Multilevel Model for Marine Science

Brendi Ang

30/10/2021

Contents

Marince Science

Exploring the dataset with basic linear model for each school

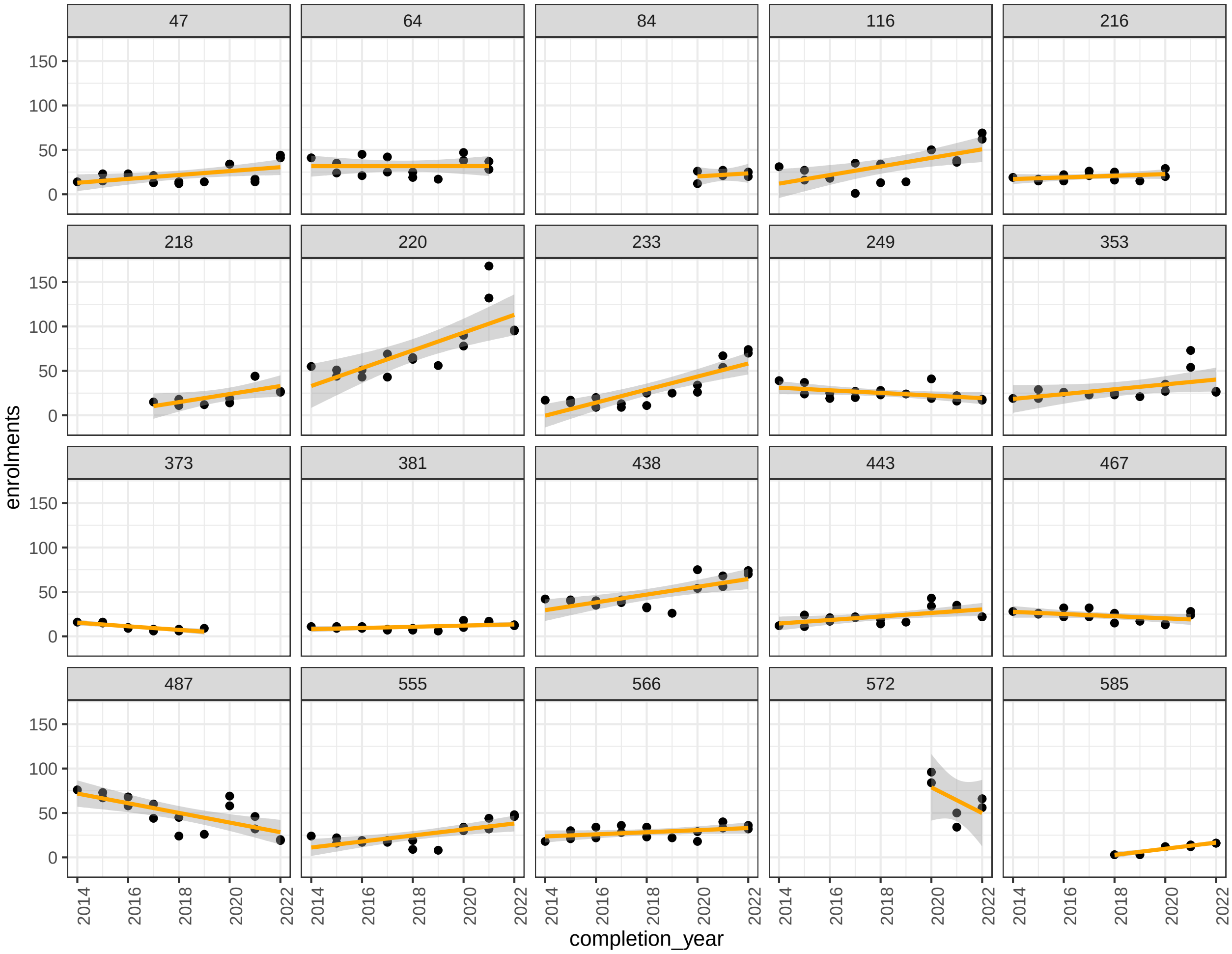


Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Marine Science

Figure 1 shows a basic linear plot for 20 randomly selected schools. The subject was introduced in 2014, and some school offered the subjects at later years while some schools discontinued the subject (*e.g.* school 373). Next, some schools (*e.g.* school 220 and 233) showed a large increase in enrolments relative to the other schools while some schools showed a decrease in enrolments over the years. The various school sizes can be seen, where some schools have less than 25 enrolments each year while some school have over 50 enrolments per cohort.

Getting the data ready for modelling

Removing zero enrolments

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

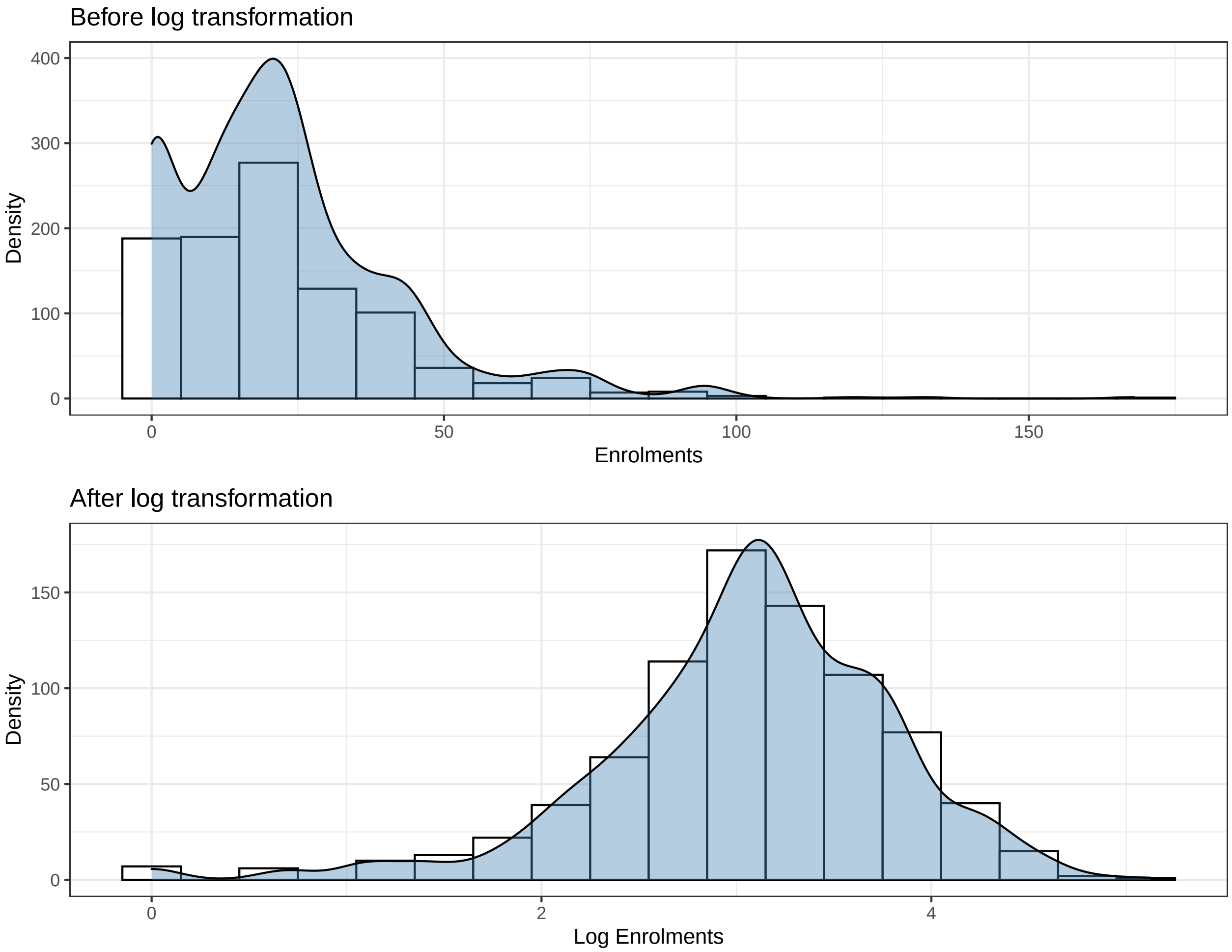


Figure 2: Effects of log transformation for response variable (enrolments) in Marine Science

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 Specialist Mathematics

	df	AIC
Model0.0: Within schools	3	1261.849
Model0.2: Schools nested within districts	4	1262.921
Model0.1: Schools nested within postcodes	4	1263.849

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)

```
## Random effects:

##  Groups         Name        Variance Std.Dev.
##  qcaa_school_id (Intercept) 0.44111  0.66416
##  Residual                   0.21735  0.46620

##
##  Fixed effects:

##             Estimate Std. Error  t value
## (Intercept) 2.955045 0.07678718 38.48358

##
##  Number of schools (level-two group) = 81
##  Number of district (level-three group) = NA
```

This model takes into account 81 schools which offer the subject. 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^*:

$$\text{level-two ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.4411}{(0.4411 + 0.2173)} = 0.6700$$

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, 67% 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) 0.4715552 0.686699
##                 year15      0.0071891 0.084788 -0.284
##  Residual                   0.1628130 0.403501
```

```
##             Estimate Std. Error  t value
## (Intercept) 2.7409337 0.08744009 31.34642
## year15      0.0464796 0.01279031  3.63397
```

```
##  Number of Level Two groups =  81
##  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.

- $\pi_{0ij} = 2.7409$: Initial status for school i (*i.e.* expected log enrolments when time = 0)
- $\pi_{1ij} = 0.0465$: Growth rate for school i
- $\epsilon_{tij} = 0.1628$: 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 15.5009 ($e^{2.7409}$) enrolments. On average, the enrolments were expected to increase by 4.7598% ($(e^{0.0465} - 1) \times 100$) per year. The estimated within-school variance decreased by 25.07% (0.2173 to 0.162813), indicating the 25.07% 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

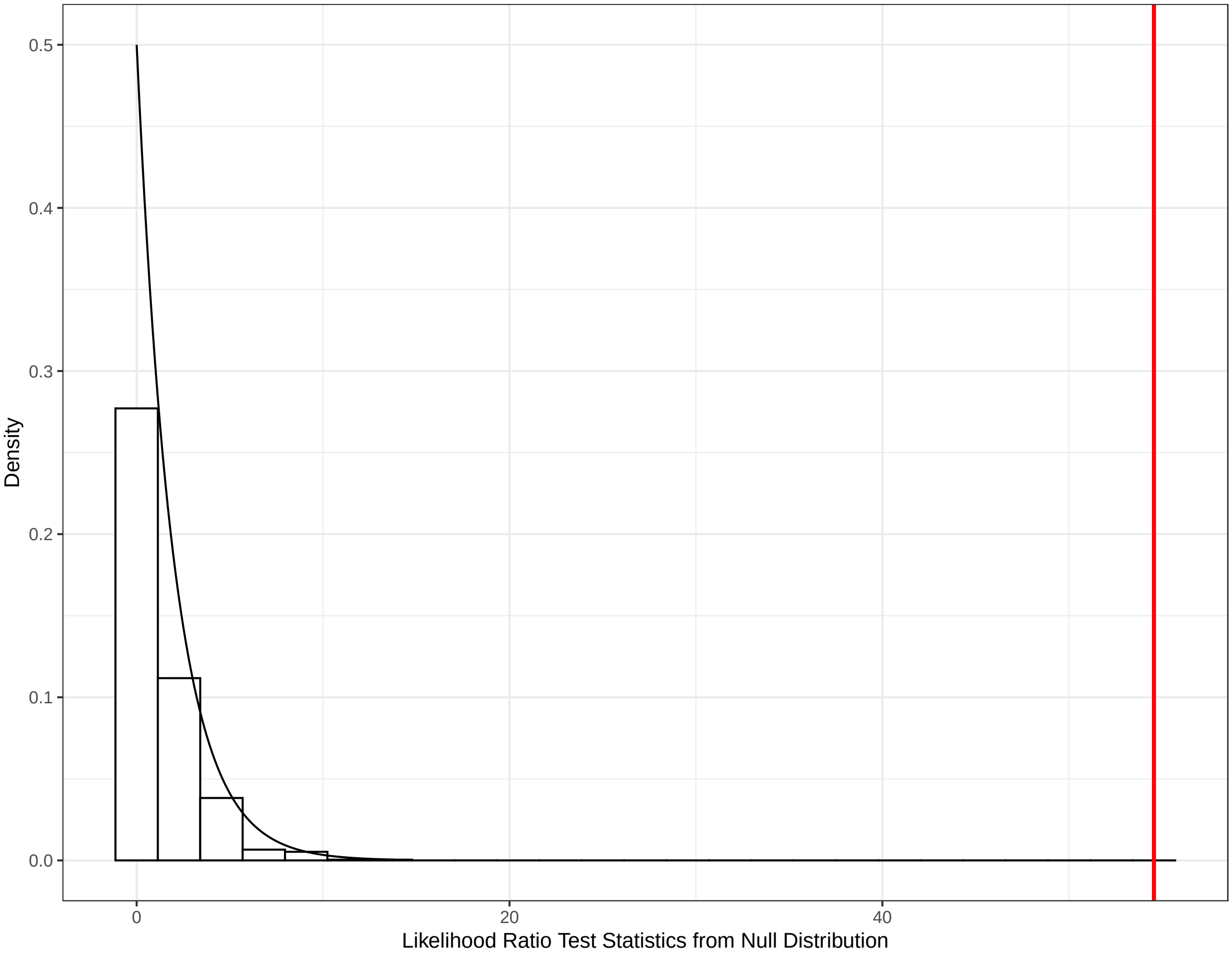
model	AIC
model4.4	1127.338
model4.2	1128.333
model4.8	1128.333
model4.10	1128.333
model4.9	1128.333
model4.3	1128.400
model4.7	1128.990
model4.5	1129.327
model4.6	1130.333
model4.1	1130.989
model4.0	1134.839

As summarise in step 6, level-two predictors `secotr` 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. `model4.4` (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)
9	1177.905	1219.838	-579.9523	1159.905	NA	NA	NA
11	1127.338	1178.590	-552.6690	1105.338	54.56667	2	0



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 **??**. There is overwhelming statistical evidence ($\chi^2 = 54.5667$ 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	0.5516981	0.8199168
cor__year15.(Intercept)|qcaa__school__id	-0.6355994	-0.0908638
sd__year15|qcaa__school__id	0.0604489	0.1065415
sigma	0.3809136	0.4236035
(Intercept)	2.4156300	3.2727234
year15	-0.0301446	0.0911383
sectorGovernment	-0.4283699	0.5614812
sectorIndependent	-0.9822681	0.1326594
unityear__12__enrolments	-0.1255390	-0.0104000
year15:sectorGovernment	-0.0252647	0.1139009
year15:sectorIndependent	-0.0999005	0.0667181

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 4), 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)

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

- Level three (districts)

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

Therefore, the composite model can be written as

$$\begin{aligned}
Y_{tij} &= \pi_{0ij} + \pi_{1ij} year92_{tij} + \epsilon_{tij} \\
&= (\beta_{00j} + \beta_{01j} sector_{ij} + \beta_{02j} unit_{ij} + u_{0ij}) + (\beta_{10j} + \beta_{11j} sector_{ij} + u_{1ij}) year15_{tij} + \epsilon_{tij} \\
&= \ [\gamma_{000} + r_{00j} + (\gamma_{010} + r_{01j}) sector_{ij} + (\gamma_{020} + r_{02j}) unit_{ij} + u_{0ij}] + \\
&\quad\ [\gamma_{100} + r_{10j} + (\gamma_{110} + r_{11j}) sector_{ij}]\, year15_{tij} + \epsilon_{tij} \\
&= \ [\gamma_{000} + \gamma_{010} sector_{ij} + \gamma_{020} unit_{ij} + \gamma_{100} year15_{tij} + \gamma_{110} sector_{ij} year15_{tij}] + \\
&\quad\ [r_{00j} + r_{01j} sector_{ij} + r_{02j} unit_{ij} + r_{10j} year15_{tij} + r_{11j} sector_{ij} year15_{tij} + u_{0ij} + \epsilon_{tij}]
\end{aligned}$$

Fixed effects

```
## Random effects:
```

```
##  Groups         Name        Variance  Std.Dev. Corr
##  qcaa_school_id (Intercept) 0.4666221 0.683097
##                 year15      0.0069507 0.083371 -0.425
##  Residual                   0.1615822 0.401973
```

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

```
##                              Estimate Std. Error    t value
## (Intercept)                2.84400284 0.21194908 13.4183310
## year15                     0.02927534 0.03094077  0.9461739
## sectorGovernment           0.03976737 0.23892386  0.1664437
## sectorIndependent         -0.44995296 0.28698959 -1.5678372
## unityear_12_enrolments    -0.06749144 0.02913496 -2.3165101
## year15:sectorGovernment    0.03977166 0.03469870  1.1462002
## year15:sectorIndependent -0.01081485 0.04262658 -0.2537113
```

```
##
##  Number of schools (level-two group) = 81
##  Number of district (level-three group) = NA
```

Based on the model output (see detailed explanation of fixed effects in step 9), the estimated mean enrolments for government schools are estimated to increase by 7.1487% ($(e^{0.0292753+0.0397717}-1)\times 100$) each year, which is 4.0573% ($(e^{0.0397717}-1)\times 100$) more than that of catholic schools.

On the other hand, independent schools are estimated to have a 1.8632% ($(e^{0.0292753-0.0108148}-1)\times 100$) increase in enrolments per year, on average. This increase is 1.0757% ($(e^{-0.0108148})-1)\times 100$) less than that of catholic schools.

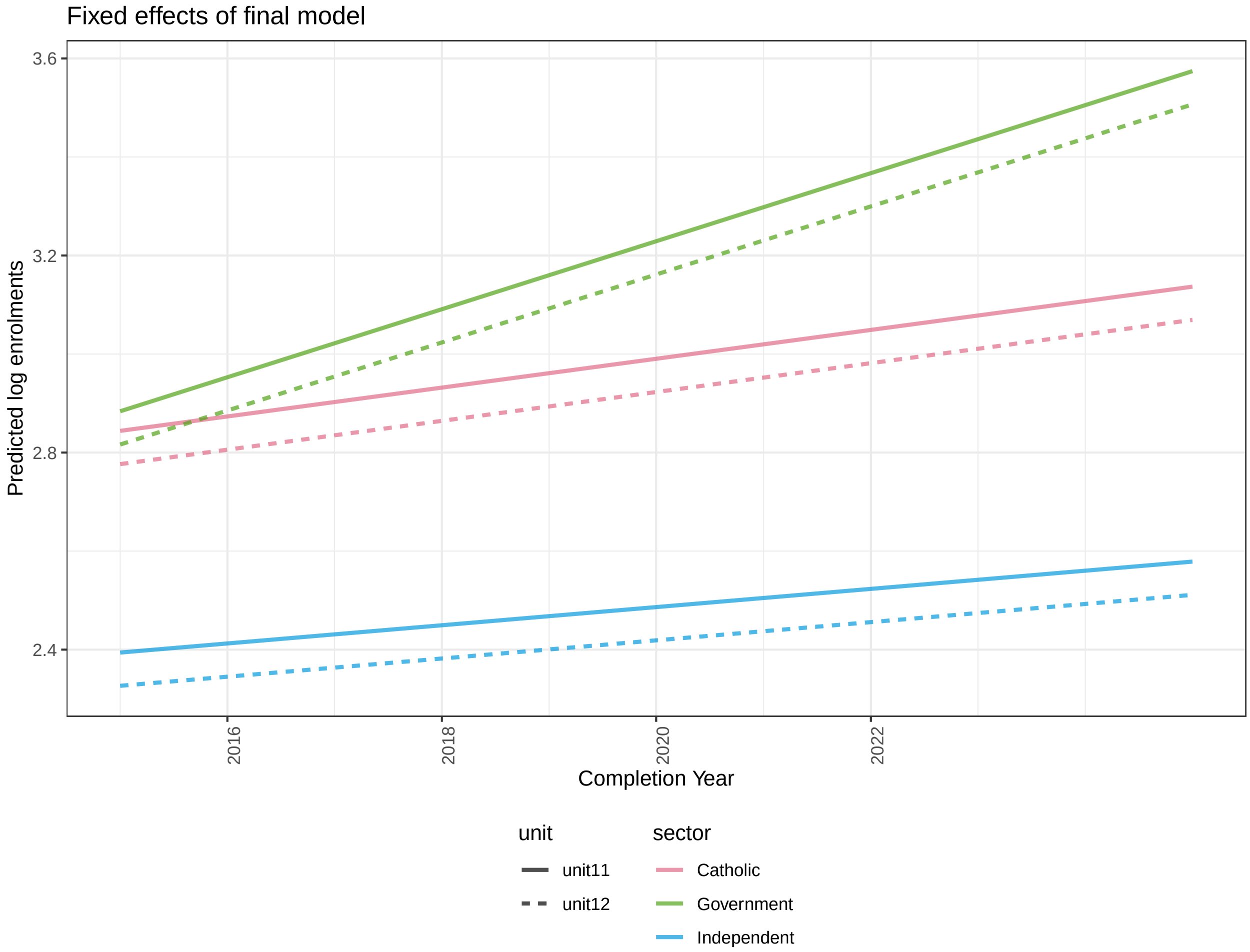


Figure 3: Fixed effects of the final model for Marine Science subject

The fixed effects are better accentuated in Figure 3. This figure shows that government schools are expected to have the largest increase in enrolments over the years, on average. For all sectors, it appears that unit 12 enrolments are consistently less than that of unit 11 enrolments, which may suggests that students are not pursuing year 12 after completing year 11 syllabus.

Random effects

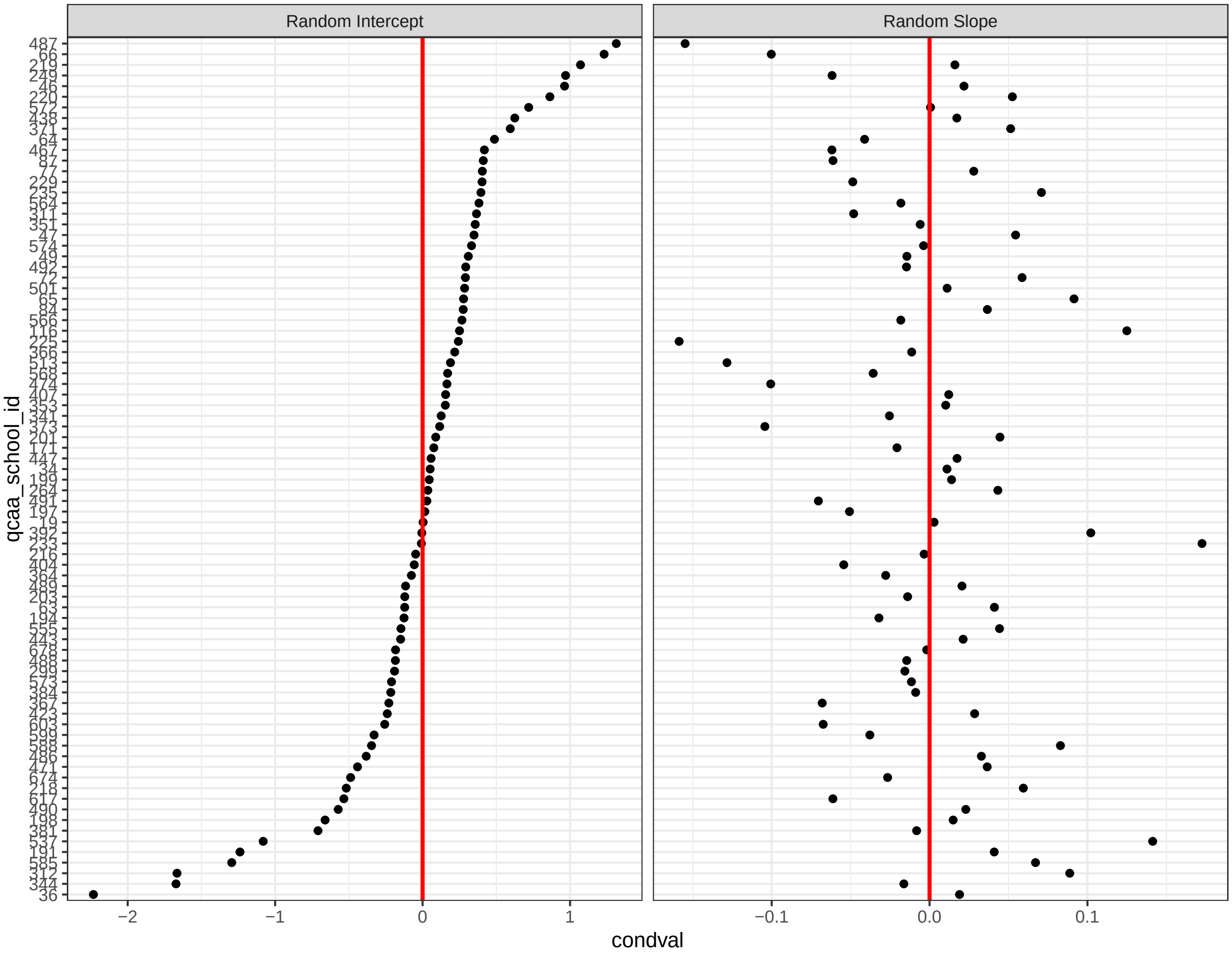


Figure 4: Random effects for all schools

Figure 4 represents the random intercept and slope for the random effects for a given school. It is manifest that there are no clear relationship between the random intercept and slope. Some schools with low enrolments when the subject was first introduced the school saw a large increase in enrolments over the years, while some showed a sharp decrease.

Predictions

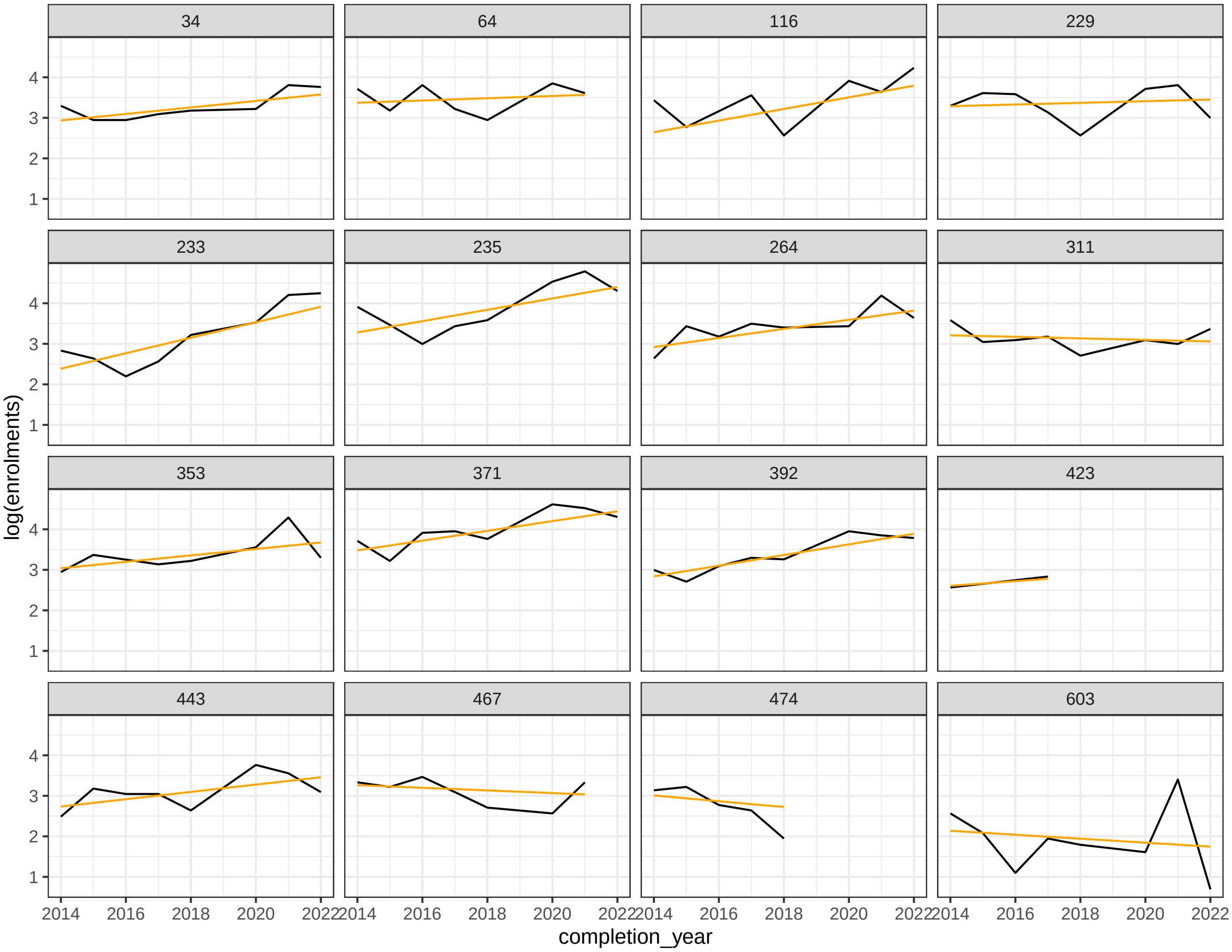


Figure 5: Model predictions for 20 randomly selected schools

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