Multilevel Model for Mathematical Methods

Brendi Ang

17/10/2021

Contents

Mathematical Methods

Exploring the dataset with basic linear model

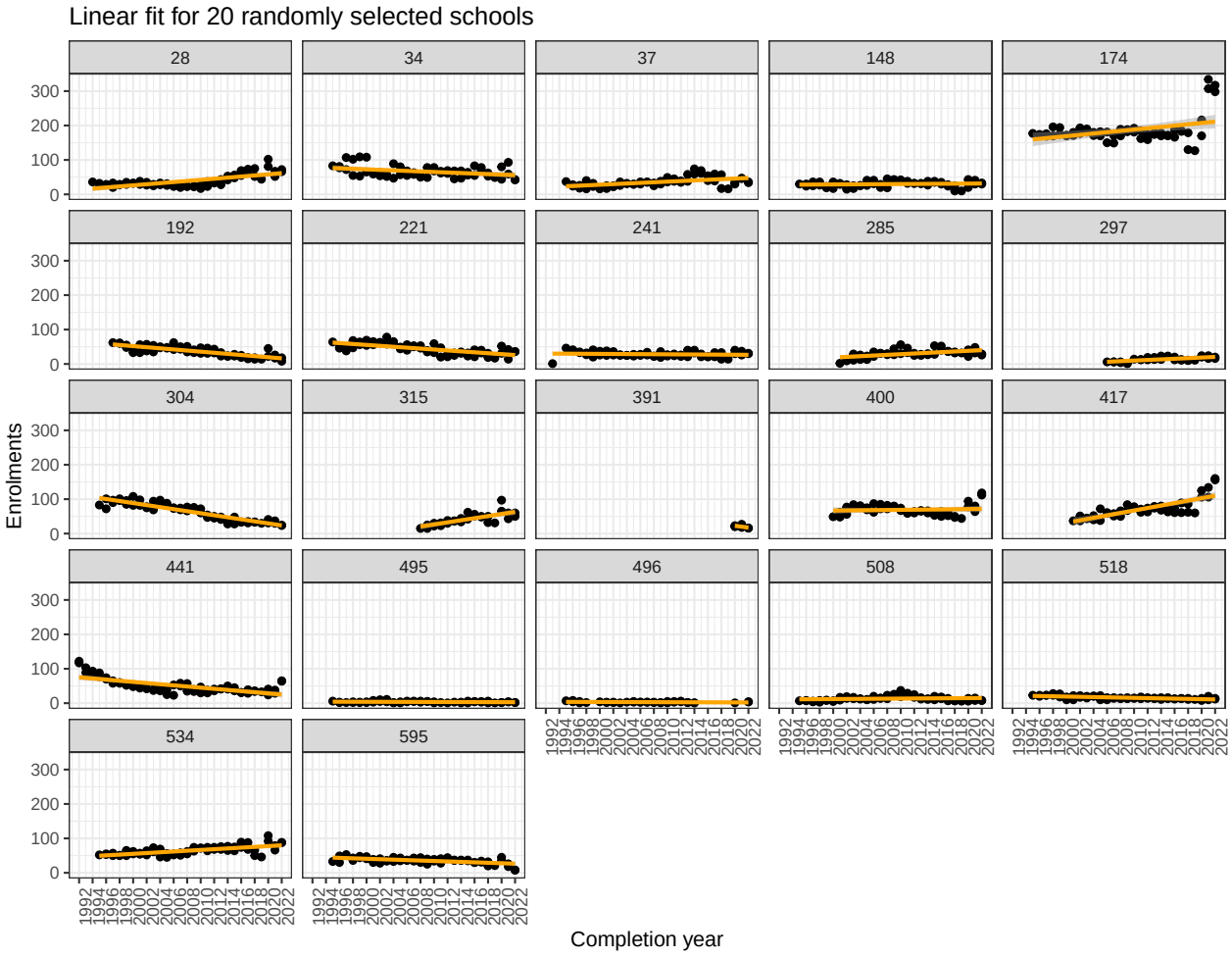

Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Mathematical Methods

Figure 1 fits a linear model for 20 randomly selected schools. As enrolments are plotted on the same y-axis, the different sizes of school is apparent. For instance, school 174 has more than 100 enrolments for mathematical methods for each cohort, while school 496 and 518 have little enrolments (< 50) for each cohort. Each school showed rather distinct trend, where school 441 and 304 showed a decrease in enrolments while schools 28, 174, 417 showed an increase in enrolments, on average. It can also be observed that school 391 (St Mary's College) only introduced the subject in 2020, as it was a relatively new school.

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

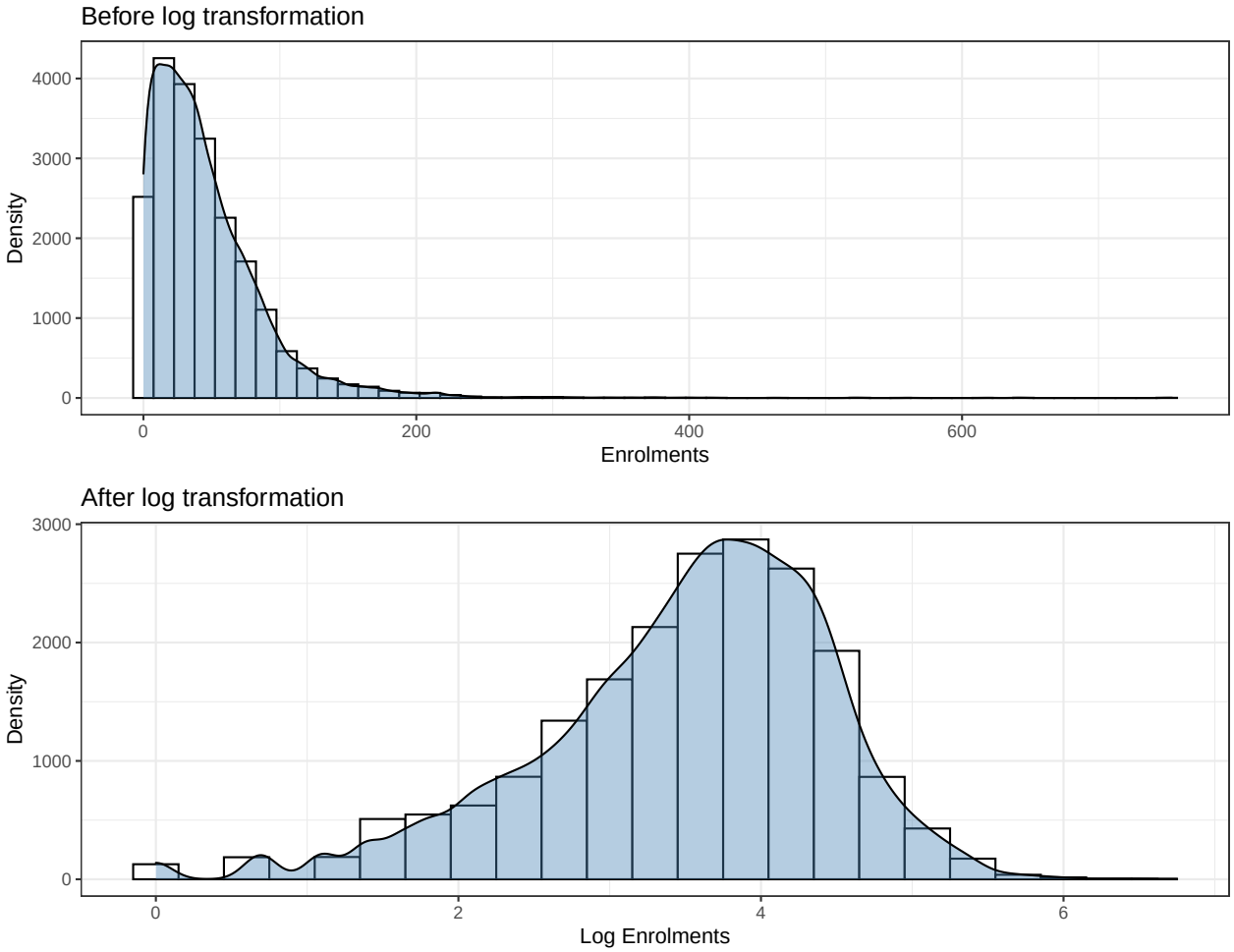

Figure 2: Effects of log transformation for response variable (enrolments) in Mathematical Methods

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 Mathematical Methods

	df	AIC
Model0.2: Schools nested within districts	4	24303.39
Model0.1: Schools nested within postcodes	4	24333.96
Model0.0: Within schools	3	24346.27

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) 0.83408  0.91328
##  qcaa_district                (Intercept) 0.14864  0.38554
##  Residual                                 0.17998  0.42424

##
##  Fixed effects:

##             Estimate Std. Error  t value
## (Intercept) 3.305797  0.1152798 28.67629

##
##  Number of schools (level-two group) = 466
##  Number of district (level-three group) = 13
```

This model will takes into account 466 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 $t^* \neq t$:

$$\text{Level-two ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.83408}{(0.83408 + 0.14864 + 0.17998)} = 0.7174$$

This can be conceptualised as the correlation between enrolments of two random draws from the same school at two different years. In other words, 71.74% 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 school j.

$$\text{Level-three ICC} = \frac{\phi_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.14864}{(0.83408 + 0.14864 + 0.17998)} = 0.1278$$

Similarly, this can be conceptualised as the correlation between enrolments of two randomly selected schools from the same district – *i.e.* 12.78% 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) 2.0601e+00 1.4353101
##                               year92      1.8682e-03 0.0432225 -0.753
##  qcaa_district                (Intercept) 6.9225e-02 0.2631056
##                               year92      5.9806e-05 0.0077335 1.000
##  Residual                                 1.2387e-01 0.3519528

##                Estimate  Std. Error   t value
## (Intercept) 3.066187193 0.100976610 30.365321
## year92      0.009110851 0.003051126  2.986062

##  Number of Level Two groups =  466
##  Number of Level Three groups =  13
```

- $\pi_{0ij} = 2.7409$: Initial status for school i in district j (*i.e.* expected log enrolments when time = 0)
- $\pi_{1ij} = 0.0465$: Growth rate for school i in district j
- $\epsilon_{tij} = 0.162813$: Variance in within-school residuals after accounting for linear growth overtime

When the subject was first introduced in 1992, schools were expected to have 15.5009 ($e^{1.7848}$) enrolments, on average. Enrolments were expected to increase by 4.7598% ($(e^{0.0465} - 1) \times 100$) per year. The estimated within-schools variance decreased by -9.5556% (0.1800 to 0.1628), implying that -9.5556% of within-school variability can be explained by the linear growth over time.

Dealing with boundary constraint

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 (@beyond-mlr).

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

```
##  Groups                          Name        Variance  Std.Dev. Corr
##  qcaa_district:qcaa_school_id (Intercept) 2.0930666 1.446743
##                               year92      0.0019446 0.044097 -0.757
##  qcaa_district                (Intercept) 0.2039055 0.451559
##  Residual                                 0.1238374 0.351905
```

```
##                Estimate Std. Error   t value
## (Intercept) 3.053472289 0.14377570 21.237749
## year92      0.009675574 0.00220878  4.380505
```

```
##  Number of Level Two groups =  466
##  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
model4.7	18176.17
model4.1	18177.29
model4.4	18177.87
model4.0	18178.63
model4.5	18178.71
model4.10	18238.84
model4.9	18238.84
model4.2	18238.85
model4.8	18238.85
model4.6	18239.98
model4.3	18241.58

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 `model4.7` (Table 2). The next section will test the selected model's random effects to build the final model.

Parametric bootstrap to test random effects

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.

Confidence interval

Table 3: 95% confidence intervals for fixed and random effects in the final model

var	2.5 %	97.5 %
sd_(Intercept)|qcaa_district:qcaa_school_id	1.2611487	1.4536475
cor_year92.(Intercept)|qcaa_district:qcaa_school_id	-0.7727809	-0.6789689
sd_year92|qcaa_district:qcaa_school_id	0.0370458	0.0430934
sd_(Intercept)|qcaa_district	0.2229889	0.6290453

var	2.5 %	97.5 %
sigma	0.3444005	0.3511923
(Intercept)	3.0378500	3.8143678
year92	-0.0002020	0.0197675
sectorGovernment	-0.3032420	0.4819675
sectorIndependent	-1.6557356	-0.8014642
unityear_12_enrolments	-0.1087469	-0.0620239
year92:sectorGovernment	-0.0257979	-0.0028113
year92:sectorIndependent	0.0132636	0.0383081
sectorGovernment:unityear_12_enrolments	-0.0550680	0.0005954
sectorIndependent:unityear_12_enrolments	-0.0385138	0.0211624

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)

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

- Level two (schools within districts) will contain new predictor(`sector`)

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

- Level three (districts)

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

Therefore, the composite model can be written as:

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

Fixed effects

```
summary(model_f)
```

```
##  Groups                        Name        Variance  Std.Dev. Corr
##  qcaa_district:qcaa_school_id (Intercept) 1.8467382 1.358947
##                               year92      0.0016073 0.040092 -0.729
##  qcaa_district                (Intercept) 0.2117140 0.460124
##  Residual                                 0.1210798 0.347965

##                                              Estimate  Std. Error    t value
## (Intercept)                                3.429618908 0.202380825 16.9463629
## year92                                     0.009468131 0.004804577  1.9706483
## sectorGovernment                           0.067490171 0.182073306  0.3706758
## sectorIndependent                         -1.246645798 0.198604067 -6.2770406
## unityear_12_enrolments                    -0.084742947 0.011705675 -7.2394753
## year92:sectorGovernment                   -0.013953478 0.005543237 -2.5172075
## year92:sectorIndependent                   0.026088675 0.006094448  4.2807280
## sectorGovernment:unityear_12_enrolments   -0.028631744 0.013514625 -2.1185748
## sectorIndependent:unityear_12_enrolments -0.009183082 0.015106467 -0.6078908

##  Number of Level Two groups =  466
##  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 school are expected to decrease by 0.4475% ($(e^{0.00946813-0.01395348}-1)*100$) each year, which is 1.3857% ($(e^{-0.01395348}-1)*100$) less than that of catholic schools. On the other hand, independent schools are predicted to have an increase of 3.6197% ($(e^{0.00946813+0.02608867}-1)*100$), which is 2.6432% more than that of catholic schools.

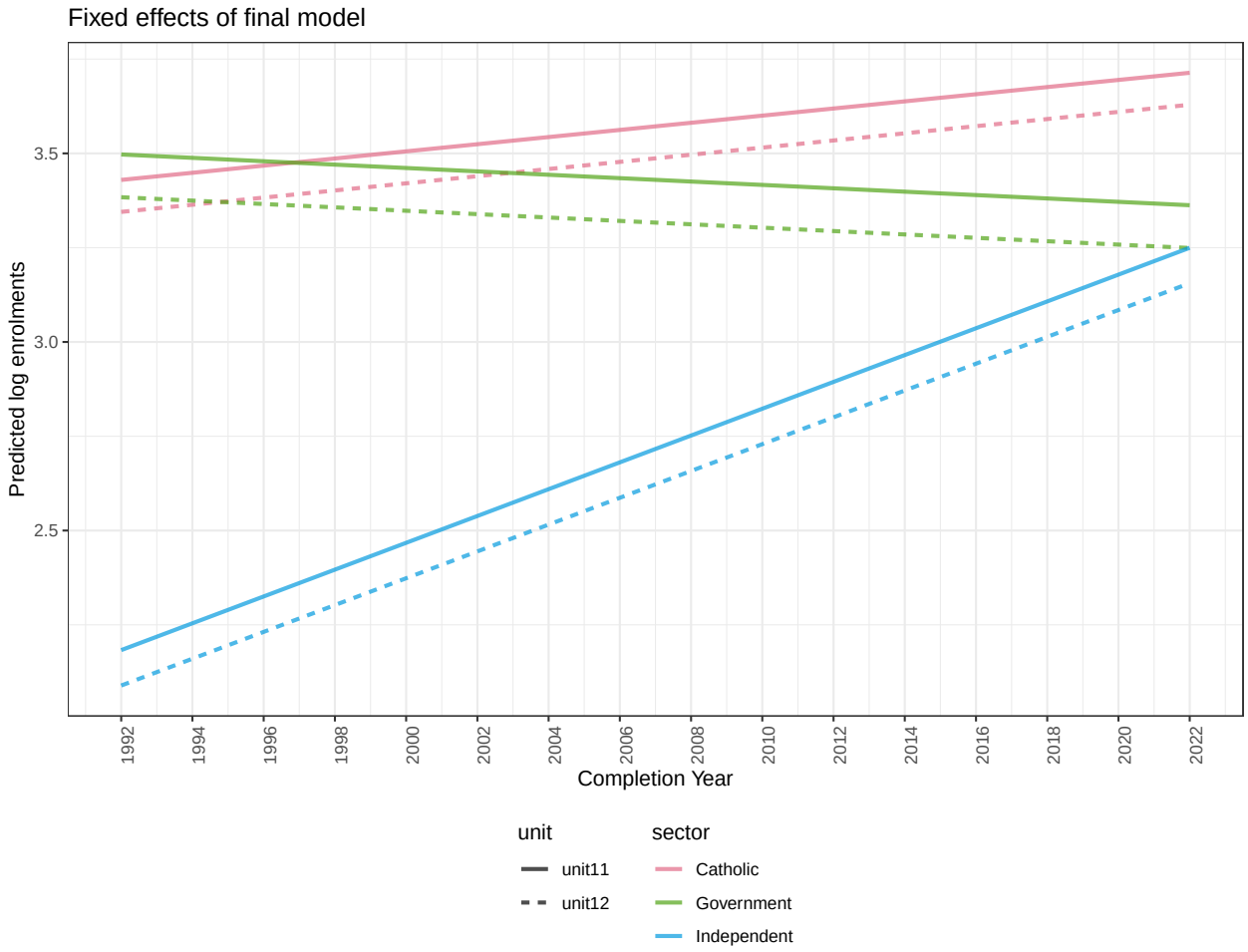

Figure 3: Fixed effects of the final model for Mathematical Methods

Based on the fixed effects (Figure 3), independent schools are expected to have the relatively highest increase in average enrolments across the years while government school showed a decrease in enrolments over the years. All sectors shows the year 11 enrolments are expected to be more than year 12 enrolments each year; This may indicate that students may be discontinuing the subject in year 11 after completing year 11 syllabus.

Random effects

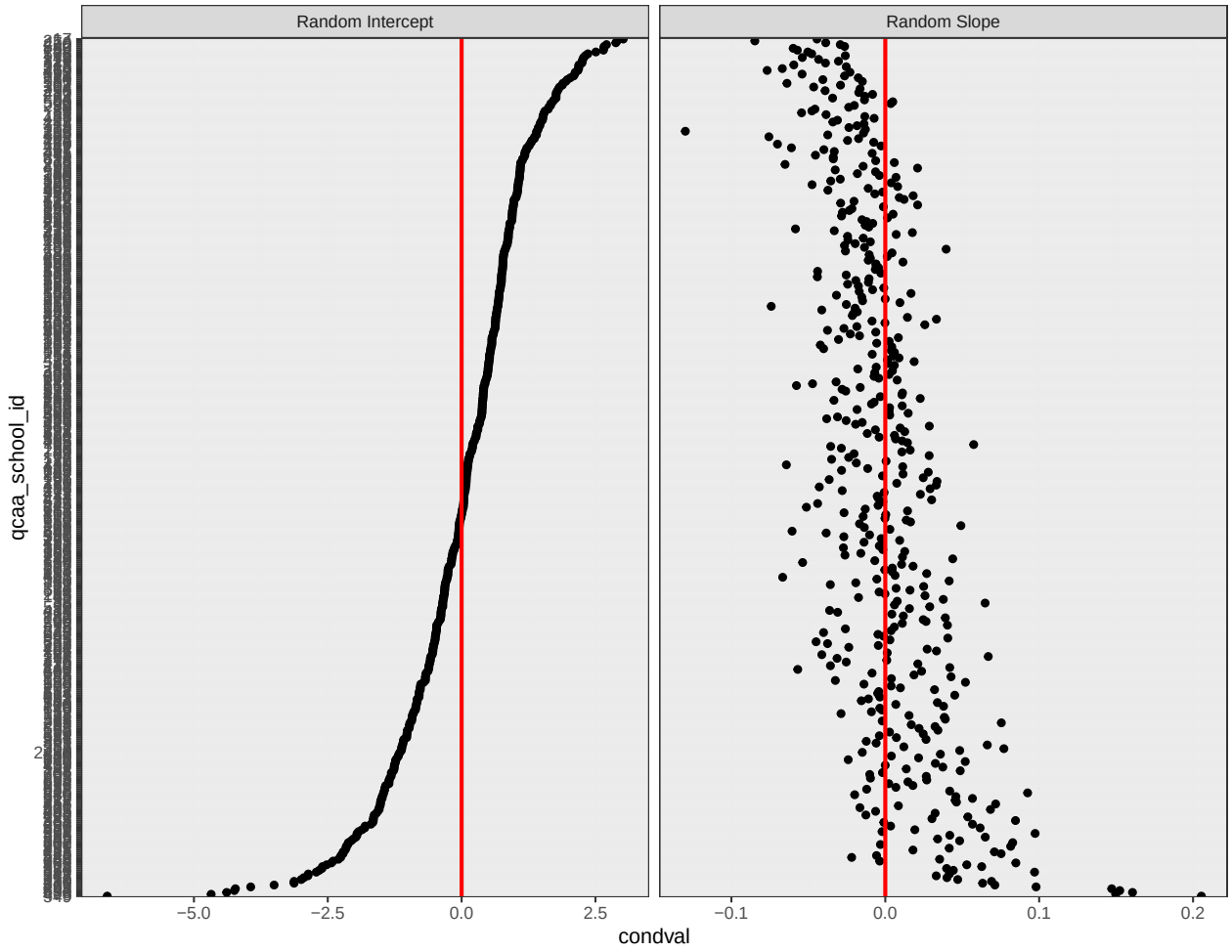

Figure 4: Random effects for all schools

Figure 4 represents the random intercept and slope of the random effects for a given school. It is manifest that the intercept and slope are negatively correlated (-0.73, as shown on the model summary output), 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.

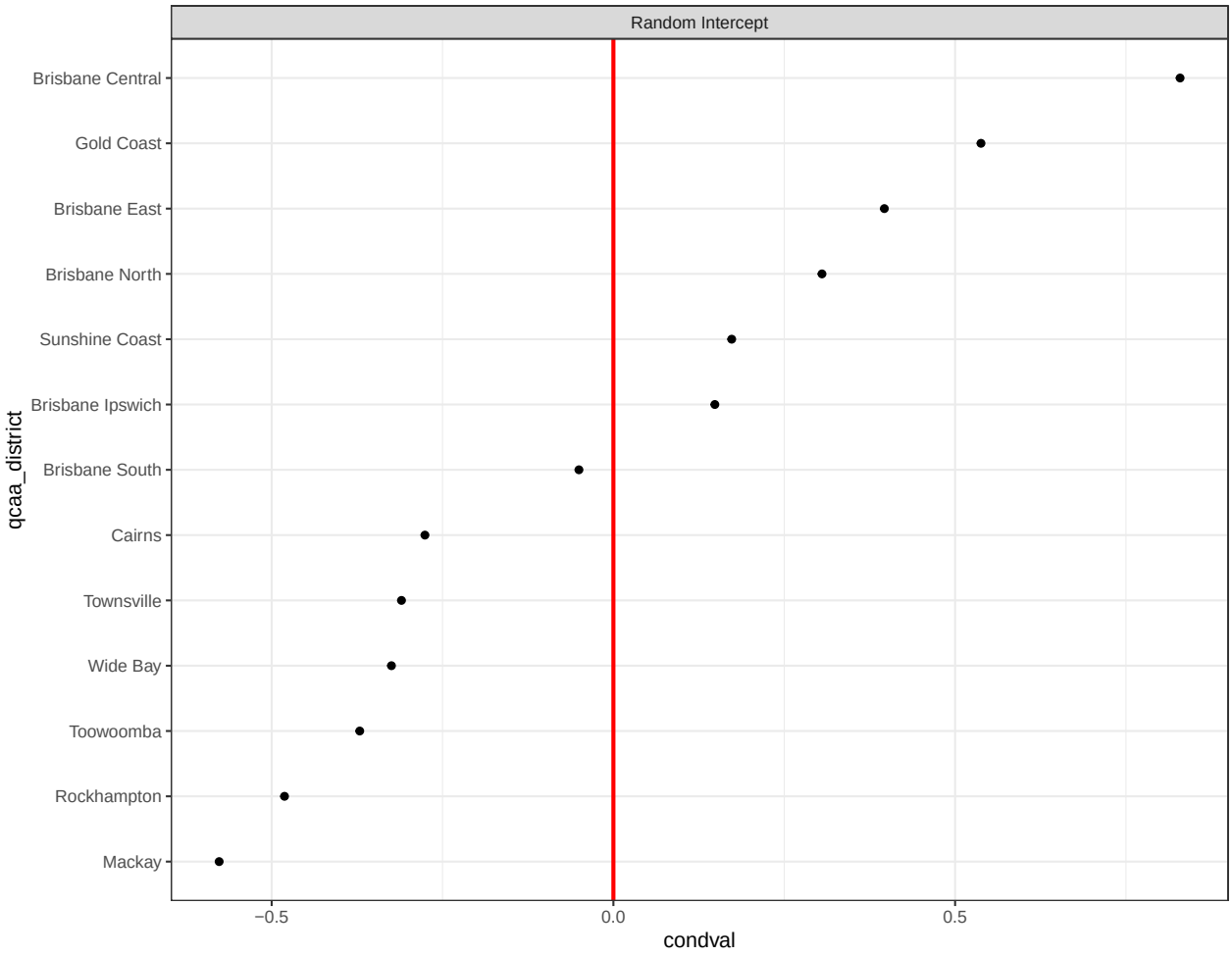

Figure 5: Random intercept for districts

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

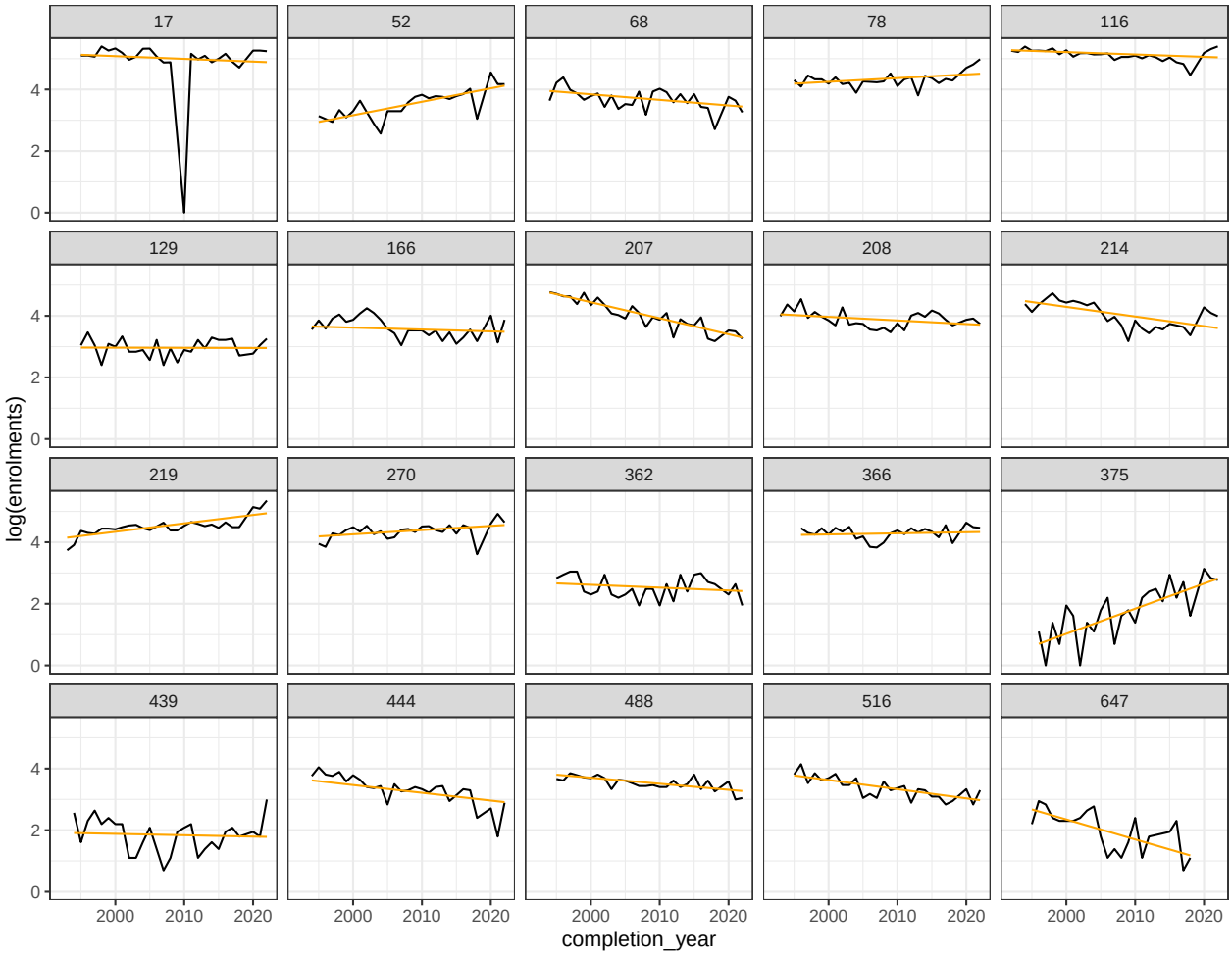

Figure 6: Model predictions for 20 randomly selected schools

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