Multilevel Model for Specialist Mathematics

Brendi Ang

17/10/2021

Contents

Specialist Mathematics

Exploring the dataset with basic linear model

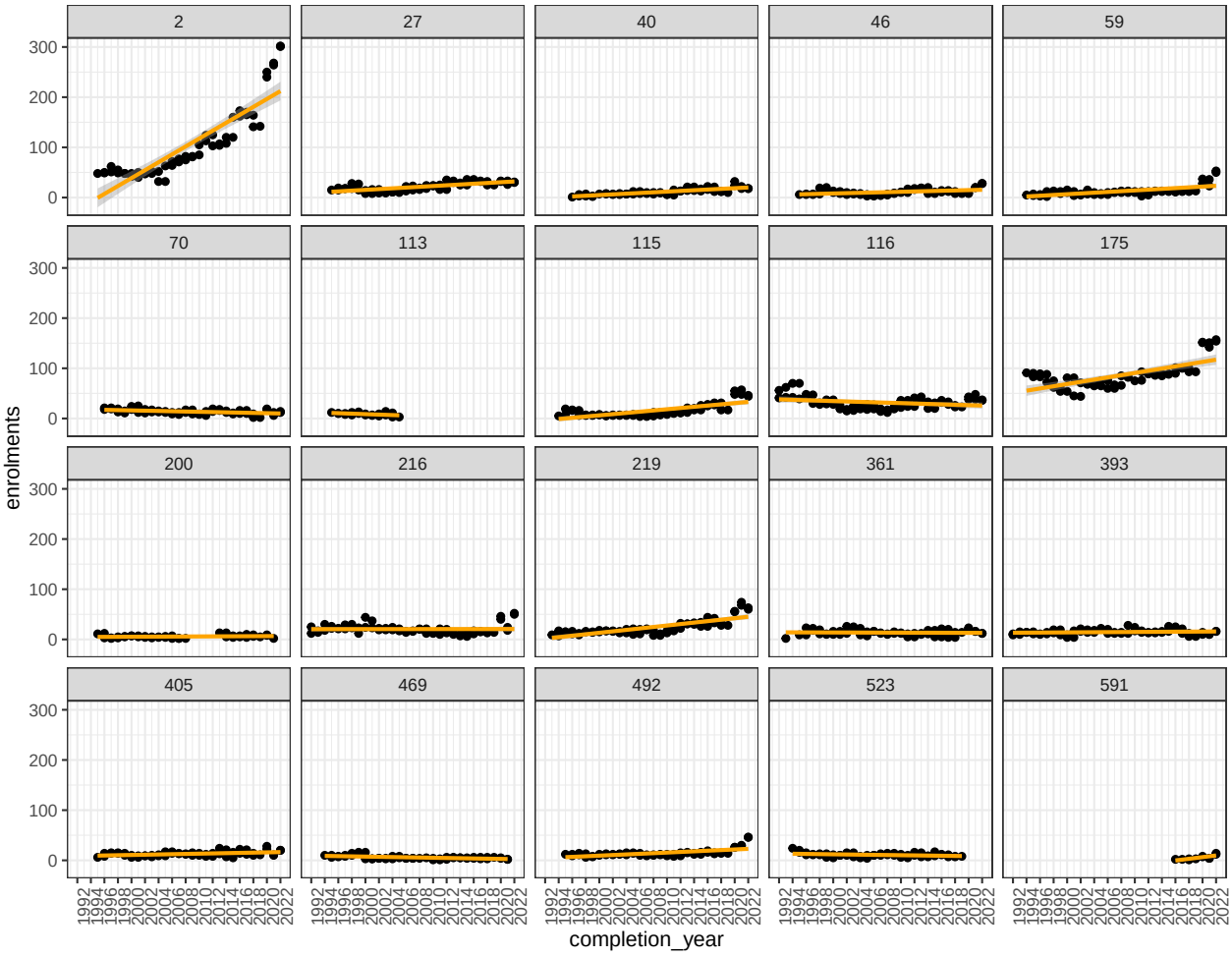

Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Specialist Mathematics subject

With reference to the first step, Figure 1 fits a linear model for enrolments for a random sample of 20 schools. A myriad of patterns can be obtained from this. For instance, enrolments in school 2 (top left) appears to have significant larger increase in enrolments over the years the specialist mathematics have been introduced. Most schools showed relatively small enrolment numbers of less than 50, however school 2 and 175 showed rather large enrolment numbers for each cohort. It also demonstrates that each school can introduce (or end) specialist mathematics at different years – *e.g.* School 591 only introduced the subject in 2016, and School 113 introduced the subject in 1995 and discontinued the subject in 2005.

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

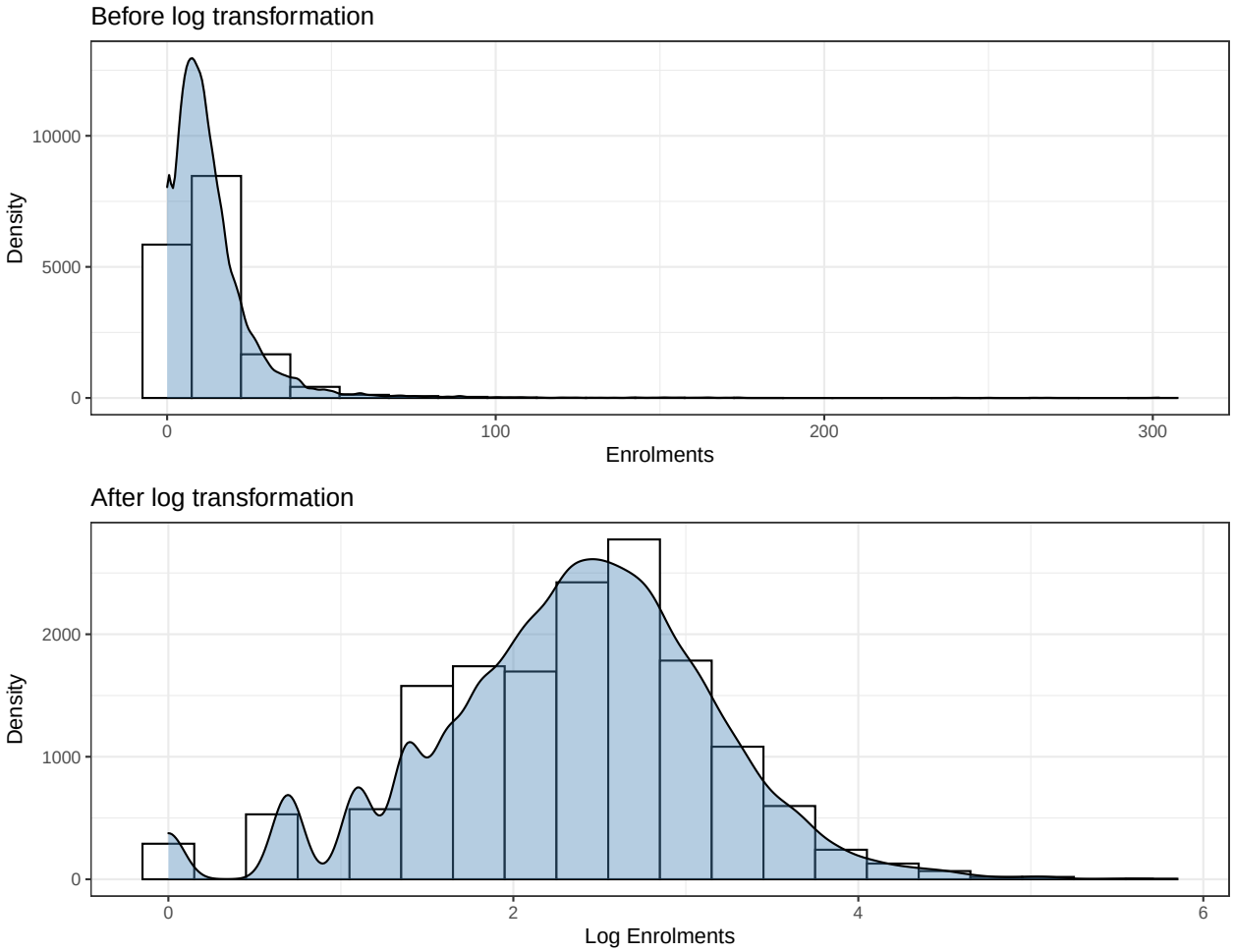

Figure 2: Effects of log transformation for response variable (enrolments) in Specialist Mathematics subject

As multilevel model assumes normality in the error terms, a log transformation is utilised to allow models to be estimated by the linear mixed models. The log transformation allows enrolment numbers to be approximately normally distributed (Figure 2).

Unconditional means model

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

	df	AIC
Model0.2: Schools nested within districts	4	25300.73
Model0.1: Schools nested within postcodes	4	25344.29
Model0.0: Within schools	3	25361.47

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)

```
summary(model0.2)
```

```
## Random effects:

##  Groups                        Name        Variance Std.Dev.
##  qcaa_school_id:qcaa_district (Intercept) 0.42131  0.64908
##  qcaa_district                (Intercept) 0.11159  0.33405
##  Residual                                 0.27720  0.52649

##
##  Fixed effects:

##             Estimate Std. Error  t value
## (Intercept) 2.120037 0.09828461 21.57039

##
##  Number of schools (level-two group) = 422
##  Number of district (level-three group) = 13
```

This model will takes into account 422 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 $t^* \neq t$:

$$\text{Level-two ICC} = \frac{\tau^2_{00}}{\tau^2_{00} + \phi^2_{00} + \sigma^2} = \frac{0.4213}{(0.4213 + 0.1116 + 0.2772)} = 0.5201$$

This can be conceptualised as the correlation between enrolments of two random draws from the same school at two different years. In other words, 52.01% 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 $t^* \neq t$.

$$\text{Level-three ICC} = \frac{\phi_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.1116}{(0.4213 + 0.1116 + 0.2772)} = 0.1378$$

Similarly, this can be conceptualised as the correlation between enrolments of two randomly selected schools from the same district – *i.e.* 11.70% 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.

```
summary(model1.0)
```

```
##  Groups                        Name        Variance  Std.Dev. Corr
##  qcaa_district:qcaa_school_id (Intercept) 0.8214386 0.906332
##                               year92      0.0011262 0.033559 -0.654
##  qcaa_district                (Intercept) 0.0000000 0.000000
##                               year92      0.0002295 0.015149   NaN
##  Residual                                 0.2092841 0.457476
```

```
##               Estimate Std. Error   t value
## (Intercept) 1.79689374 0.04801983 37.419825
## year92      0.01647784 0.00461389  3.571355
```

```
##  Number of Level Two groups =  422
##  Number of Level Three groups =  13
```

- $\pi_{0ij} = 1.7969$: Initial status for school i in district j (*i.e.* expected log enrolments when time = 0)
- $\pi_{1ij} = 0.0165$: Growth rate for school i in district j
- $\epsilon_{tij} = 0.2093$: Variance in within-school residuals after accounting for linear growth overtime

When the subject was first introduced in 1992, schools were expected to have 6.0309 ($e^{1.7848}$) enrolments, on average. This enrolments were rather low as there were only a small fraction of schools that offered the subject in 1992 (as demonstrated in Figure **??**). On average, enrolments were expected to increase by 1.6614% ($(e^{0.0164778} - 1) \times 100$) per year. The estimated within-schools variance decreased by 24.45% (0.2772 to 0.2093), implying that 24.5% of within-school variability can be explained by the linear growth over time.

Dealing with boundary constraints

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*).

```
summary(model1.1)
```

```
##  Groups                        Name        Variance  Std.Dev. Corr
##  qcaa_district:qcaa_school_id (Intercept) 0.8433365 0.918334
##                               year92      0.0012537 0.035407 -0.679
##  qcaa_district                (Intercept) 0.1201703 0.346656
##  Residual                                 0.2093102 0.457504
```

```
##               Estimate  Std. Error   t value
## (Intercept) 1.75561841 0.107914664 16.268581
## year92      0.01832492 0.001964398  9.328521
```

```
##  Number of Level Two groups =  422
##  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 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
model4.4	22025.38
model4.5	22027.29
model4.7	22028.11
model4.1	22030.06
model4.0	22033.52
model4.3	22044.92
model4.2	22045.67
model4.8	22045.67
model4.10	22045.67
model4.9	22045.67
model4.6	22047.61

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.

Testing random effects with parametric bootstrpa

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.

Table 3: Parametric Bootstrap to compare larger and smaller, nested model

npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr_boot(>Chisq)
10	24312.80	24389.11	-12146.40	24292.80	NA	NA	NA
12	22025.38	22116.96	-11000.69	22001.38	2291.42	2	0

Confidence interval

Table 4: 95% confidence intervals for fixed and random effects in the final model

var	2.5 %	97.5 %
sd_(Intercept)|qcaa_district:qcaa_school_id	0.8475964	0.9886445
cor_year92.(Intercept)|qcaa_district:qcaa_school_id	-0.7244002	-0.5994735
sd_year92|qcaa_district:qcaa_school_id	0.0314348	0.0370023
sd_(Intercept)|qcaa_district	0.1917148	0.5116679
sigma	0.4517649	0.4620813
(Intercept)	1.5792349	2.1738440
sectorGovernment	-0.2351881	0.2926828
sectorIndependent	-0.7884265	-0.1855885
unityear_12_enrolments	-0.0559129	-0.0286231
year92	0.0058831	0.0231169
sectorGovernment:year92	-0.0122945	0.0079922
sectorIndependent:year92	0.0074664	0.0310211

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 fixed and random effects all exclude 0 (Table 4), indicating that they're different from 0 in the population (*i.e.* statistically significant).

Interpreting the final model

Composite model

- Level one (measurement variable)

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

- Level two (schools within districts)

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

- Level three (districts)

$$\beta_{00j} = \gamma_{000} + r_{00j}$$
$$\beta_{01j} = \gamma_{010} + r_{01j}$$
$$\beta_{02j} = \gamma_{020} + r_{02j}$$
$$\beta_{10j} = \gamma_{100}$$
$$\beta_{11j} = \gamma_{110}$$

Therefore, the composite model can be written as

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

Fixed effects

```
summary(model_f)
```

```
##  Groups                        Name        Variance  Std.Dev. Corr
##  qcaa_district:qcaa_school_id (Intercept) 0.8370257 0.914891
##                               year92      0.0011733 0.034254 -0.666
##  qcaa_district                (Intercept) 0.1262182 0.355272
##  Residual                                 0.2087225 0.456862
```

```
##                             Estimate  Std. Error    t value
## (Intercept)               1.88540434 0.150880944 12.4959740
## sectorGovernment          0.02044649 0.131787850  0.1551470
## sectorIndependent        -0.48119714 0.148800988 -3.2338303
## unityear_12_enrolments   -0.04230158 0.007419052 -5.7017495
## year92                    0.01448120 0.004293656  3.3726968
## sectorGovernment:year92  -0.00194999 0.005033040 -0.3874377
## sectorIndependent:year92  0.01893009 0.005655238  3.3473547
```

```
##  Number of Level Two groups =  422
##  Number of Level Three groups =  13
```

Based on the model output (see detailed explanation of fixed effects in step 9), the estimated mean enrolments for government schools are estimated to be 2.06% ($(e^{0.02044} - 1) \times 100$) more than that of catholic schools when the subject is first introduced in 1992 (*i.e.* larger initial status). Government schools are estimated to have a mean increase in enrolments of 1.2610% ($(e^{(0.01448 - 0.0019)} - 1) \times 100$) per year, which is -0.1948% ($(e^{(} - 0.00194999) - 1) * 100)$) less than that of catholic schools.

On the other hand, independent schools have an estimated mean enrolments of 38% less than that of catholic schools when the subject is first introduced in 1992. However, this low initially low enrolments is matched with an a mean increase of 3.3976% per year, which is 1.9110% more than that of catholic schools per year.

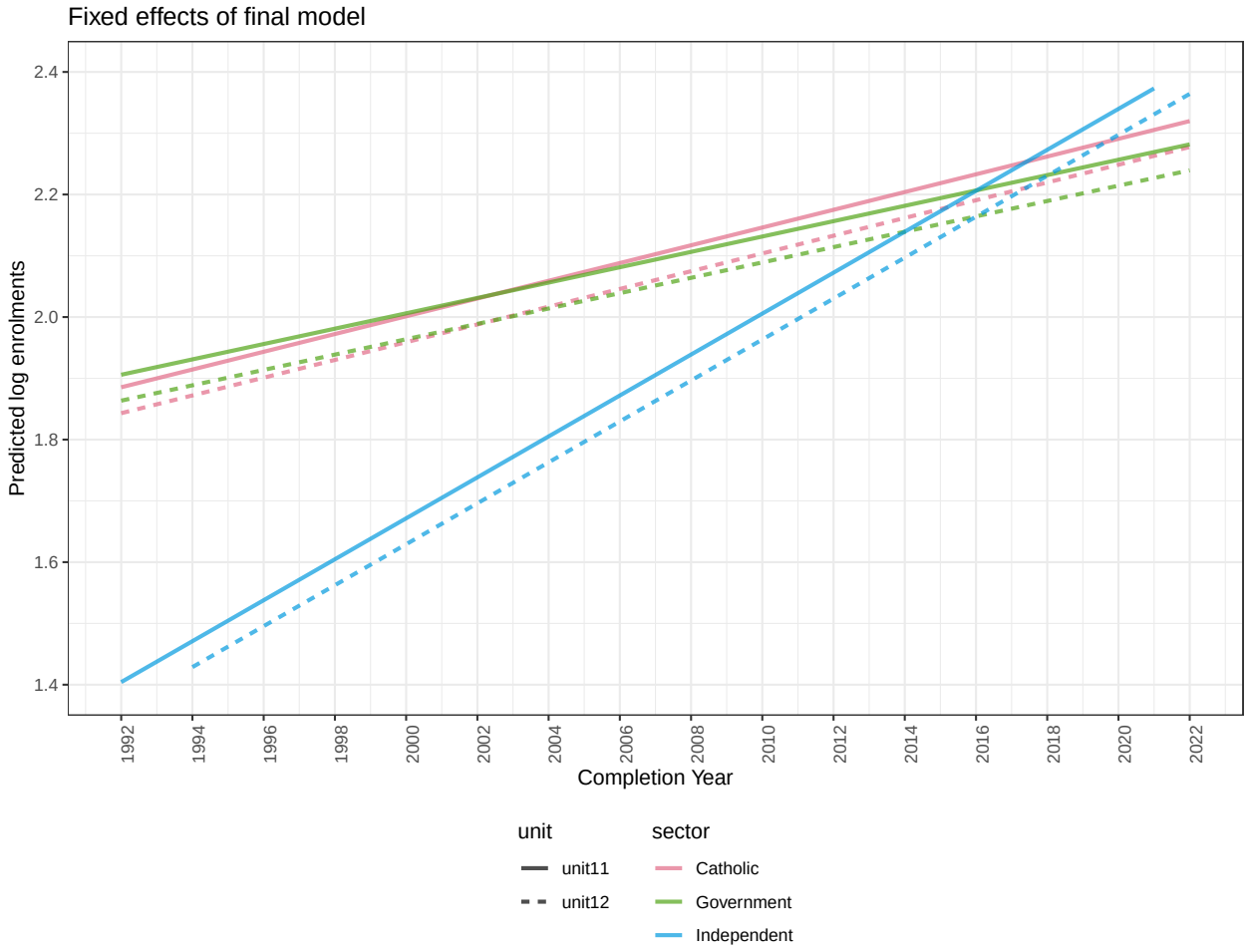

Figure 3: Fixed effects of the final model for Specialist Mathematics subject

Based on the fixed effects, independent schools are expected to have highest log enrolments after 2020 (Figure 3. In all sectors, unit 11 enrolments appears to be marginally larger than unit 12, which may imply that students are taking the subject in year 10, so lesser students are enrolled in the subject in year 11. Government sectors appears to have the highest enrolment numbers initially (in year 1992), but have increases at a relatively slower rate as compared to the other sectors.

To be noted, these fixed effects considers all schools within the different sectors; Therefore, the many small government schools with little enrolments in the subject (as seen in Figure **??**) may 'down weight' the mean fixed effects of all schools.

Random effects

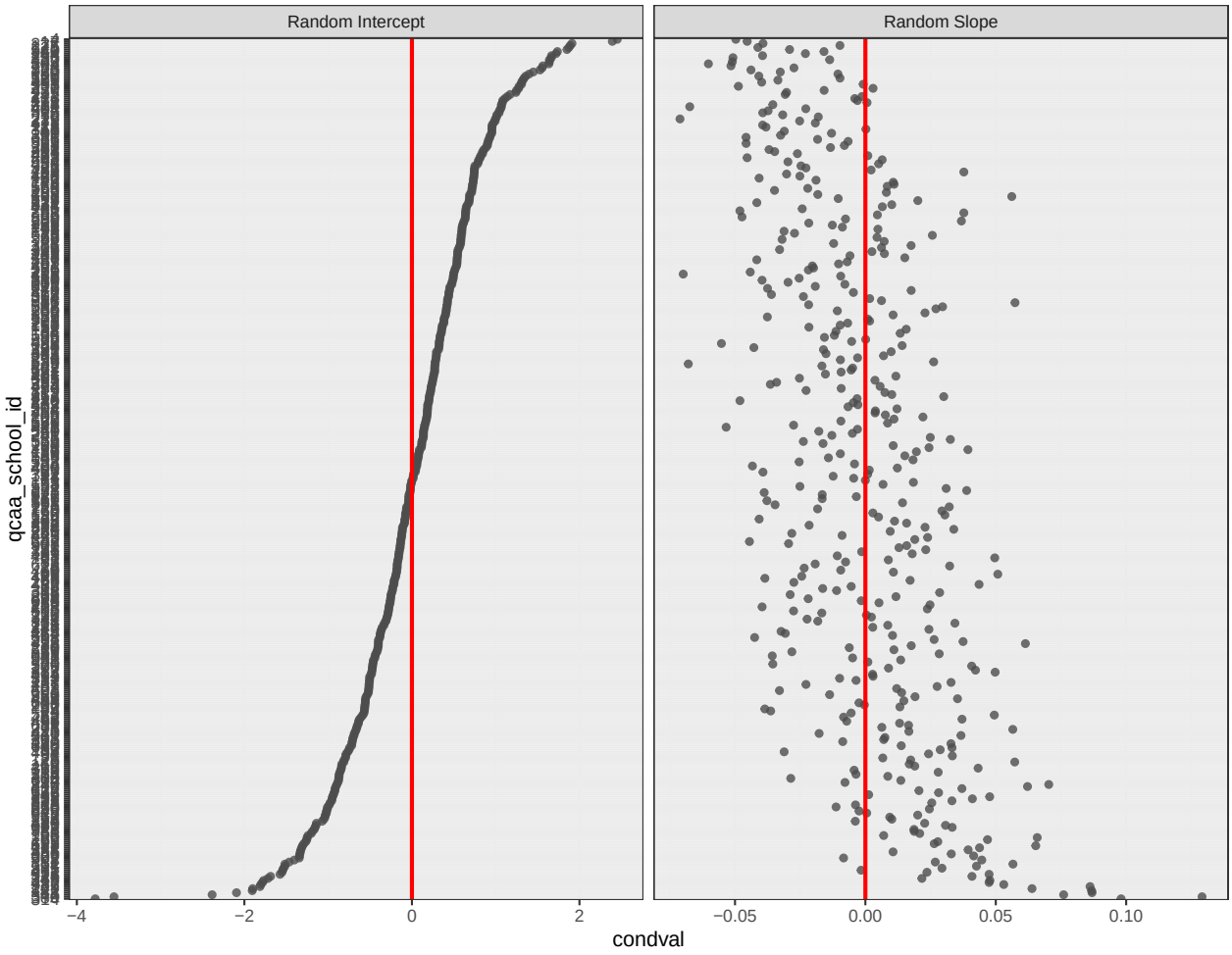

Figure 4: Random effects for all schools

Figure 4 represents the random intercept and slope of the random effects for a given school. It is a manifest that the intercept and slope are negatively correlated, where a large intercept is associated with a smaller random slope. 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.

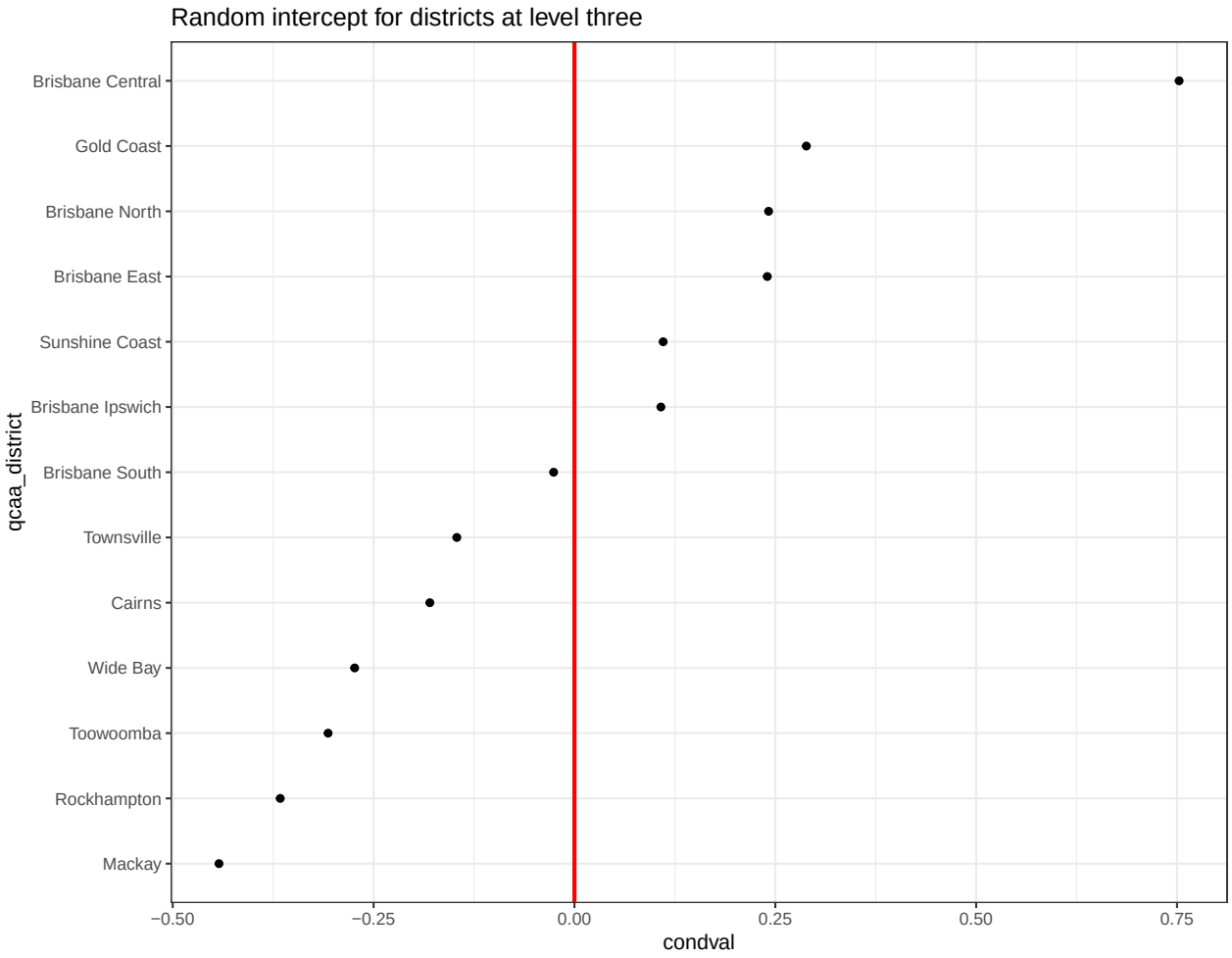

Figure 5: Random intercept for districts

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 Brisbane Central has the largest enrolments, on average.

Predictions

Figure 6: Model predictions for 20 randomly selected schools

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