effects

Table 2: AIC for all possible models with different combinations of fixed effects

| model | AIC |
|---|---|
| model4.0 | 15044.37 |
| model4.1 | 15051.92 |
| model4.7 | 15083.63 |
| model4.5 | 15133.73 |
| model4.6 | 15143.43 |
| model4.4 | 15173.32 |
| model4.9 | 15175.07 |
| model4.2 | 15175.07 |
| model4.8 | 15175.07 |
| model4.10 | 15175.07 |
| model4.3 | 15226.17 |

As highlighted in step 6, `sector` and `unit` will be added as predictors to the model. The largest possible model will be fitted, before removing fixed effects one by one while recording the AIC for each model. In this case, `model4.0` corresponds to the largest possible model while `model4.10` is the smallest possible model. The model with the optimal (lowest) AIC is `model4.4` (Table 2). The next section will test the selected model's random effects to build the final model.

## Parametric bootstrap to test random effects

This step will not be undertaken, as the random slope will not be included at level three as a boundary constraint was found in the unconditional growth model, indicating that the model will be overfitted if random slopes were included at level three.

## Confidence interval

Table 3: 95% confidence intervals for fixed and random effects in the final model

| var | 2.5 % | 97.5 % |
|---|---|---|
| sd_(Intercept)\|qcaa_district:qcaa_school_id | 1.5242174 | 1.7454448 |
| cor_year92.(Intercept)\|qcaa_district:qcaa_school_id | -0.8618092 | -0.8004080 |
| sd_year92\|qcaa_district:qcaa_school_id | 0.0460150 | 0.0533202 |
| sd_(Intercept)\|qcaa_district | 0.1979323 | 0.5344465 |

| var | 2.5 % | 97.5 % |
|---|---|---|
| sigma | 0.3098662 | 0.3158212 |
| (Intercept) | 2.6985826 | 3.5356257 |
| year92 | 0.0288684 | 0.0511286 |
| sectorGovernment | 0.2532983 | 1.0794554 |
| sectorIndependent | -2.1037594 | -1.1831583 |
| unityear_12_enrolments | -0.0161797 | 0.0802241 |
| year92:sectorGovernment | -0.0434640 | -0.0178734 |
| year92:sectorIndependent | 0.0139084 | 0.0424983 |
| year92:unityear_12_enrolments | -0.0018883 | 0.0030005 |
| sectorGovernment:unityear_12_enrolments | -0.2128654 | -0.0980218 |
| sectorIndependent:unityear_12_enrolments | -0.0886348 | 0.0386853 |
| year92:sectorGovernment:unityear_12_enrolments | 0.0011711 | 0.0070114 |
| year92:sectorIndependent:unityear_12_enrolments | -0.0021703 | 0.0038465 |

The parametric bootstrap is utilised to construct confidence intervals (detailed explanation in step 8) for the random effects. If the confidence intervals between the random effects does not include 0, it provides statistical evidence that the p-value is less than 0.5. In other words, it suggests that the random effects and the correlation between the random effects are significant at the 5% level. The confidence interval for the random effects all exclude 0 (Table 3), indicating that they're different from 0 in the population (*i.e.* statistically significant).

## Interpreting final model

### Composite model

- Level one (measurement variable)

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} year92_{tij} + \epsilon_{tij}$$

- Level two (schools within districts) will contain new predictor(`sector`)

$$\pi_{0ij} = \beta_{00j} + \beta_{01j} sector_{ij} + \beta_{02j} unit_{ij} + \beta_{03j} sector_{ij} unit_{ij} + u_{0ij}$$
$$\pi_{1ij} = \beta_{10j} + \beta_{11j} sector_{ij} + \beta_{12j} unit_{ij} + \beta_{13j} sector_{ij} unit_{ij} + u_{1ij}$$

- Level three (districts)

$$\beta_{00j} = \gamma_{000} + r_{00j}$$
$$\beta_{01j} = \gamma_{010} + r_{01j}$$
$$\beta_{02j} = \gamma_{020} + r_{02j}$$
$$\beta_{03j} = \gamma_{030} + r_{03j}$$
$$\beta_{10j} = \gamma_{100}$$
$$\beta_{11j} = \gamma_{110}$$
$$\beta_{12j} = \gamma_{120}$$
$$\beta_{13j} = \gamma_{130}$$

The composite model can therefore be written as:

$$\begin{aligned}
Y_{tij} &= \pi_{0ij} + \pi_{1ij} year92_{tij} + \epsilon_{tij} \\
&= (\beta_{00j} + \beta_{01j} sector_{ij} + \beta_{02j} unit_{ij} + \beta_{03j} sector_{ij} unit_{ij} + u_{0ij}) + (\beta_{10j} + \beta_{11j} sector_{ij} + \beta_{12j} unit_{ij} + \beta_{13j} sector_{ij} unit_{ij} + \\
&= [\gamma_{000} + r_{00j} + (\gamma_{010} + r_{01j}) sector_{ij} + (\gamma_{020} + r_{02j}) unit_{ij} + (\gamma_{030} + r_{03j}) sector_{ij} unit_{ij} + u_{0ij}] + \\
&\quad [\gamma_{100} + \gamma_{110} sector_{ij} + \gamma_{120} unit_{ij} + \gamma_{130} sector_{ij} unit_{ij} + u_{1ij}]\, year92_{tij} + \epsilon_{tij} \\
&= [\gamma_{000} + \gamma_{010} sector_{ij} + \gamma_{020} unit_{ij} + \gamma_{030} sector_{ij} unit_{ij} + \gamma_{100} year92_{tij} + \gamma_{110} year92_{tij} sector_{ij} + \gamma_{120} unit_{ij} year92_{tij} + \\
&\quad \left[r_{00j} + r_{01j} sector_{ij} + r_{02j} unit_{ij} + r_{03j} sector_{ij} unit_{ij} + u_{0ij]} + u_{1ij} year92_{tij} + \epsilon_{tij}\right]
\end{aligned}$$

**Fixed effects**

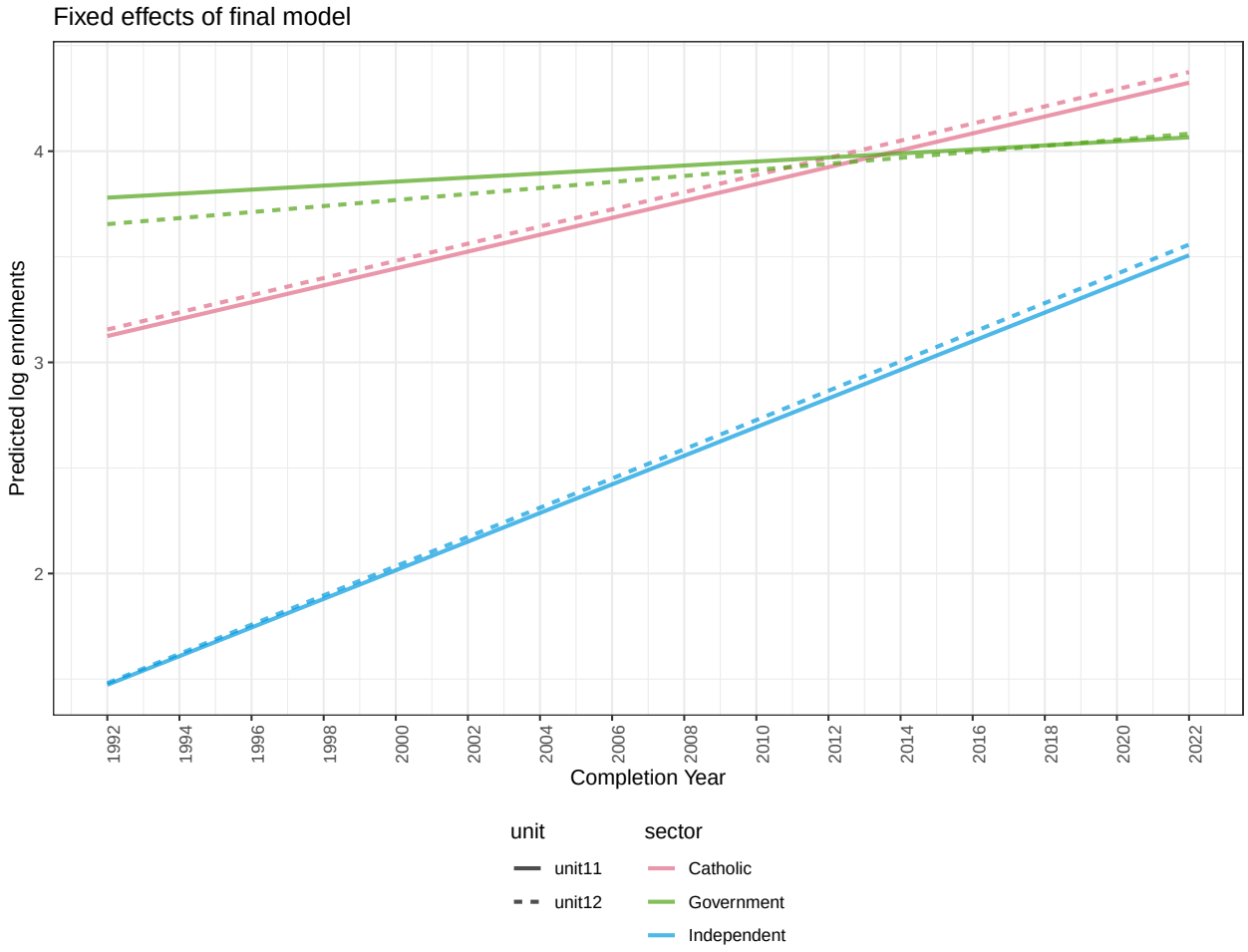


Figure 3: Fixed effects of the final model for General Mathematics subject

With a three-way interaction, it is easier to visualise the fixed effects (Figure 3). When the subject was first introduced in 1992, independent schools appears to have the least enrolments, on average. This low enrolment number was matched with a relatively large increase in enrolments per year, as seen by the slope. Although government schools have high enrolments initially, the rate of change in enrolments over the years increased relatively slow compared to the other sectors. Year 11 and year 12 units have similar enrolment numbers, on average.

**Random effects**

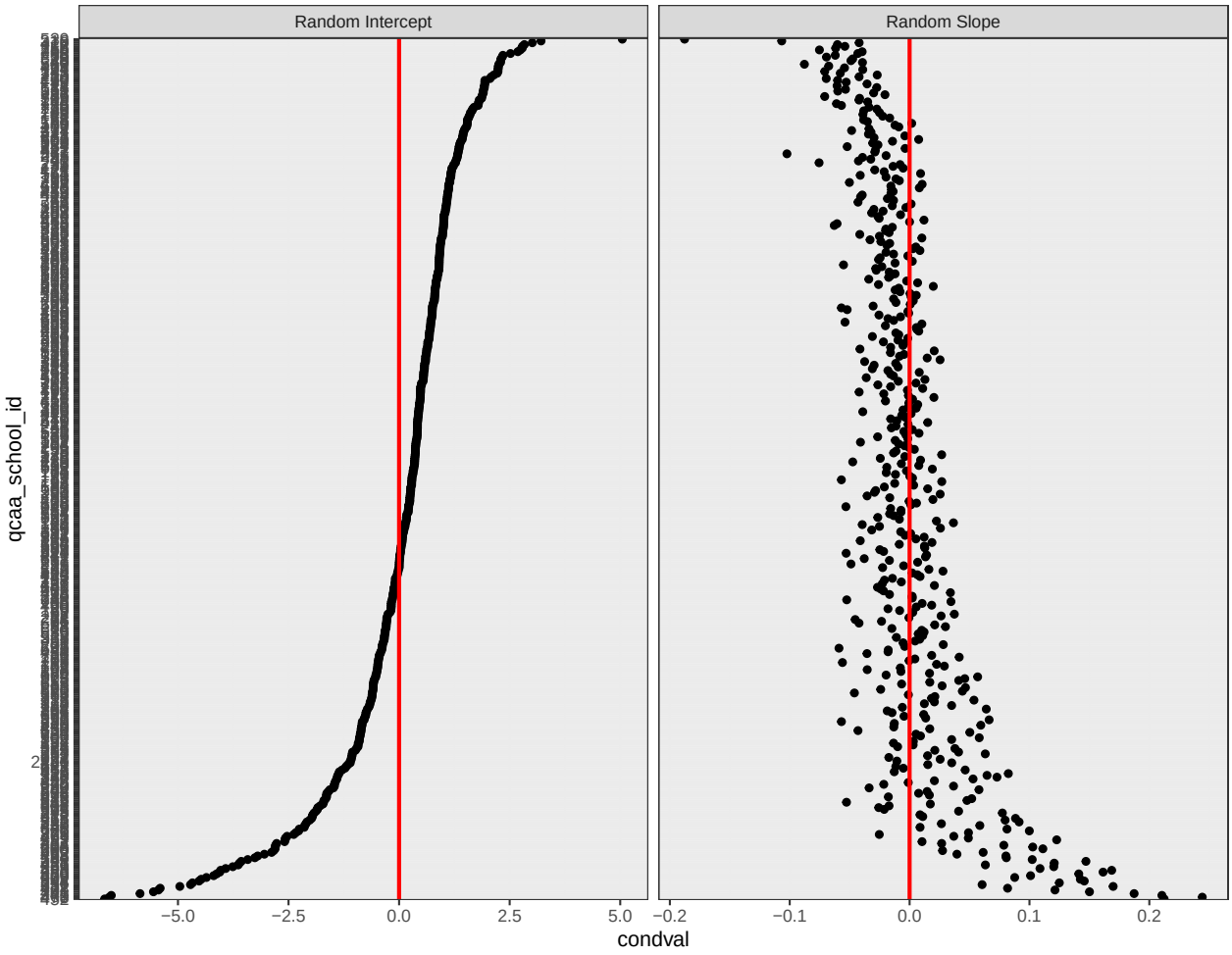


Figure 4: Random effects for all schools

A large negative correlation between the random intercept and slope at the school level is apparent (Figure 4. This suggests that a larger school is associated with a smaller increase (decrease) in enrolments over the years while smaller schools are predicted to have large increase in enrolments over the years.

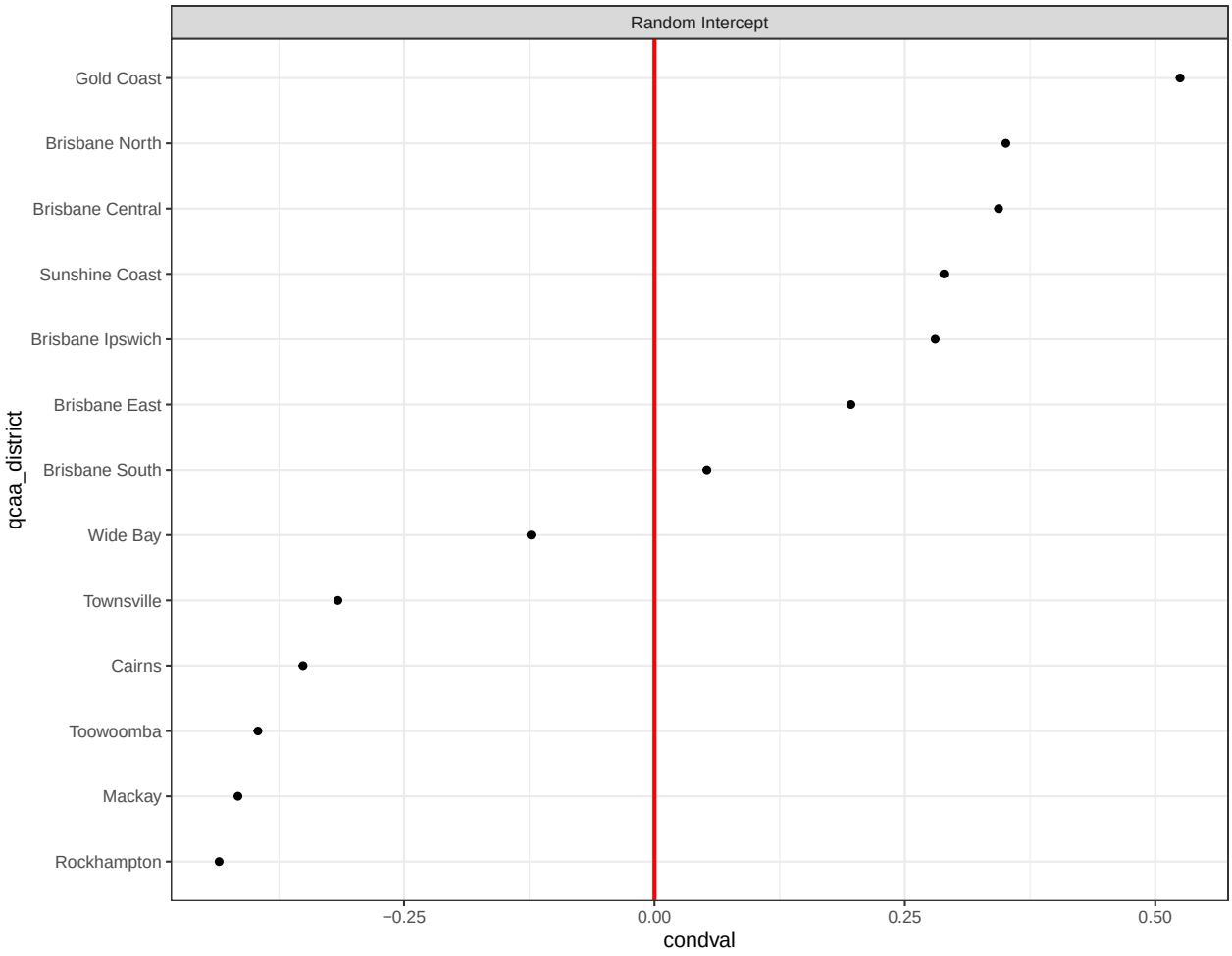


Figure 5: Random intercept for districts

As the random slopes are removed, all districts are predicted to have the same increase in enrolments over the years; And as was discussed previously, this was a reasonable assumption or an otherwise perfect correlation with random slope and intercept will be fitted. Figure 5 demonstrates that schools in Gold Coast has the largest enrolments, on average.

**Predictions**

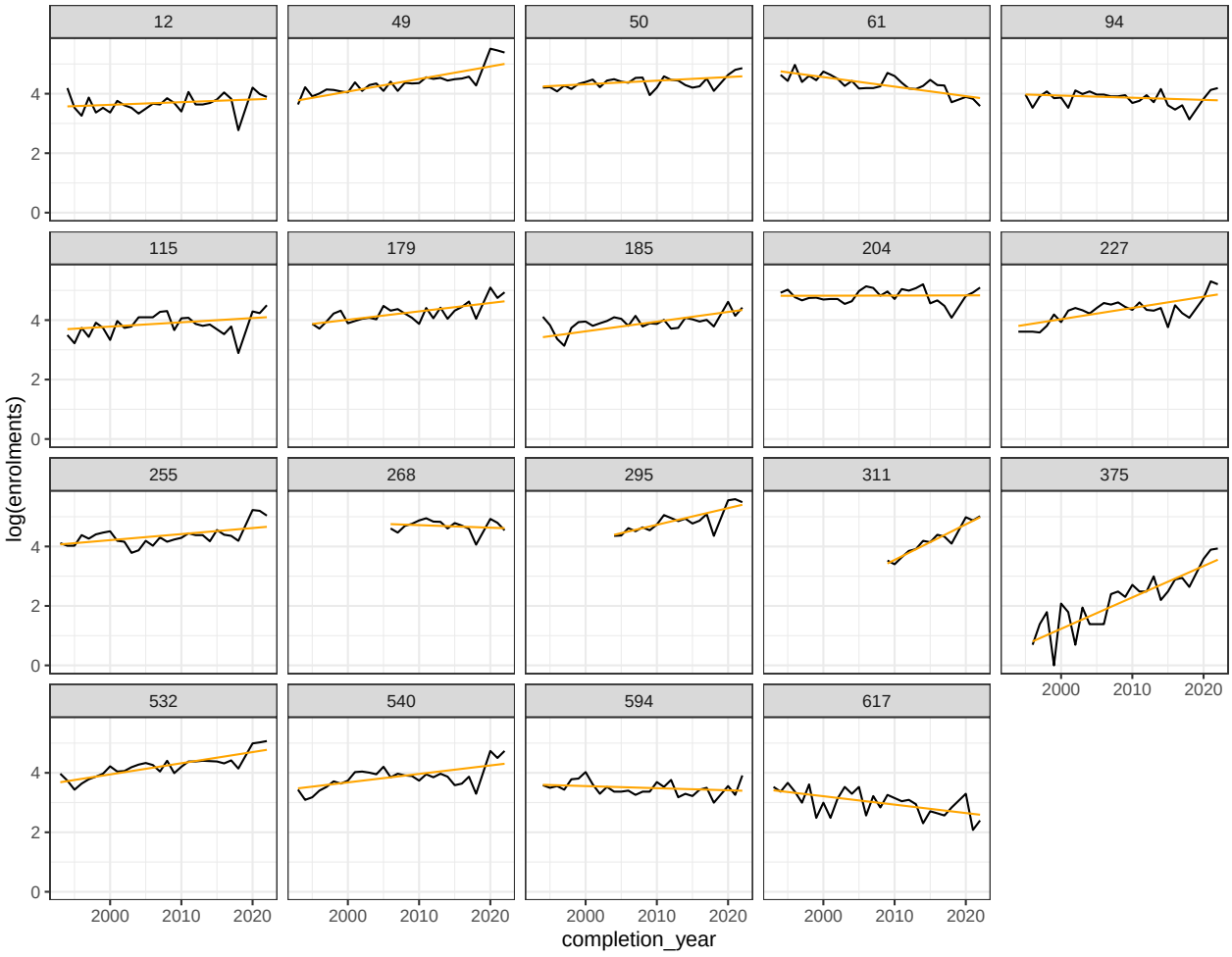


Figure 6: Model predictions for year 11 enrolments for 20 randomly selected schools

Figure 6 above shows the predictions for 20 randomly selected schools.