Multilevel Model for Science in Practice

Brendi Ang

17/10/2021

Contents

Science in Practice

Exploring the dataset with basic linear model

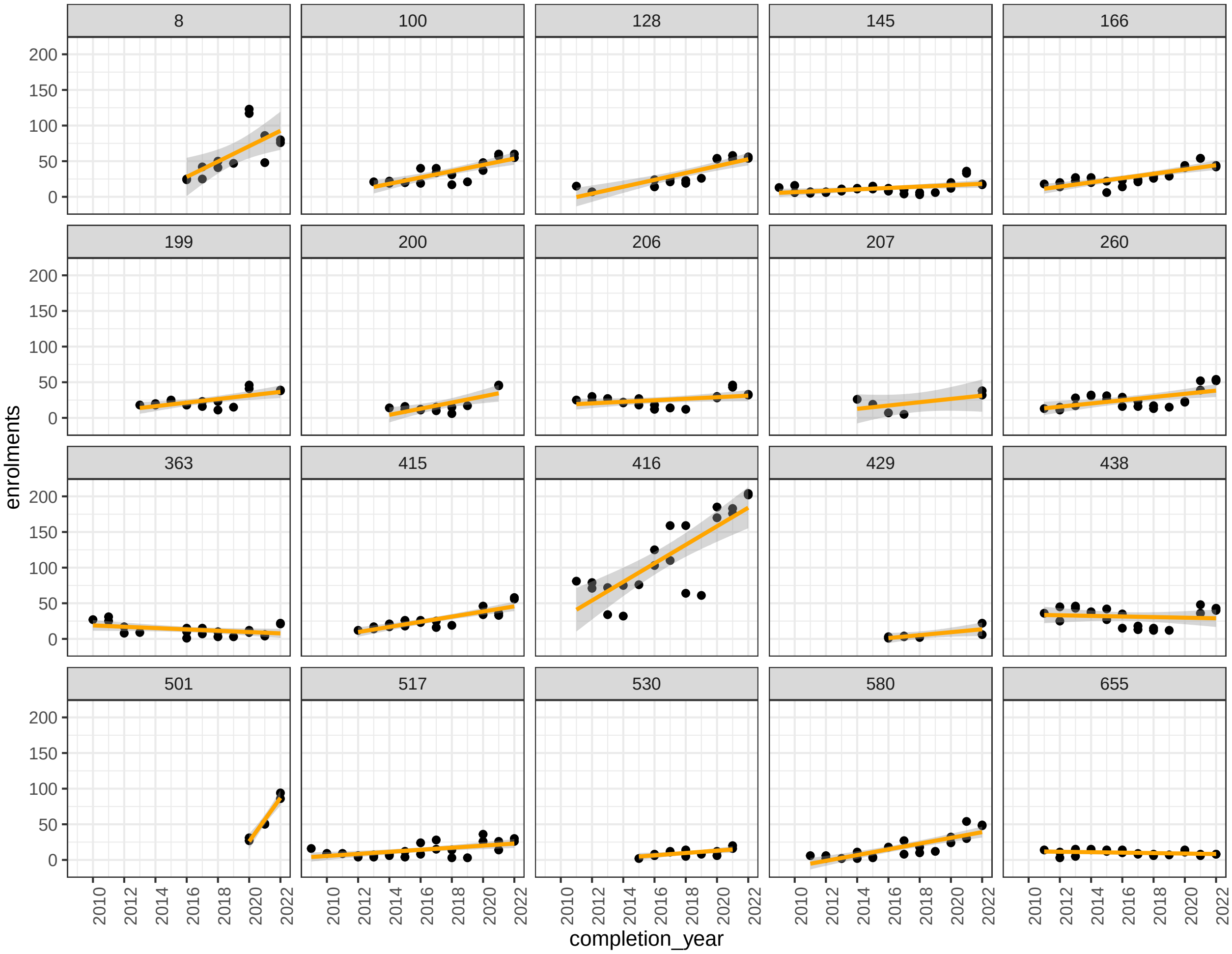

Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Specialist Mathematics subject

With reference to the first step, Figure 1 fits a linear model for enrolments for a random sample of 20 schools to provide insights on the enrolments trends in Science in Practice. As demonstrated, cohort size for each school vary, with smaller cohorts in schools 517, 530, 655 (bottom) which has enrolment numbers of approximately less than 50 each year; and larger schools such as school 416, which has more than 200 enrolments in a single cohort by the end of 2021. Some schools showed a steep increase in enrolments (*e.g.* school 416 and 501), while some schools showed rather constant or decreasing enrolments trends (*e.g.* school 655).

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

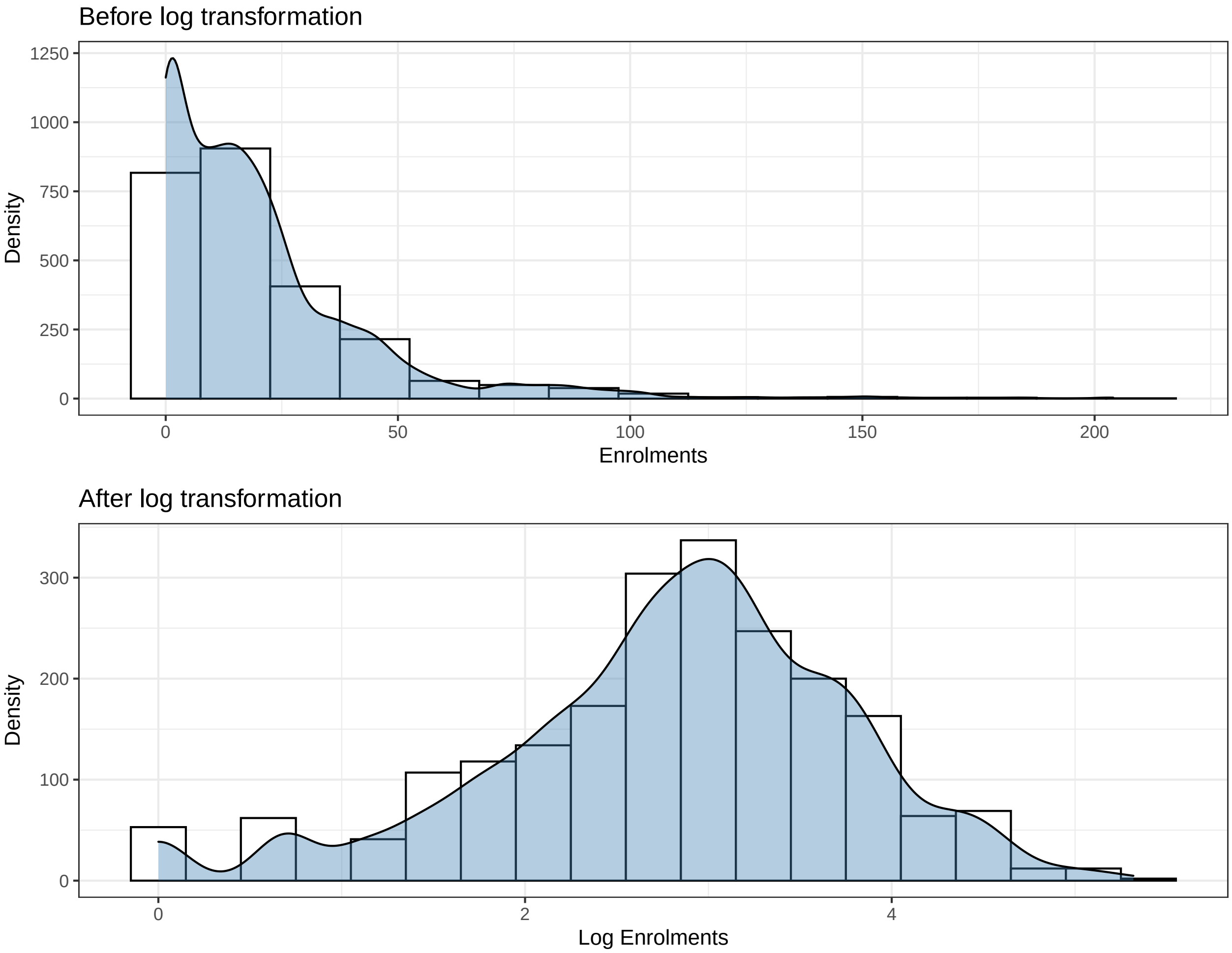

Figure 2: Effects of log transformation for response variable (enrolments) in Specialist Mathematics subject

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 Earth and Environmental Science

	df	AIC
Model0.0: Within schools	3	4705.261
Model0.2: Schools nested within districts	4	4705.552
Model0.1: Schools nested within postcodes	4	4707.261

As outlined in step 3, the three candidate models are fitted and their AIC is shown in Table 1. Based on the AIC, the two-level model (`model0.0`) is the superior model and will be used in the subsequent analysis.

Intraclass correlation (ICC)

```
## Random effects:
```

```
##  Groups         Name        Variance Std.Dev.
##  qcaa_school_id (Intercept) 0.62634  0.79142
##  Residual                   0.47198  0.68701
```

```
##
##  Fixed effects:
```

```
##             Estimate Std. Error  t value
## (Intercept)  2.75816 0.06012933 45.87046
```

```
##
##  Number of schools (level-two group) = 196
##  Number of district (level-three group) = NA
```

This model takes into account 196 schools. For a two-level multilevel model, the level two intraclass correlation coefficient (ICC) can be computed using the model output above. The **level-two ICC** is the correlation between a school i in time t and time t^*:

$$\text{level-two ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.6263}{(0.6263 + 0.4720)} = 0.5702$$

This can be conceptualised as the correlation between the enrolments of a selected school at two randomly drawn year (*i.e.* two randomly selected cohort from the same school). In other words, 57.02% of the total variability is attributable to the differences in enrolments within schools at different time periods.

Unconditional Growth model

```
##  Groups         Name        Variance Std.Dev. Corr
##  qcaa_school_id (Intercept) 1.108130 1.05268
##                 year10      0.010253 0.10126  -0.720
##  Residual                   0.306377 0.55351
```

```
##              Estimate  Std. Error  t value
## (Intercept) 1.8442381 0.097944037 18.82951
## year10      0.1004847 0.009731646 10.32556
```

```
##  Number of Level Two groups =  196
##  Number of Level Three groups =  NA
```

The next step involves incorporating the linear growth of time into the model. The model output is shown above.

- $\pi_{0ij} = 1.8442$: Initial status for school i (*i.e.* expected log enrolments when time = 0)
- $\pi_{1ij} = 0.1005$: Growth rate for school i
- $\epsilon_{tij} = 0.3064$: Variance in within-school residuals after accounting for linear growth overtime

When the subject was first introduced in 2010, schools were expected to have an average of 6.3230 ($e^{1.8442}$) enrolments, which is a relatively low number as compared to the other mathematics and science subjects. On average, the enrolments were expected to increase by 10.5724% ($(e^{0.1005} - 1) \times 100$) per year. The estimated within-school variance decreased by 16.70% (0.4720 to 0.30638), indicating the 35.089% can be explained by the linear growth in time.

Testing fixed effects

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

model	npar	AIC	BIC	logLik
model4.1	14	4022.482	4101.041	-1997.241
model4.6	12	4024.466	4091.802	-2000.233
model4.9	13	4025.268	4098.215	-1999.634
model4.7	13	4025.268	4098.215	-1999.634
model4.8	13	4025.268	4098.215	-1999.634
model4.0	16	4025.315	4115.096	-1996.657
model4.5	12	4025.633	4092.969	-2000.816
model4.3	10	4027.180	4083.293	-2003.590
model4.2	11	4027.303	4089.028	-2002.651
model4.4	11	4027.477	4089.202	-2002.739
model4.10	11	4031.656	4093.381	-2004.828

As summarise in step 6, level-two predictors `sector` and `unit` will be added to the model. The largest possible model (`model4.0`) will first be fitted, before iteratively removing fixed effects one at a time (with `model4.10` being the smallest of all 10 candidate models), whilst recording the AIC for each model. `model4.1` appears to have the optimal (smallest) AIC (Table 2), and will be used in the next section in building the final model.

Parametric bootstrap to test random effects

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

npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr_boot(>Chisq)
12	4227.105	4294.442	-2101.553	4203.105	NA	NA	NA
14	4022.482	4101.041	-1997.241	3994.482	208.6234	2	0

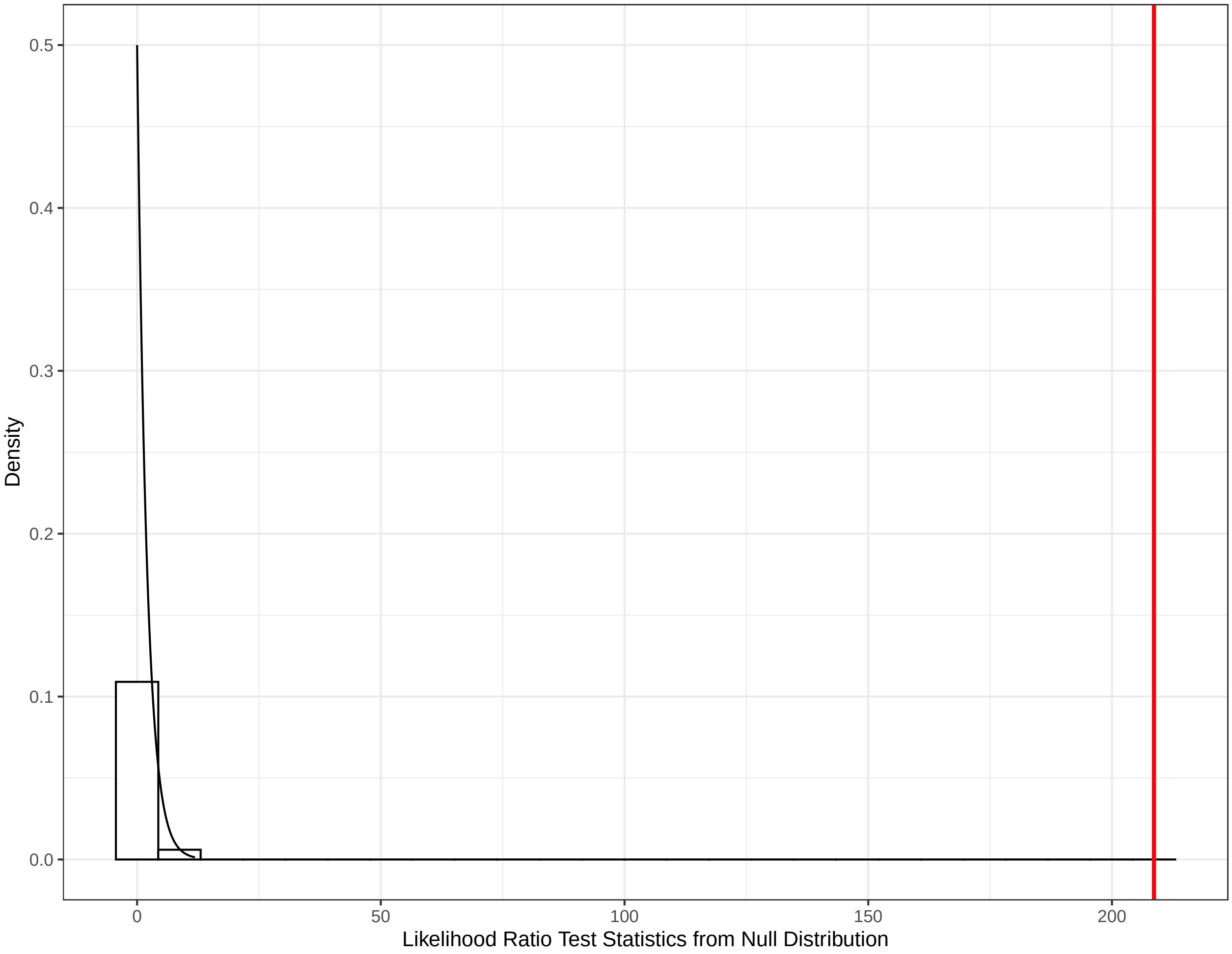

Figure 3: Histogram of likelihood ratio test statistic, with a red vertical line indicating the likelihood ratio test statistic for the actual model

The parametric bootstrap is used to approximate the likelihood ratio test statistic to produce a more accurate p-value by simulating data under the null hypothesis (detailed explanation can be found in step 7. The p-value indicates the proportion of times in which the bootstrap test statistic is greater than the observed test statistic. Figure 3 displays the likelihood ratio test statistic from the null distribution, with the red line indicates the likelihood ratio test statistic using the actual data.

There is overwhelming statistical evidence ($\chi^2 = 208.623$ and p-value $= 0$ from Table 3) that the larger model (including random slope at level two) is the better model.

Confidence interval

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

var	2.5 %	97.5 %
sd_(Intercept)|qcaa_school_id	1.0791392	1.5883744
cor_year92.(Intercept)|qcaa_school_id	-0.9079310	-0.6751860
sd_year92|qcaa_school_id	0.0393192	0.0678080
sigma	0.4746765	0.5167394
(Intercept)	0.5979313	2.9539799
year92	-0.0293459	0.0774094
sectorGovernment	-1.6512267	0.7654638
sectorIndependent	-0.4903149	2.2137154
unityear_12_enrolments	-0.0447976	0.0733969
year92:sectorGovernment	-0.0366734	0.0751267
year92:sectorIndependent	-0.0819655	0.0454768

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, indicating that they're different from 0 in the population (*i.e.* statistically significant).

Interpreting final model

Composite model

- Level one (measurement variable)

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} year10_{tij} + \epsilon_{tij}$$

- Level two (schools within districts)

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

Therefore, the composite model can be written as

$$\begin{aligned} Y_{tij} &= \pi_{0ij} + \pi_{1ij} year10_{tij} + \epsilon_{tij} \\ &= (\beta_{00j} + \beta_{01j} sector_{ij} + \beta_{02j} unit_{ij} + \beta_{03j} sector_{ij} unit_{ij} + u_{0ij}) + \\ &\quad (\beta_{10j} + \beta_{11j} sector_{ij} + \beta_{12} unit_{ij} + u_{1ij}) year10_{tij} + \epsilon_{tij} \\ &= \left[\beta_{00j} + \beta_{01} sector_{ij} + \beta_{02} unit_{ij} + \beta_{03j} sector_{ij} unit_{ij} + \beta_{10j} year10_{tij} + \beta_{11j} sector_{ij} year_{tij} + \beta_{12} unit_{ij} year10_{tij}\right] \left[u_{0ij} + \right. \end{aligned}$$

Fixed Effects

```
summary(model)
```

```
##  Groups         Name        Variance  Std.Dev. Corr
##  qcaa_school_id (Intercept) 1.0471207 1.023289
##                 year10      0.0097253 0.098617 -0.802
##  Residual                   0.3036501 0.551045
```

```
##                                              Estimate  Std. Error   t value
## (Intercept)                                1.92746582 0.296641578  6.497625
## year10                                     0.03386364 0.030295845  1.117765
## sectorGovernment                           0.11738215 0.315770024  0.371733
## sectorIndependent                         -0.78920786 0.393905776 -2.003545
## unityear_12_enrolments                    -0.02478714 0.091733860 -0.270207
## year10:sectorGovernment                    0.07898180 0.032078517  2.462140
## year10:sectorIndependent                   0.06729049 0.038678772  1.739727
## year10:unityear_12_enrolments              0.01479030 0.006765064  2.186277
## sectorGovernment:unityear_12_enrolments  -0.15761038 0.082352025 -1.913862
## sectorIndependent:unityear_12_enrolments -0.27451131 0.102332610 -2.682540
```

```
##  Number of Level Two groups =  196
##  Number of Level Three groups =  NA
```

Based on the model output, government schools are estimated to have a mean increase of 11.9459% $((e^{0.0248299+0.0197221}-1)*\times 100)$ per year, which is 8.21846% $((e^{0.0789818}-1)*\times 100)$ greater than the increase in enrolments in catholic schools. Independent schools showed a large increase in enrolments (10.6447%) per year, on average. This increase in enrolments is 6.96062% $((exp^{-0.0221572}-1)*100)$ more than that of catholic schools.

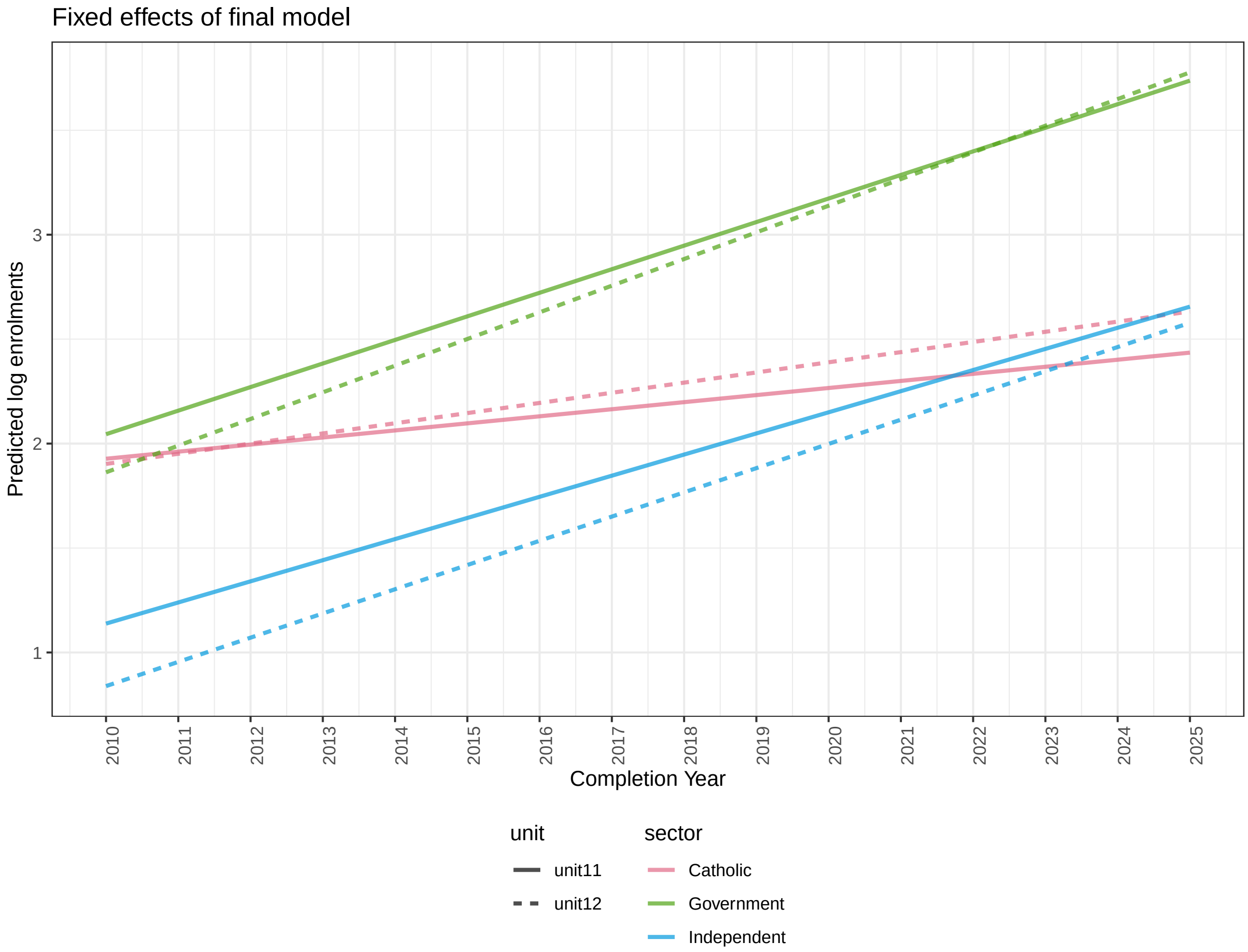

Figure 4: Fixed effects of the final model for Science in Practice

The results can be better visualised in Figure 4. Catholic schools are estimated to have smallest growth (as shown by the gentle slope) relative to the other two sectors. By the same token, year 12 enrolments appears to be increasing at a faster rate than that of year 11 units, this may indicate that students may be opting to take the unit in year 10 and re-enrol in year 12.

Random effects

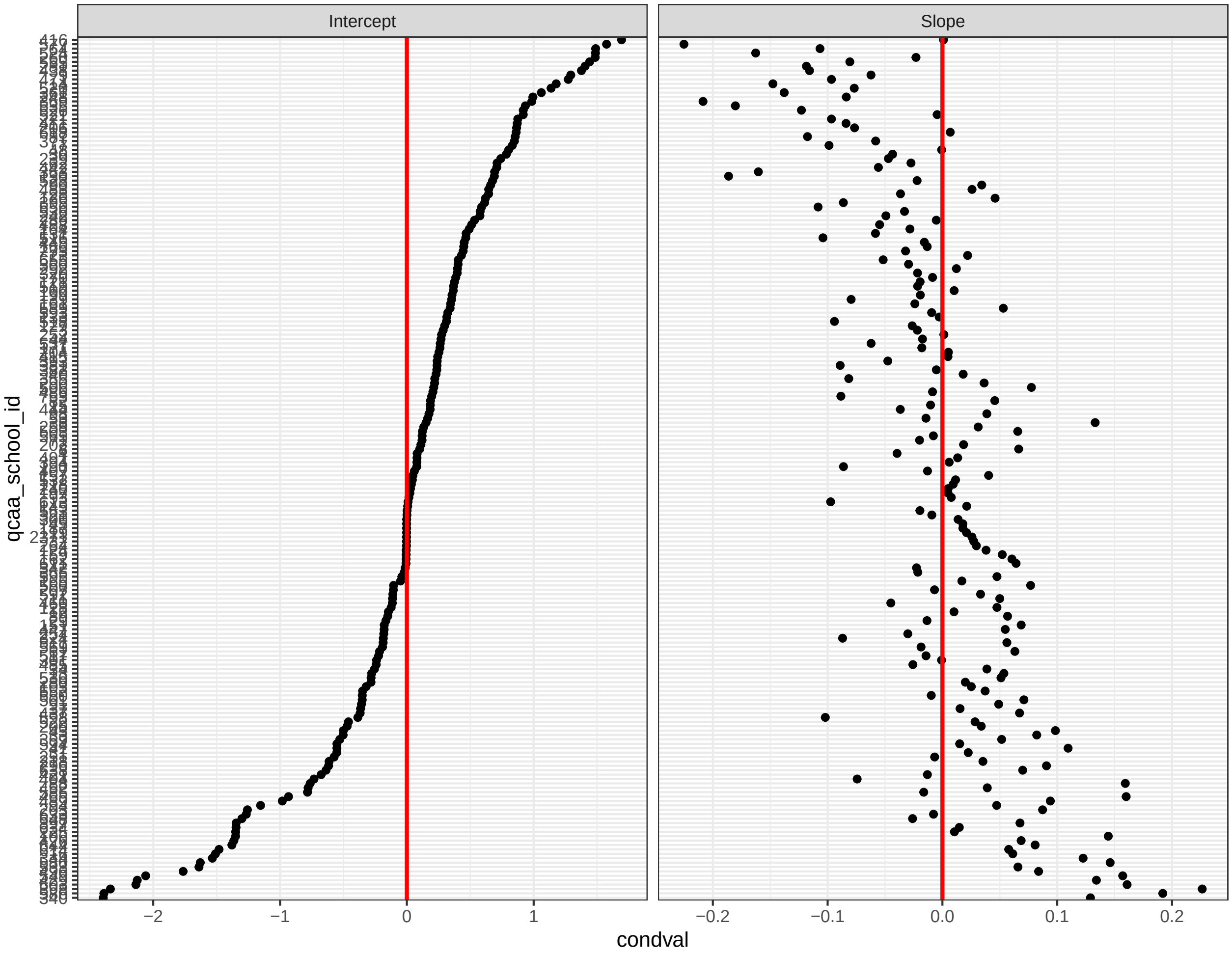

Figure 5: Random effects for all schools

Figure 5 shows the random effects for all 196 schools that offered the subject. There is a clear negative correlation (-0.80) between the random intercept and the random slope, which indicates that in general, schools with lesser enrolments are generally matched with a larger increase in enrolments over the years.

Predictions

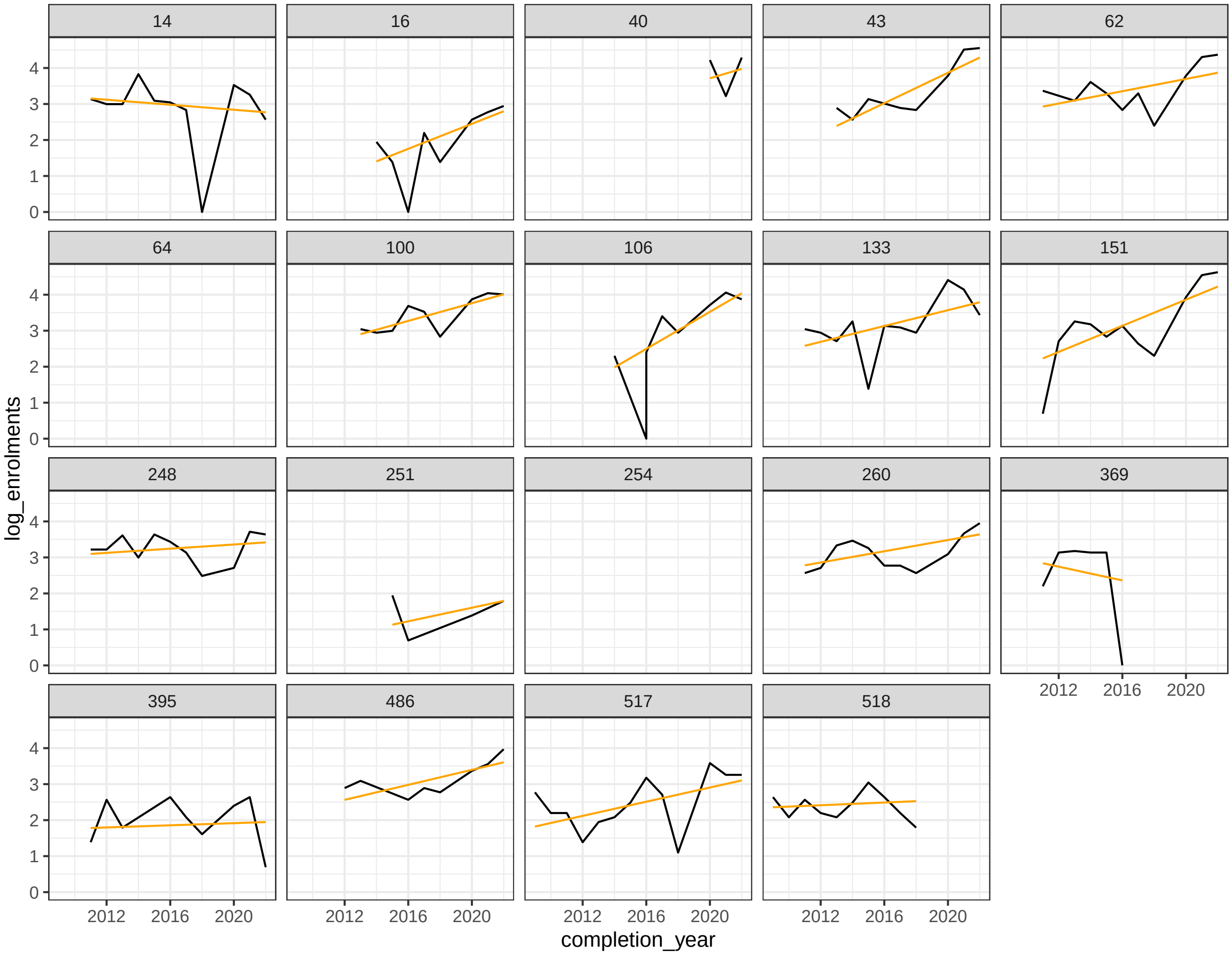

Figure 6: Model predictions for year 11 enrolments for 20 randomly selected schools

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