Multilevel Model for Aquatic Practices

Brendi Ang

17/10/2021

Contents

Aquatic Practices

Exploring the dataset with basic linear model for each school

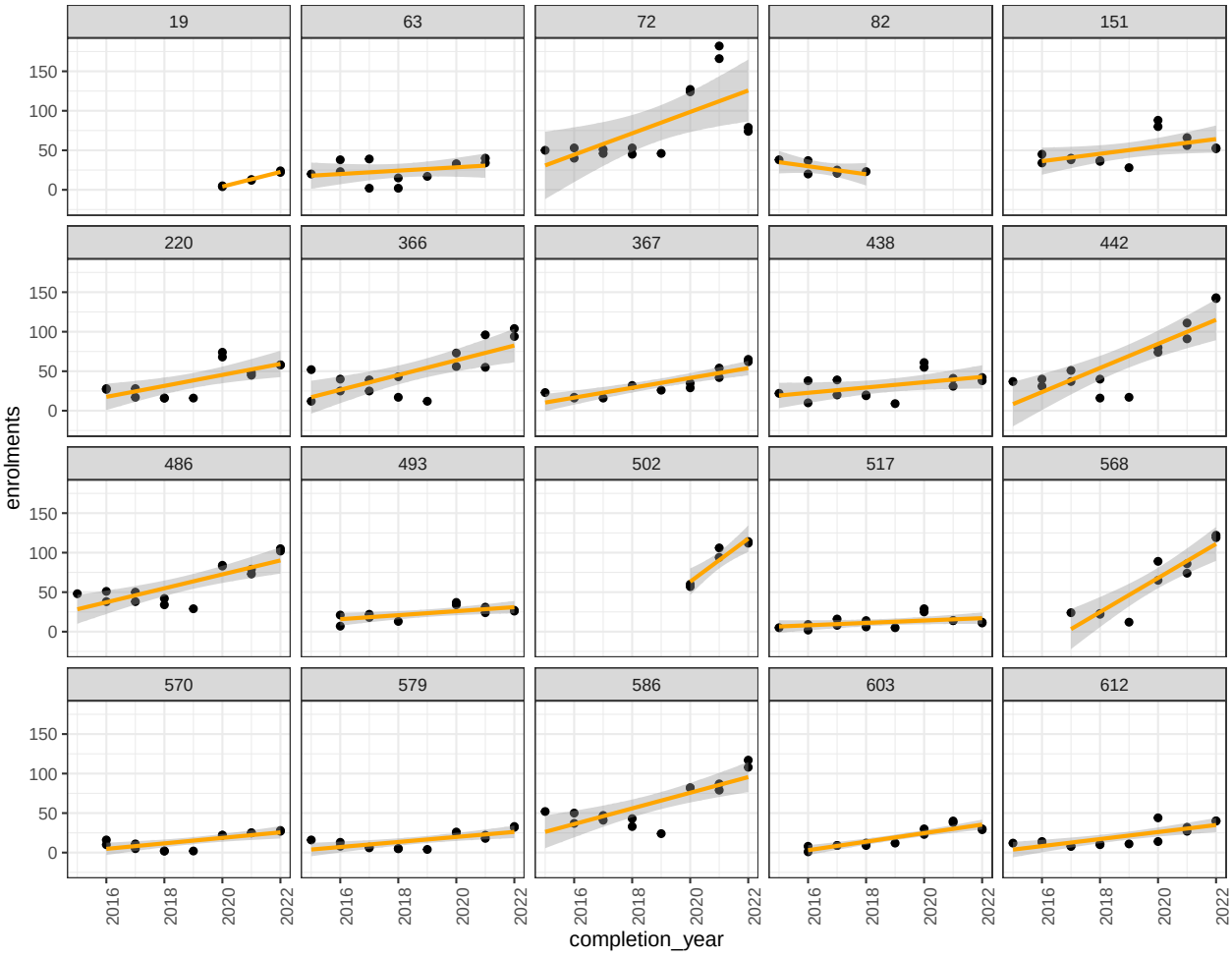


Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Aquatic Practices

As shown previously (Figure **??**), Aquatic Practices is a relatively new subject introduced in 2015. The plot above (Figure 1) displays a linear model fitted for each school (explained in step 1), primarily to provide an at-a-glance visualisation to the enrolment trends in each school.

In the randomly selected schools, the various school sizes is distinct, where there were relatively larger schools such as school 72 and 442, which showed enrolments of over 100 students while schools 570 and 579 consistently had enrolments below 50 for each year. Some schools showed a stark increase in enrolments (*e.g.* school 72 and 442), while some school showed rather constant growth enrolments (*e.g.* School 493 and 517).

Getting the data ready for modelling

Removing graduating cohort 2019

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. For these reasons, all completion years with zero enrolments will also be removed for modelling.

Linearise response variable using log transformation

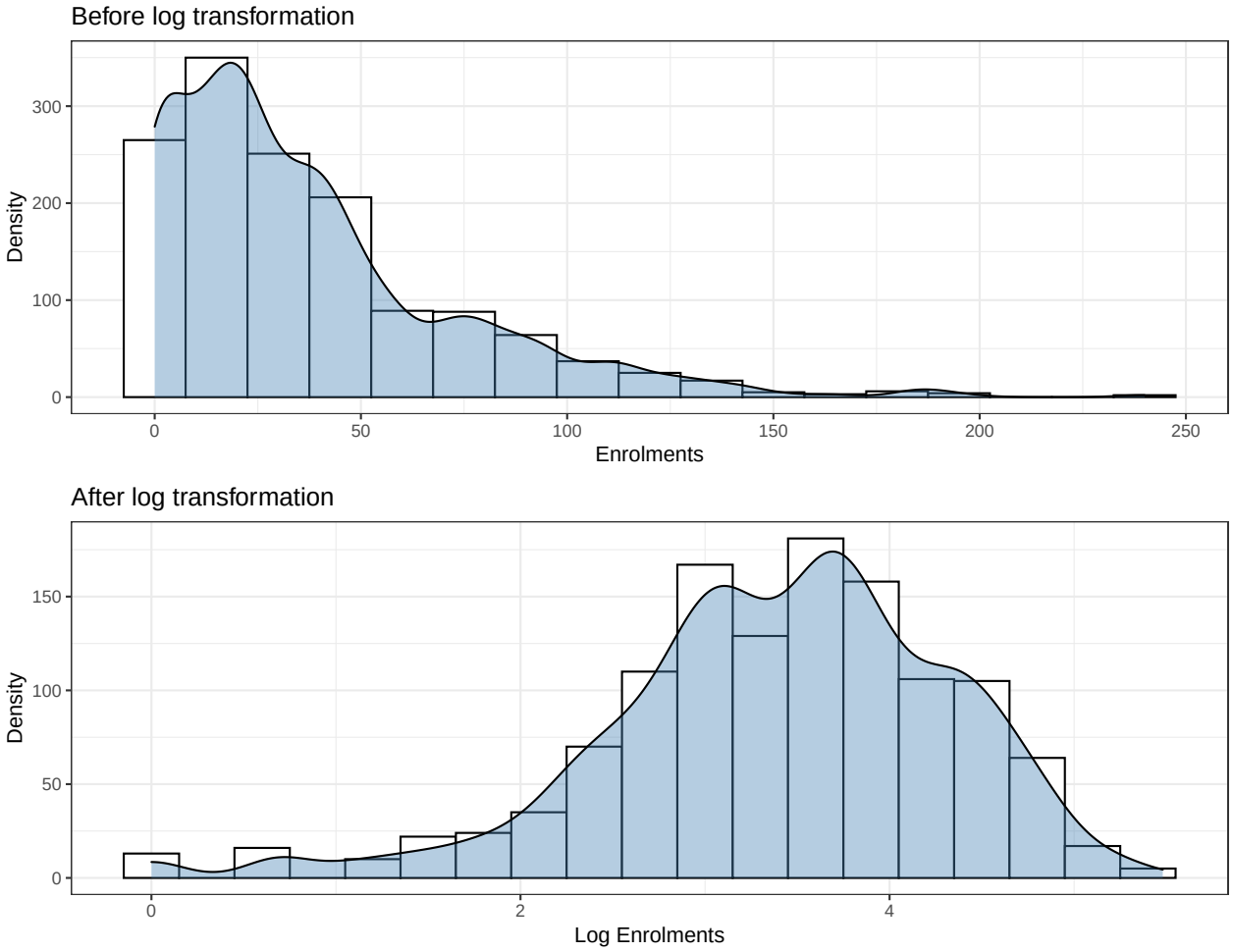


Figure 2: Effects of log transformation for response variable (enrolments) in Aquatic Practices

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 Aquatic Practices

	df	AIC
Model0.0: Within schools	3	2505.504
Model0.2: Schools nested within districts	4	2506.955
Model0.1: Schools nested within postcodes	4	2507.105

As underlined 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 best model and will be used in the subsequent analysis.

Intraclass correlation (ICC)

```
summary(model0.0)
```

```
## Random effects:

##  Groups         Name        Variance Std.Dev.
##  qcaa_school_id (Intercept) 0.69382  0.83296
##  Residual                   0.38336  0.61916

##
##  Fixed effects:

##             Estimate Std. Error t value
## (Intercept) 3.308375 0.07911121 41.8193

##
##  Number of schools (level-two group) = 120
##  Number of district (level-three group) = NA
```

This model takes into account 120 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^*:

$$\text{Level-two ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.6938}{(0.6938 + 0.3834)} = 0.6441$$

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, 64.41% of the total variability is attributable to the differences in enrolments within schools at different time periods.

Unconditional growth model

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

```
##  Groups         Name        Variance  Std.Dev. Corr
##  qcaa_school_id (Intercept) 0.9608954 0.980253
##                 year15      0.0069467 0.083347 -0.534
##  Residual                   0.1964707 0.443250

##              Estimate Std. Error  t value
## (Intercept) 2.5658807 0.09690281 26.47891
## year15      0.1791002 0.01069684 16.74328

##  Number of Level Two groups =  120
##  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.

- $\pi_{0ij} = 2.5659$: Initial status for school i (*i.e.* expected log enrolments when time = 0)
- $\pi_{1ij} = 0.0353$: Growth rate for school i
- $\epsilon_{tij} = 0.2458$: Variance in within-school residuals after accounting for linear growth overtime

When the subject was first introduced in 2015, schools were expected to have an average of 13.0124 ($e^{2.56588}$) enrolments. On average, the enrolments were expected to increase by 19.614% ($(e^{0.0353} - 1) \times 100$) per year. The estimated within-school variance decreased by 48.75% (0.3834 to 0.1965), indicating the 48.75% 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
model4.5	12	1869.409	1929.989	-922.7043
model4.1	14	1871.986	1942.663	-921.9928
model4.4	11	1874.380	1929.912	-926.1898
model4.0	16	1874.525	1955.299	-921.2624
model4.7	13	1876.474	1942.103	-925.2370
model4.3	10	1881.588	1932.072	-930.7942
model4.6	12	1884.298	1944.879	-930.1491
model4.9	11	1888.825	1944.357	-933.4124
model4.10	11	1888.825	1944.357	-933.4124
model4.2	11	1888.825	1944.357	-933.4124
model4.8	11	1888.825	1944.357	-933.4124

As summarised 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. `model4.5` (Table 2) appears to have the optimal (smallest) AIC, 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)
10	1907.775	1958.259	-943.8874	1887.775	NA	NA	NA
12	1869.409	1929.989	-922.7043	1845.409	42.36629	2	0

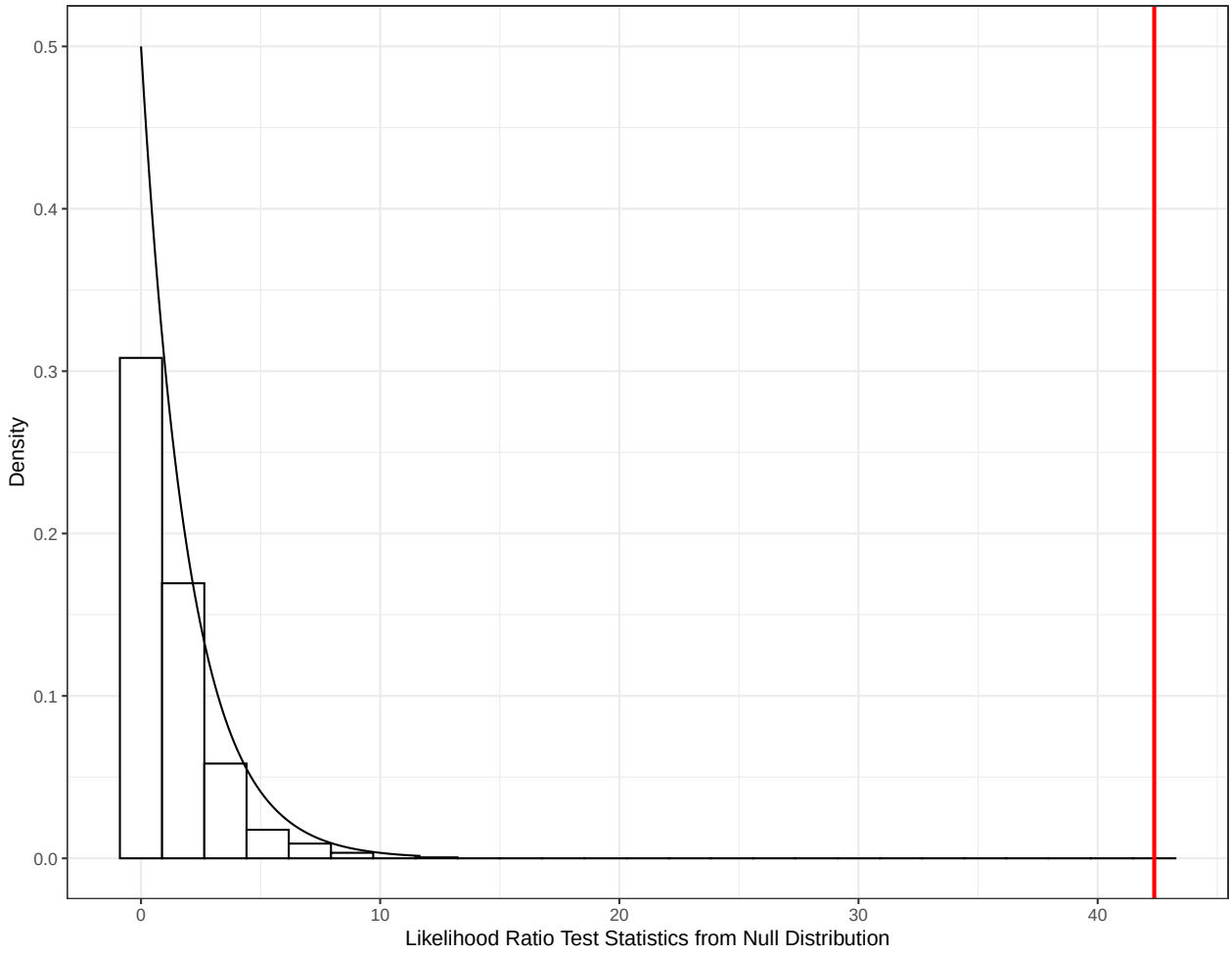


Figure 3: Histogram of likelihood ratio test statistic, with a red vertical line indicating the likelihood ratio test statistic for the actual model

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 (as indicated by the red line in Figure 3. There is overwhelming statistical evidence ($\chi^2 = 42.3663$ and p-value $= 0$ – see 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	0.7694684	1.0336552
cor_year15.(Intercept)|qcaa_school_id	-0.7092237	-0.3083679
sd_year15|qcaa_school_id	0.0557664	0.0927001
sigma	0.4187891	0.4583143
(Intercept)	2.3796251	3.5004024
year15	-0.0076337	0.1346439
sectorGovernment	-0.7578332	0.4846176
sectorIndependent	-2.4092098	-0.8681366
unityear_12_enrolments	-0.3080638	-0.1124896
year15:sectorGovernment	0.0376980	0.1846632
year15:sectorIndependent	0.0957980	0.2969844
year15:unityear_12_enrolments	0.0095620	0.0524932

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).

Composite model

- Level one (measurement variable)

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} year15_{tij} + \epsilon_{tij}$$

- Level two (schools within postcodes)

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

Therefore, the composite model can be written as

$$\begin{aligned} Y_{tij} &= \pi_{0ij} + \pi_{1ij} year15_{tij} + \epsilon_{tij} \\ &= (\beta_{00j} + \beta_{01} sector_{ij} + \beta_{02} unit_{ij} + u_{0ij}) + (\beta_{10j} + \beta_{11j} sector_{ij} + \beta_{12j} unit_{ij} + u_{1ij}) year15_{tij} + \epsilon_{tij} \\ &= [\beta_{00j} + \beta_{01} sector_{ij} + \beta_{02} unit_{ij} + \beta_{10j} year15_{tij} + \beta_{11j} sector_{ij} year15_{tij} + \beta_{12j} unit_{ij} year15_{tij}] [u_{0ij} + u_{1ij} + \epsilon_{tij}] \end{aligned}$$

Fixed effects

```
summary(model_f)
```

```
##  Groups         Name        Variance Std.Dev. Corr
##  qcaa_school_id (Intercept) 0.802676 0.895922
##                 year15      0.005632 0.075047 -0.537
##  Residual                   0.192796 0.439085

##                                    Estimate Std. Error    t value
## (Intercept)                     2.96690311 0.28548780 10.3923990
## year15                          0.05859201 0.03459941  1.6934394
## sectorGovernment               -0.14916989 0.30209047 -0.4937921
## sectorIndependent              -1.64252982 0.39342178 -4.1749844
## unityear_12_enrolments         -0.20422624 0.05216529 -3.9149833
## year15:sectorGovernment         0.11332298 0.03601076  3.1469199
## year15:sectorIndependent        0.19500364 0.04787424  4.0732474
## year15:unityear_12_enrolments   0.03002428 0.01136349  2.6421712

##  Number of Level Two groups =  120
##  Number of Level Three groups =  NA
```

Based on the model output, the estimated mean enrolments for government schools are estimated to be 13.8577% ($(e^{-0.1491699}-1)\times 100$) less than that of catholic schools when the subject was first introduced in 2015. Government schools are also expected to have an average increase of 18.7577% ($(e^{0.0585920+0.1133230}-1)\times 100$), 11.9994% ($(e^{0.1133230}-1)\times 100$) more than that of catholic schools per year, .

On the other hand, independent schools are estimated to have an average enrolments of 80.651% ($(e^{-1.6425298}-1)\times 100$) less than that of catholic schools. However, this small initial status is matched with a 28.8651% ($e^{.0585920+0.1950036}-1)\times 100$ increase in enrolments per year, on average. This increase is 21.5315% ($(e^{0.1950036}-1)\times 100$) greater than that of catholic schools.

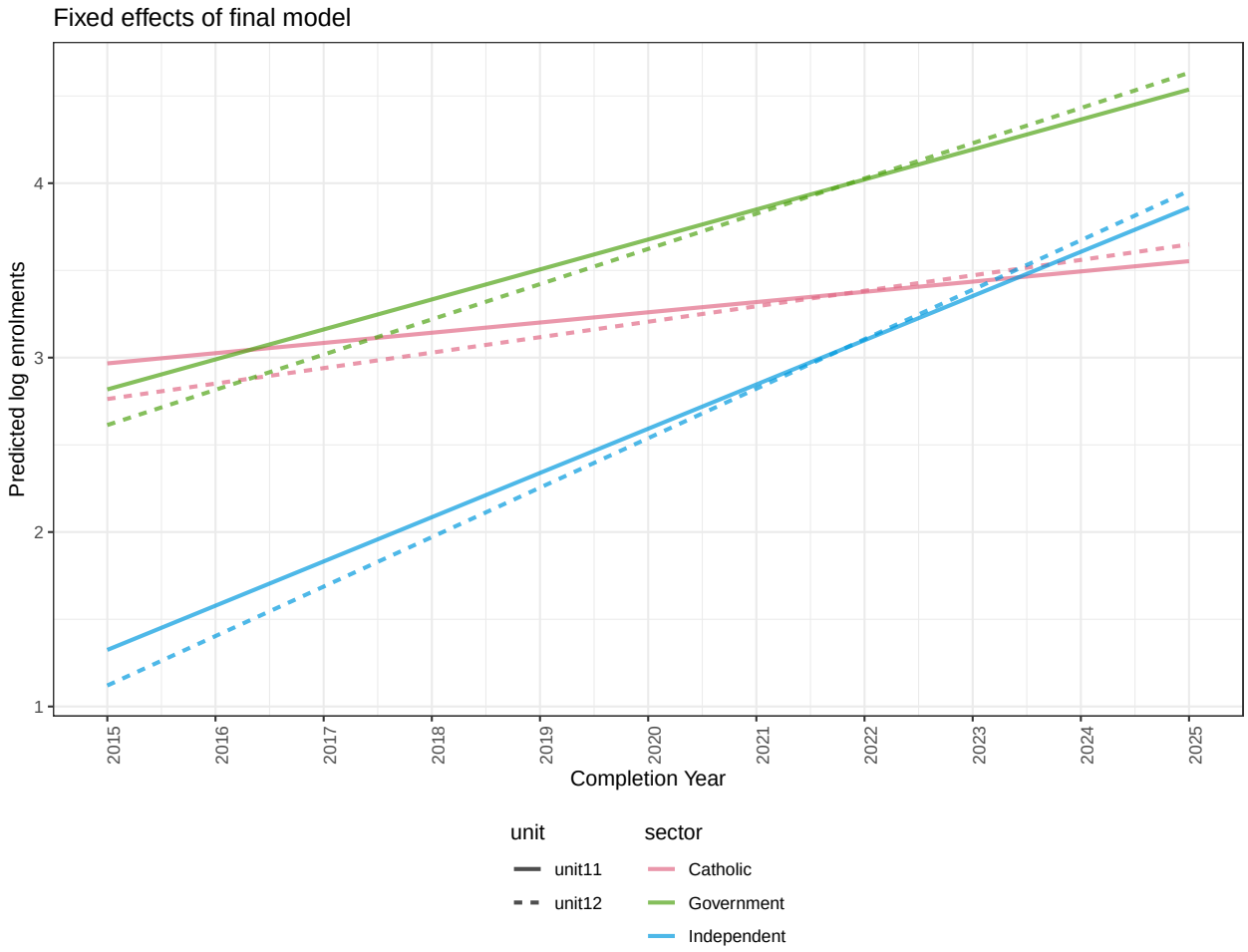
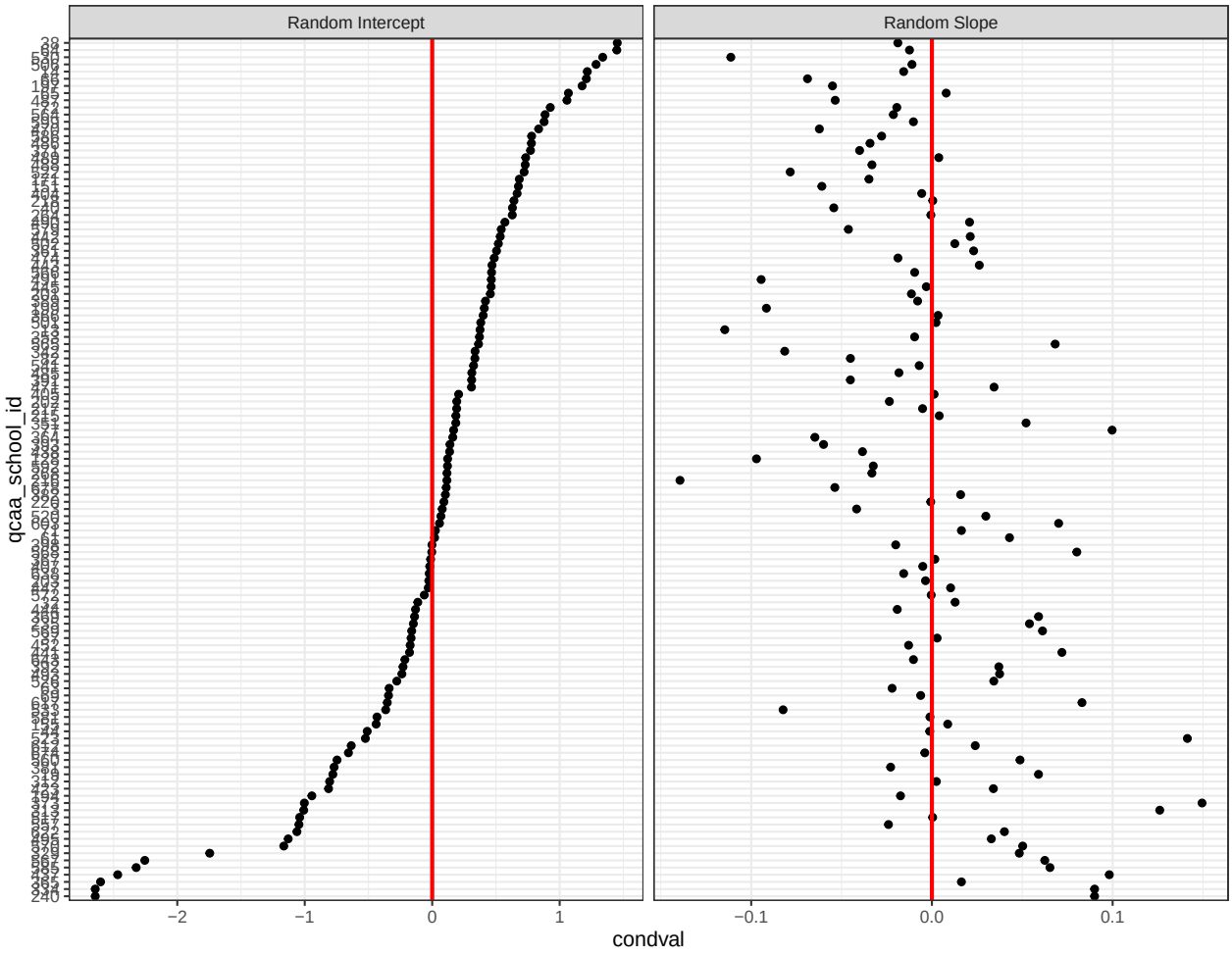


Figure 4: Fixed effects of the final model for Aquatic Practices subject

The fixed effects can be visualised in Figure 4. As mentioned, catholic schools had the highest enrolments score in 2015, but had a slow increase in enrolments over the years; By 2018, government schools had greater enrolments, on average than catholic schools, and it is expected that independent schools have larger enrolments numbers, on average, than catholic schools. In all sectors, unit 12 appears to increase at a higher rate than that of unit 11, as shown by the steeper slope.

Random effects



In the random effects, there is a some negative correlation between the random intercept and slope, where schools with lower enrolments when subject was first introduced are matched with a higher increase (decrease) in enrolments over the years. However, this negative relationship was not relatively strong, where the correlation between the random intercept and slope is only at -0.54, as shown in the model output.

Predictions

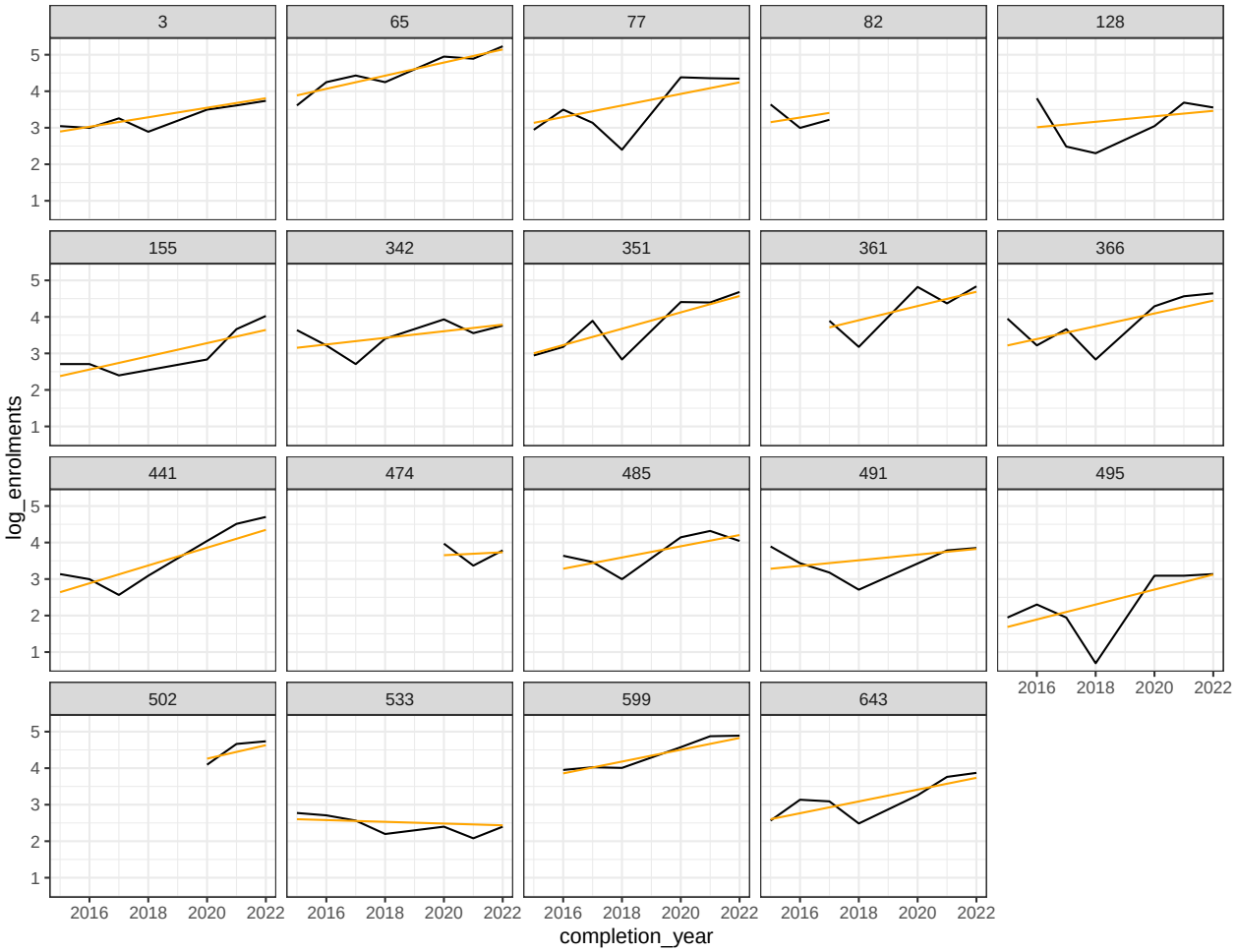


Figure 5: Model predictions for year 11 enrolments for 20 randomly selected schools

Figure 5 above shows the predictions for 20 randomly selected schools.