Multilevel Model for Biology

Brendi Ang

17/10/2021

Contents

Biology

Exploring the dataset with basic linear model for each school

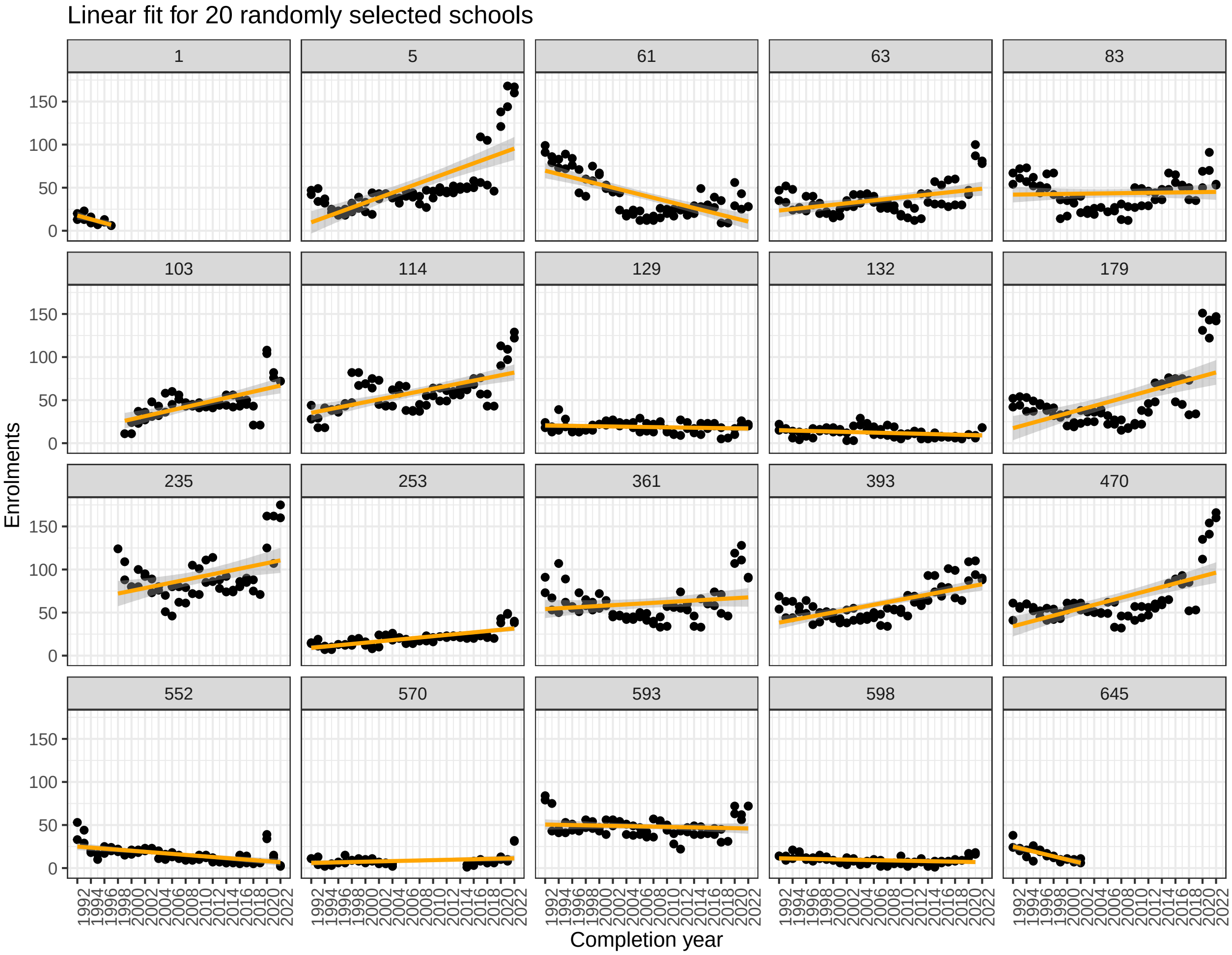

Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Biology subject

As described in step 1, a basic linear model was plotted for each school to provide insights of the enrolment trends for each school. Figure 1 demonstrates Biology may be introduced in schools in later years (*e.g.* school 235) and schools may have removed the subject and in some cases, school 1 and school 645 did not exist after 1997. School 570 showed a halt in the subject from 2005 to 2014. Schools can also vary greatly in enrolment sizes, for instance, schools 129, 132 and 570 had less than 50 enrolments for every cohort while some schools (*e.g.* school 235) have relatively larger cohort size. Some schools also showed a general decrease in enrolments across the years while some schools such at school 5 and 470 showed a significant increase in enrolments over the years.

Getting the data ready for modelling

Removing graduating cohort 2019

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.

Linearise relationship using log transformation

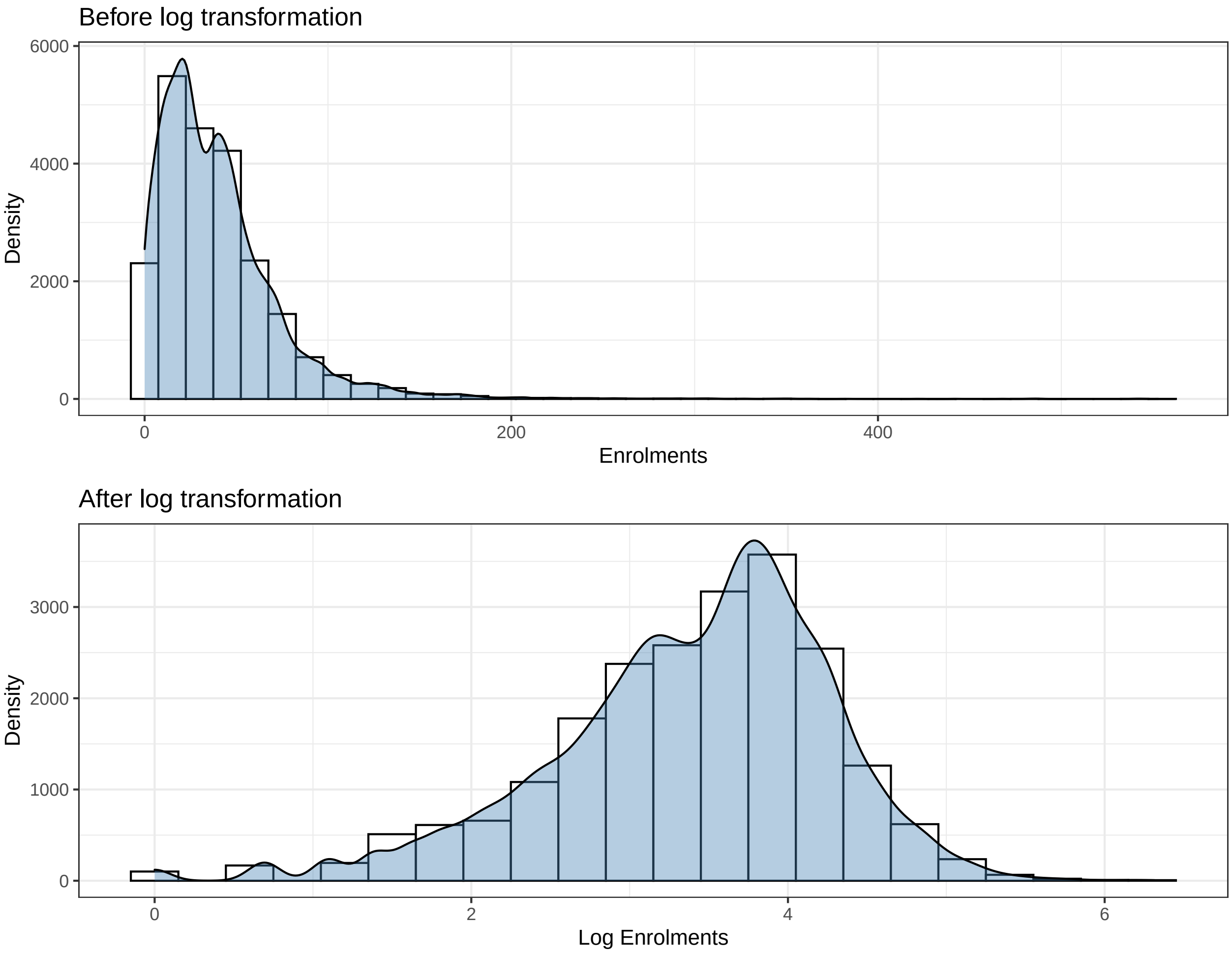

Figure 2: Effects of log transformation for response variable (enrolments) in Biology subject

A log transformation was used on the response variable (`enrolments`) to allow model to be estimated by the multilevel model, which assumes normality in the error terms. As shown in Figure 2.

Unconditional means model

Table 1: AIC values for all candidate models for Biology

	df	AIC
Model0.2: Schools nested within districts	4	30247.41
Model0.1: Schools nested within postcodes	4	30284.86
Model0.0: Within schools	3	30290.46

Referring back to step 3, three candidate models are fitted, with the AIC shown in Table 1. `Model0.2`, corresponding to having schools nested within districts is the best model, with optimised (lowest) AIC 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) 0.61534  0.78444
##  qcaa_district                (Intercept) 0.11133  0.33366
##  Residual                                 0.22037  0.46943

##
##  Fixed effects:

##             Estimate Std. Error  t value
## (Intercept) 3.271578 0.09968013 32.82077

##
##  Number of schools (level-two group) = 468
##  Number of district (level-three group) = 13
```

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 a certain district k in time t and in time t^*:

$$\text{Level-two ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.6153}{(0.6153 + 0.1113 + 0.2204)} = 0.6497$$

This can be conceptualised as the correlation between enrolments of two random draws from the same school at two different years. In other words, 64.97% of the total variability is attributable to the differences between schools from the same district rather than changes over time within schools.

The **level-three ICC** refers to the correlation between different schools i and i^* from a specific school j.

$$\text{Level-three ICC} = \frac{\phi_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.1113}{(0.6153 + 0.1113 + 0.1113)} = 0.1175$$

Similarly, it can be inferred that the correlation between enrolments of two randomly selected schools from different districts are 11.75%, where the total variability can be attributed to the difference between districts.

Unconditional growth model

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

```
##  Groups                       Name        Variance   Std.Dev.  Corr
##  qcaa_district:qcaa_school_id (Intercept) 2.1466e+00 1.4651162
##                               year92      2.3778e-03 0.0487625 -0.805
##  qcaa_district                (Intercept) 6.5640e-02 0.2562035
##                               year92      6.0662e-05 0.0077886 0.409
##  Residual                                 1.5031e-01 0.3876979
```

```
##               Estimate  Std. Error   t value
## (Intercept) 2.68260254 0.101014731 26.556548
## year92      0.02593676 0.003253587  7.971744
```

```
##  Number of Level Two groups =  468
##  Number of Level Three groups =  13
```

The unconditional growth model adds the systematic changes over time, the model specification can be found in step 4. This allows for assessing within-school variability which can be attributed to the linear changes over time. Based on the model output:

- $\pi_{0ij} = 2.6826$: Initial status for school i in district j (*i.e.* expected log enrolments when time = 0)
- $\pi_{1ij} = 0.0259$: Growth rate for school i in district j
- $\epsilon_{tij} = 0.1503$: Variance in within-school residuals after accounting for linear growth overtime

Biology was first introduced in 1992, and schools are expected to have 14.62 ($e^{2.68260}$), on average. Furthermore, enrolments were expected to increase by 2.628% ($(e^{0.02594} - 1) \times 100$) every year. The estimated within-school variance decrease by 24.49% (0.2204 to 0.1503), implying that 24.49% of the within-school variability can be explained by the linear growth over time.

Testing fixed effects

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

model	npar	AIC	BIC	logLik
model4.1	17	23783.03	23918.34	-11874.51
model4.0	19	23784.32	23935.55	-11873.16
model4.7	16	23784.71	23912.06	-11876.35
model4.5	15	23789.67	23909.07	-11879.84
model4.4	14	23791.07	23902.50	-11881.53
model4.9	14	23853.67	23965.11	-11912.84
model4.10	14	23853.67	23965.11	-11912.84
model4.2	14	23853.67	23965.11	-11912.84
model4.8	14	23853.67	23965.11	-11912.84
model4.3	13	23858.62	23962.10	-11916.31
model4.6	15	23859.14	23978.53	-11914.57

As detailed in step 6, level-two predictors (`sector` and `unit`) are added to the model. The largest possible model will be fitted, before removing each fixed effect one by one whilst recording the AIC for each model. `model4.0` corresponds to the largest model while `model4.10` is the smallest possible model. The model with the optimal (lowest) AIC is `model4.1`, and will be used in subsequent sections.

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)
15	23781.72	23901.12	-11875.86	23751.72	NA	NA	NA
17	23783.03	23918.35	-11874.51	23749.03	2.694776	2	0.126

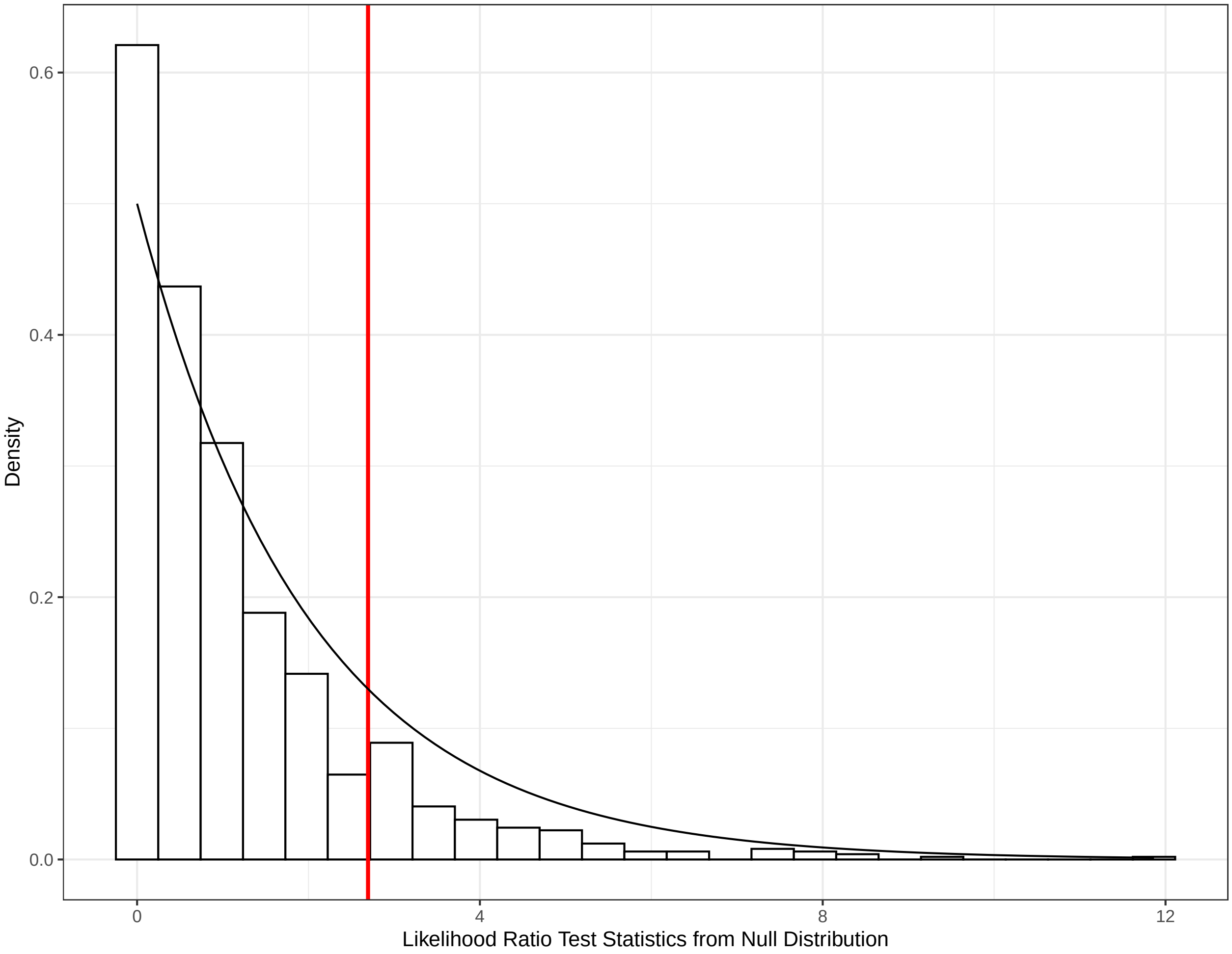

Figure 3: Histogram of likelihood ratio test statistic, with a red vertical line indicating the likelihood ratio test statistic for the actual model

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. Figure 3 displays the likelihood ratio test statistic from the null distribution, with the red line representing the likelihood ratio test statistic using the actual data. The p-value of 14.5% (Table 3) indicates the proportion of times in which the bootstrap test statistic is greater than the observed test statistic. The large estimated p-value is $0.145 < 0.05$ fails to reject the null hypothesis at the 5% level, indicating that the smaller model (without random slope at level three) is preferred.

Confidence interval

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

var	2.5 %	97.5 %
sd_(Intercept)|school_postcode:qcaa_school_id	0.8530526	1.2182627
cor_year94.(Intercept)|school_postcode:qcaa_school_id	-0.9690616	-0.8565112
sd_year94|school_postcode:qcaa_school_id	0.0315056	0.0460868
sd_(Intercept)|school_postcode	0.0000009	0.4581153
sigma	0.4876907	0.5140716
(Intercept)	1.4227924	2.2504155
year94	0.0005497	0.0184106
sectorGovernment	-0.3117705	0.4153108
sectorIndependent	-0.0628937	0.7950850
unityear_12_enrolments	-0.2747165	-0.1289871
year94:unityear_12_enrolments	0.0026651	0.0107254

The parametric bootstrap is utilised to construct confidence intervals (as detailed in step 8). If the confidence intervals for 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 95% confidence interval is shown above (Table 4), and the random effects all exclude 0, further reiterating that they are statistically significant at the 5% level. Some fixed effects such as `unityear_12_enrolments` were insignificant, suggesting that there were no differences between unit 11 and unit 12 units.

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)

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

- Level three (districts)

$$
\begin{aligned}
\beta_{00j} &= \gamma_{000} + r_{00j} \\
\beta_{01j} &= \gamma_{010} + r_{01j} \\
\beta_{02j} &= \gamma_{020} + r_{02j} \\
\beta_{03j} &= \gamma_{030} + r_{03j} \\
\beta_{10j} &= \gamma_{100} \\
\beta_{11j} &= \gamma_{110}
\end{aligned}
$$

Therefore, the composite model can be written as

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

Fixed effects

```
summary(model_f)
```

```
##  Groups                       Name        Variance  Std.Dev. Corr
##  qcaa_district:qcaa_school_id (Intercept) 1.8582181 1.363165
##                               year92      0.0020144 0.044882 -0.777
##  qcaa_district                (Intercept) 0.1468611 0.383225
##  Residual                                 0.1493206 0.386420
```

```
##                                             Estimate   Std. Error    t value
## (Intercept)                              2.967847727 0.1906944799 15.5633646
## year92                                   0.026446106 0.0053039215  4.9861420
## sectorGovernment                         0.170207789 0.1829943554  0.9301259
## sectorIndependent                       -1.266171057 0.2020482537 -6.2666766
## unityear_12_enrolments                  -0.003237843 0.0157576995 -0.2054769
## year92:sectorGovernment                 -0.016269000 0.0061142842 -2.6608184
## year92:sectorIndependent                 0.031103798 0.0067695649  4.5946524
## year92:unityear_12_enrolments           -0.001164564 0.0006060924 -1.9214299
## sectorGovernment:unityear_12_enrolments -0.046618866 0.0143680795 -3.2446136
## sectorIndependent:unityear_12_enrolments -0.039104708 0.0162432846 -2.4074384
##  Number of Level Two groups =  468
##  Number of Level Three groups =  13
```

Using the model output above (see step 9 for detailed explanation on fixed effects), the estimated increase in mean enrolments for government schools is 1.0229% ($e^{0.0264-0.0162}-1) \times 100$), which is 1.6402% (($e^{0.01626900}-1) \times 100$) less than that of catholic schools. On the other hand, the mean enrolments for independent schools are estimated to increase by 5.9238% ($e^{0.0264-0.0162}-1) \times 100$) each year, which is 3.1592% more than catholic schools.

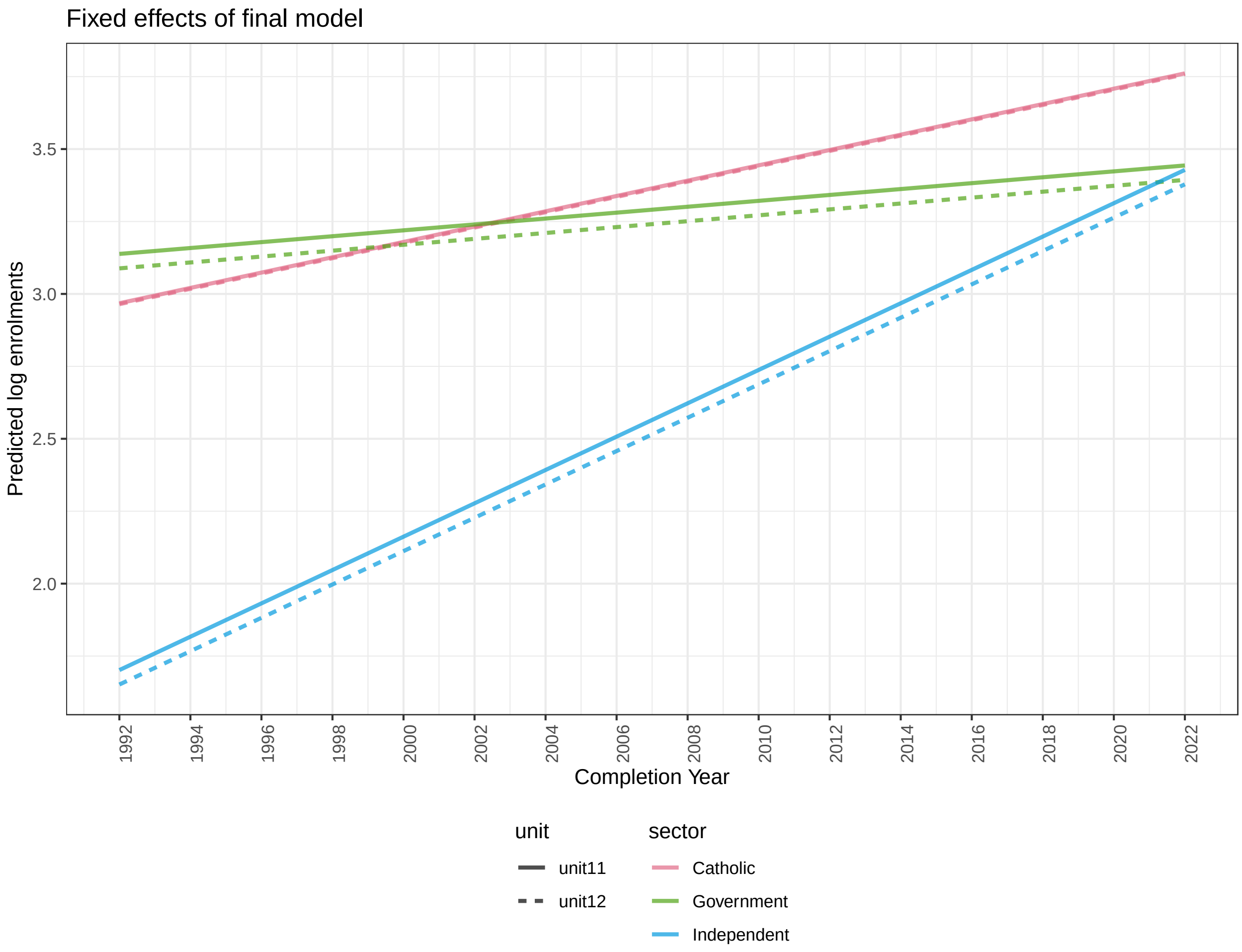

Figure 4: Fixed effects of the final model for Biology subject

The fixed effects can be better visualised in Figure 4, independent schools appears to have the highest average increase in enrolments per year. It appears that after 2022, enrolments in independent schools are predicted to be higher than government schools, on average. Unit 11 enrolments only appears to be marginally smaller than unit 12 enrolments for government and independent schools, but appears to be the same for catholic schools. Government schools starts of (in 1992) with the highest enrolments on average, but it is matched with a slow increase in enrolments over the years.

Random effects

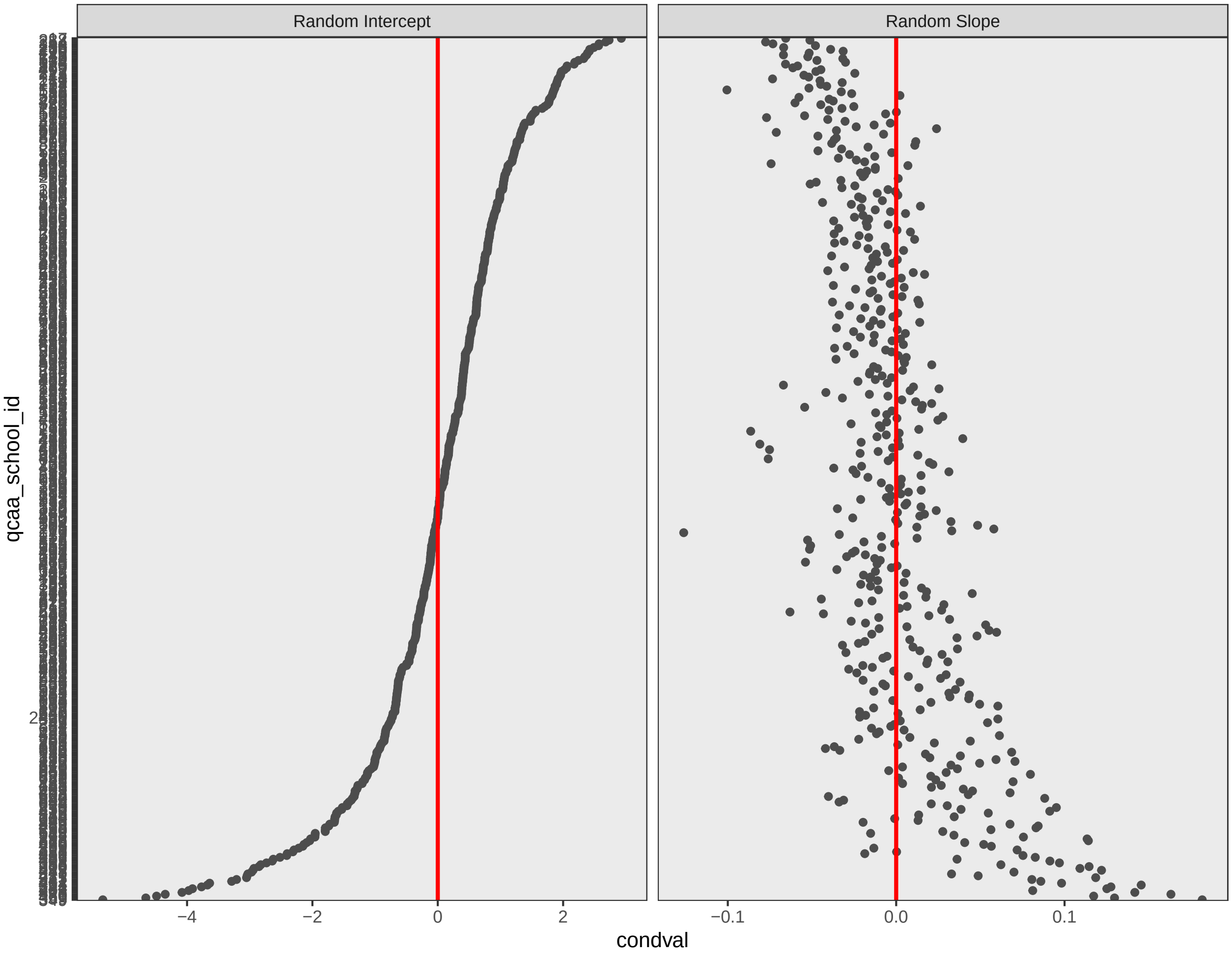

Figure 5: Random effects for schools

Figure 5 displays the random effects for a given school. It is apparent that the random intercepts and slopes are negatively correlated, where a large intercept is associated with a smaller random slope. This indicates that a larger school is associated with a smaller increase (decrease) in enrolments over the years while smaller schools are predicted to have larger increase in enrolments over the years.

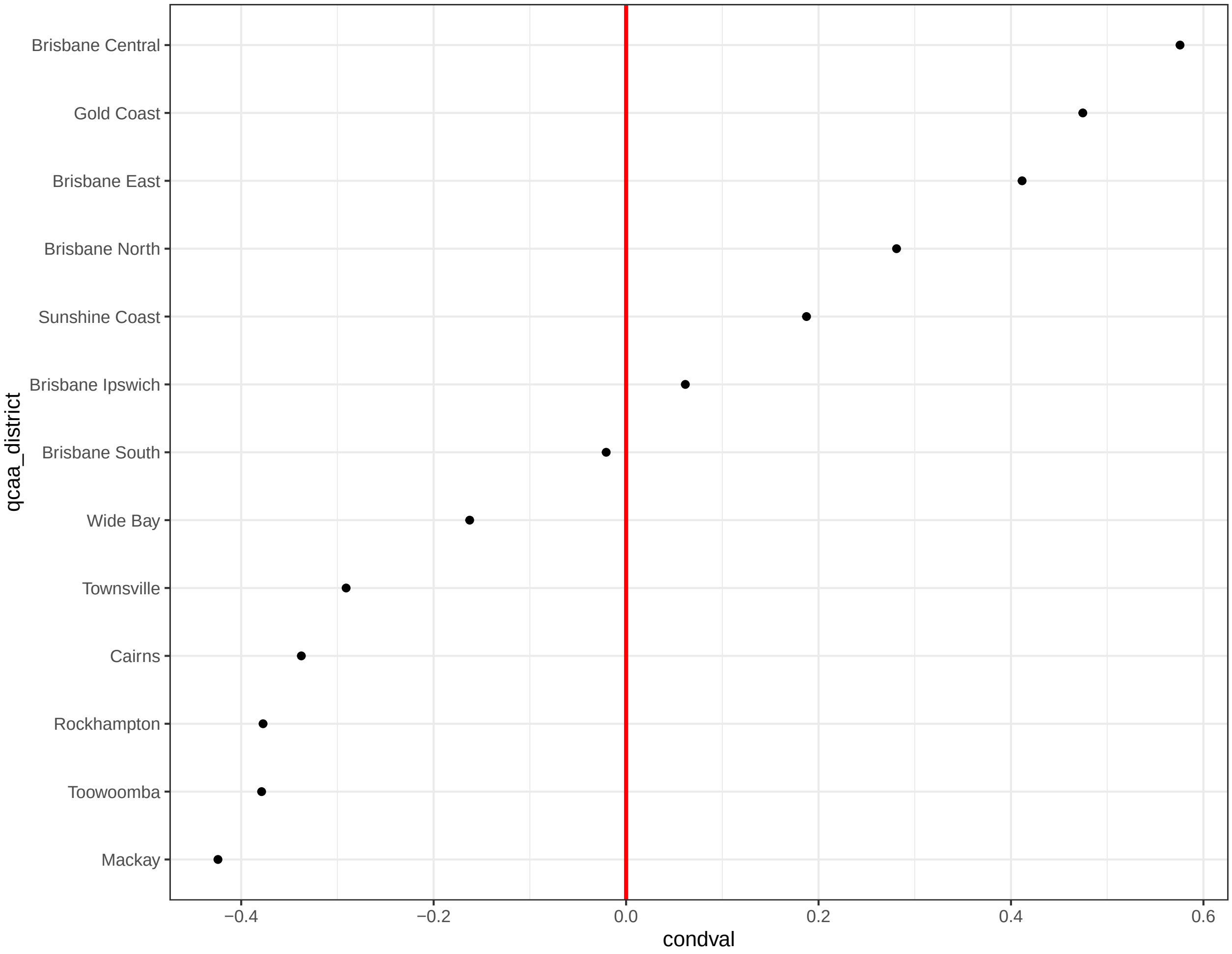

Figure 6: Random intercept for districts

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 6 demonstrates that schools in Brisbane Central has the largest enrolments while Mackay have the lowest enrolments in Biology subject, on average.

Predictions

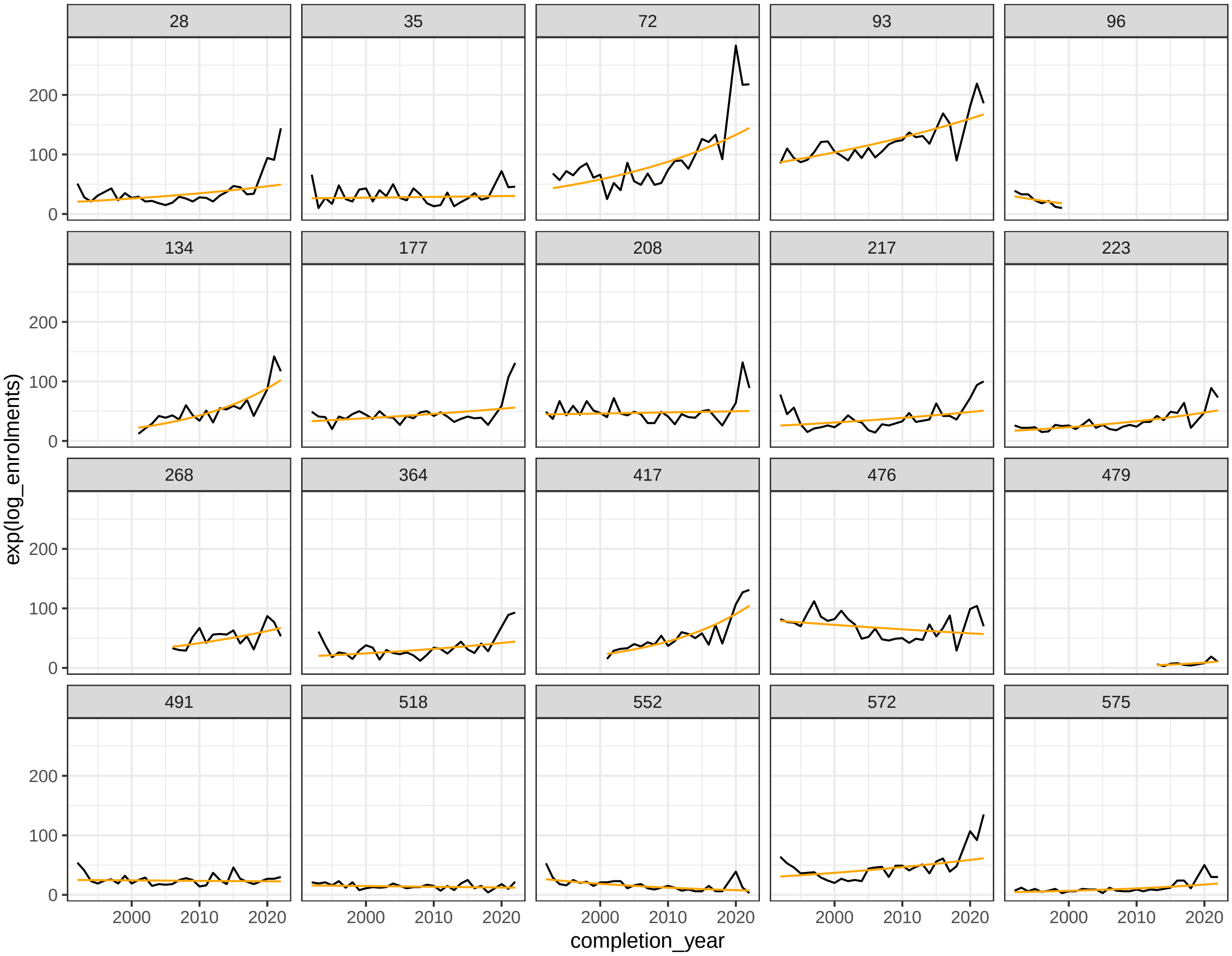

Figure 7: Model predictions for 20 randomly selected schools

Figure 7 above shows the predictions for 20 randomly selected schools.