Multilevel Model for Chemistry

Brendi Ang

17/10/2021

Contents

Chemistry

Exploring the dataset with basic linear model

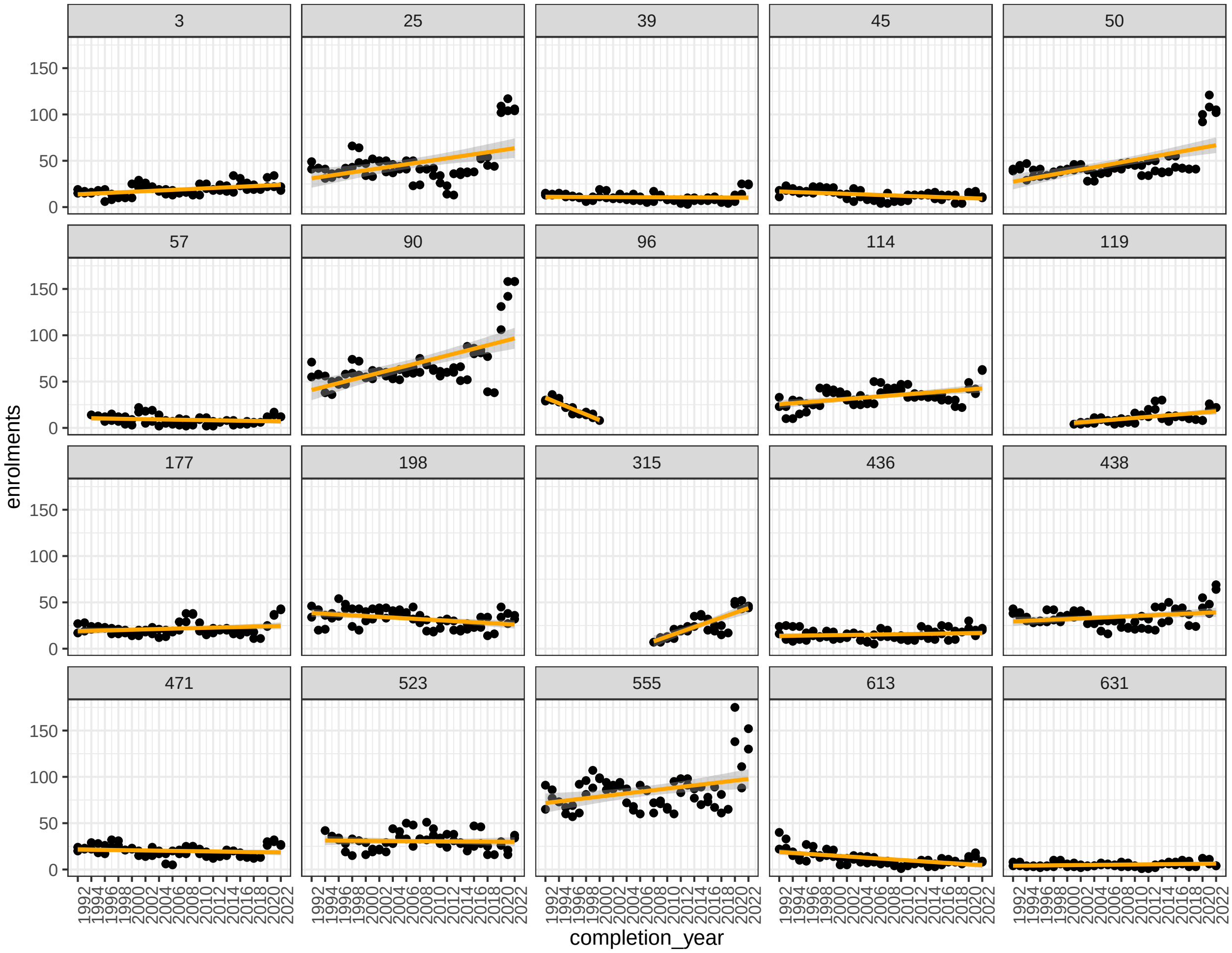

Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Chemistry subject

With reference to the first step, Figure 1 fits a linear model for enrolments for a random sample of 20 schools. The various school sizes is apparent; To illustrate, school 631 and 647 had enrolments of less than 20 each year, while schools such as schools 90 and 555 consistently showed enrolments of approximately more than 50 students each year. Some schools discontinued the subjects (*e.g.* school 96) while other schools only introduced the subject in the later years (*e.g.* school 315).

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

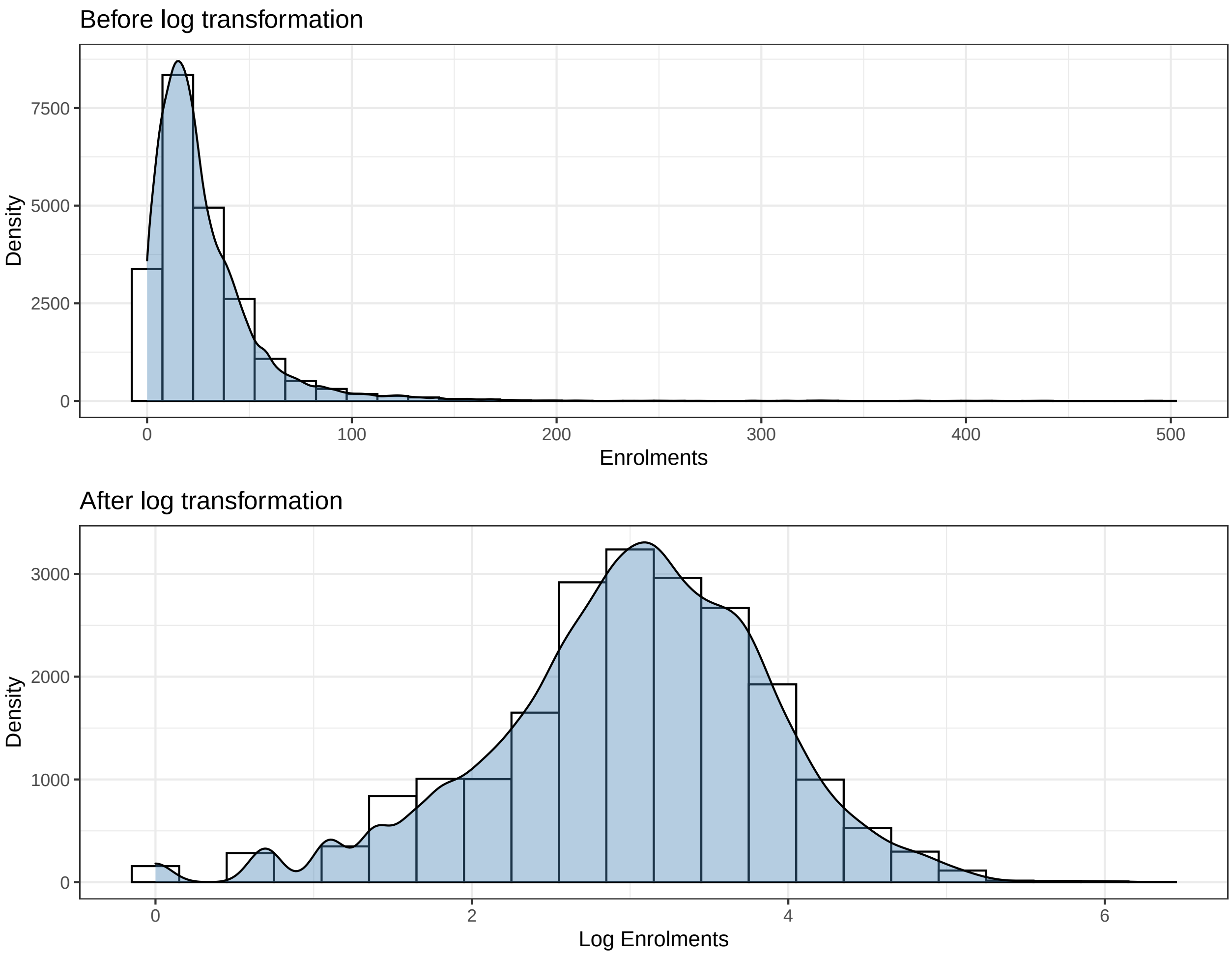

Figure 2: Effects of log transformation for response variable (enrolments) in Chemistry subject

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 Chemistry

	df	AIC
Model0.2: Schools nested within districts	4	28379.04
Model0.1: Schools nested within postcodes	4	28401.11
Model0.0: Within schools	3	28422.33

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) 0.58885  0.76736
##  qcaa_district                (Intercept) 0.10767  0.32813
##  Residual                                 0.20940  0.45760

##
##  Fixed effects:

##             Estimate Std. Error t value
## (Intercept) 2.883525 0.09814118 29.3814

##
##  Number of schools (level-two group) = 454
##  Number of district (level-three group) = 13
```

This model takes into account 454 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 $t^* \neq t$:

$$\text{Level-two ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.5888}{(0.5888 + 0.1077 + 0.2094)} = 0.6450$$

This can be conceptualised as the correlation between enrolments of two random draws from the same school at two different years. In other words, 64.50% 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 district j in time t and time $t^* \neq t$.

$$\text{Level-three ICC} = \frac{\phi_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.1077}{(0.5888 + 0.1077 + 0.2094)} = 0.1189$$

Similarly, this can be conceptualised as the correlation between enrolments of two randomly selected schools from the same district – *i.e.* 11.70% 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) 1.7739e+00 1.3318683
##                                year92      1.8155e-03 0.0426081 -0.790
##  qcaa_district                (Intercept) 6.1621e-02 0.2482367
##                                year92      6.1531e-05 0.0078442 0.512
##  Residual                                  1.4798e-01 0.3846785

##               Estimate  Std. Error   t value
## (Intercept) 2.40033451 0.095359577 25.171405
## year92      0.02165636 0.003068893  7.056736

##  Number of Level Two groups =  454
##  Number of Level Three groups =  13
```

- $\pi_{0ij} = 2.4003$: Initial status for school i in district j (*i.e.* expected log enrolments when time = 0)
- $\pi_{1ij} = 0.0217$: Growth rate for school i in district j
- $\epsilon_{tij} = 0.1480$: Variance in within-school residuals after accounting for linear growth overtime

When the subject was first introduced in 1992, schools were expected to have 11.0265 ($e^{1.7848}$) enrolments, on average. Furthermore, on average, enrolments were expected to increase by 2.1893% ($(e^{0.0216564} - 1) \times 100$) per year. The estimated within-schools variance decreased by 29.3326% (0.2094 to 0.14797754), implying that 29.3326% of 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	AIC
model4.7	22651.34
model4.1	22651.36
model4.4	22653.66
model4.5	22654.01
model4.0	22656.29
model4.2	22718.07
model4.8	22718.07
model4.9	22718.07
model4.10	22718.07
model4.6	22718.07
model4.3	22720.78

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 `model4.4` (Table 2). The next section will test the selected model's random effects to build 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)
14	22651.66	22762.71	-11311.83	22623.66	NA	NA	NA
16	22651.34	22778.26	-11309.67	22619.34	4.311858	2	0.04

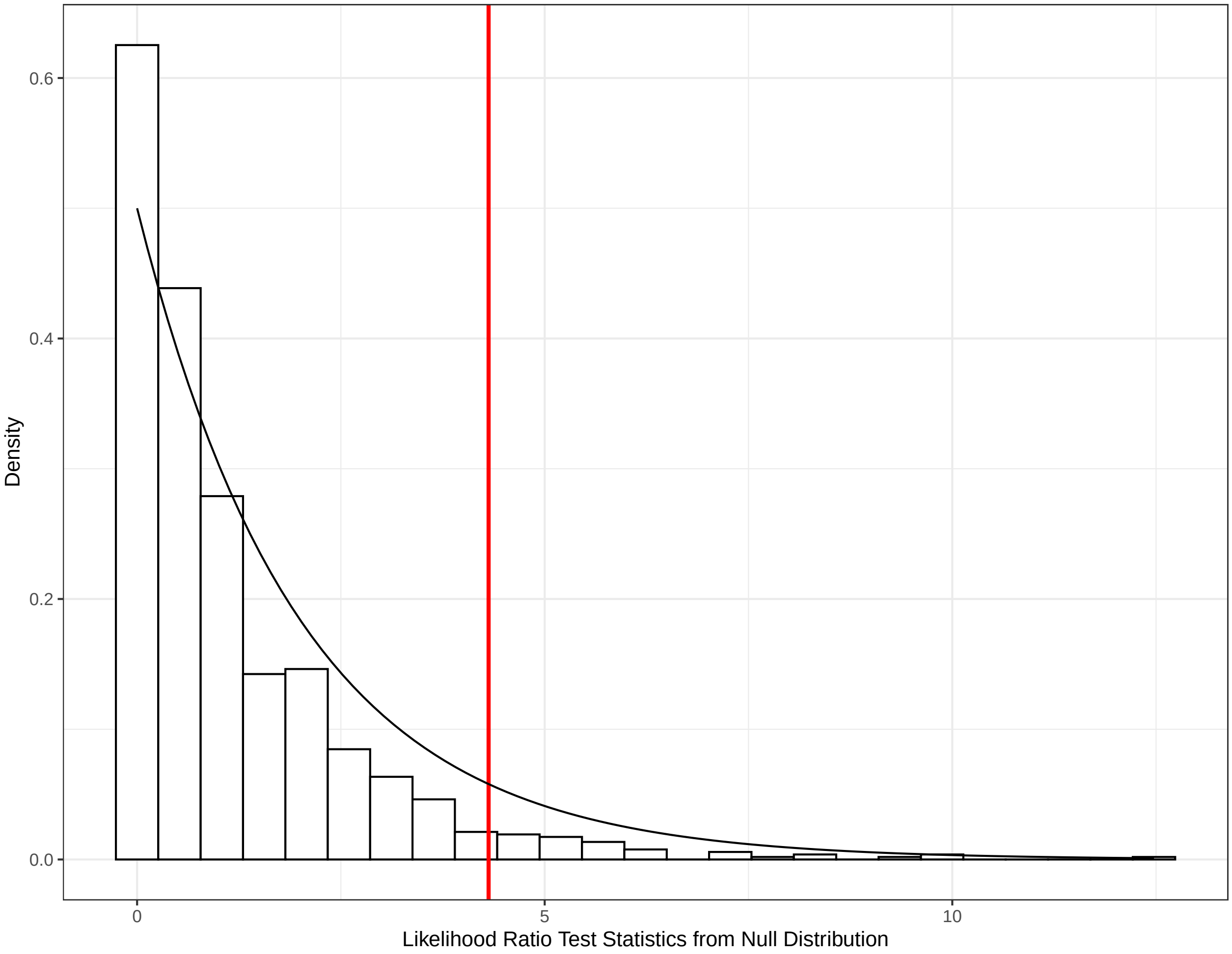

Figure 3: Histogram of likelihood ratio test statistic, with a red vertical line indicating the likelihood ratio test statistic for the actual model

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 0.04% (Table 3) indicates the proportion of times in which the bootstrap test statistic is greater than the observed test statistic. The estimated p-value is $0.04 < 0.05$ fails to reject the null hypothesis at the 5% level, indicating that the larger 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)|qcaa_district:qcaa_school_id	1.1802941	1.3535015
cor_year92.(Intercept)|qcaa_district:qcaa_school_id	-0.8005691	-0.7167320
sd_year92|qcaa_district:qcaa_school_id	0.0362919	0.0424250
sd_(Intercept)|qcaa_district	0.0854839	0.4588320
cor_year92.(Intercept)|qcaa_district	-0.6818153	1.0000000
sd_year92|qcaa_district	0.0006963	0.0113808
sigma	0.3792006	0.3871578
(Intercept)	2.3667074	3.0418792
year92	0.0071077	0.0269531
sectorGovernment	-0.2661168	0.3889485
sectorIndependent	-1.4463471	-0.6984020
unityear_12_enrolments	-0.0646203	-0.0173266
year92:sectorGovernment	-0.0209668	0.0013685
year92:sectorIndependent	0.0189922	0.0427861
sectorGovernment:unityear_12_enrolments	-0.0624251	-0.0069597
sectorIndependent:unityear_12_enrolments	-0.0527715	0.0100726

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.

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)

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

Therefore, the composite model can be written as

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

Fixed effects

```
summary(model_f)
```

```
##  Groups                       Name        Variance   Std.Dev.  Corr
##  qcaa_district:qcaa_school_id (Intercept) 1.6097e+00 1.2687521
##                               year92      1.5320e-03 0.0391402 -0.765
##  qcaa_district                (Intercept) 7.7313e-02 0.2780516
##                               year92      3.3268e-05 0.0057678 0.621
##  Residual                                 1.4672e-01 0.3830370
```

```
##                                             Estimate  Std. Error    t value
## (Intercept)                               2.69358309 0.165949162 16.2313751
## year92                                    0.01832166 0.004966819  3.6888110
## sectorGovernment                          0.06422409 0.170475913  0.3767341
## sectorIndependent                        -1.04537987 0.188584541 -5.5432957
## unityear_12_enrolments                   -0.04196115 0.012373547 -3.3911980
## year92:sectorGovernment                  -0.01003111 0.005433864 -1.8460368
## year92:sectorIndependent                  0.03050230 0.006036915  5.0526299
## sectorGovernment:unityear_12_enrolments  -0.03559772 0.014308159 -2.4879313
## sectorIndependent:unityear_12_enrolments -0.02235697 0.016211174 -1.3791088
##  Number of Level Two groups =  454
##  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 0.832496% $((e^{0.0183-0.0100} - 1) \times 100)$, which is 1.0082% $((e^{0.0100} - 1) \times 100)$ less than that of catholic schools. On the other hand, the mean enrolments for independent schools are estimated to increase by 5.0036% $(e^{0.01832+0.03050} - 1) \times 100)$ each year, which is 3.0972% $((e^{0.0305022} - 1) \times 100)$ more than catholic schools.

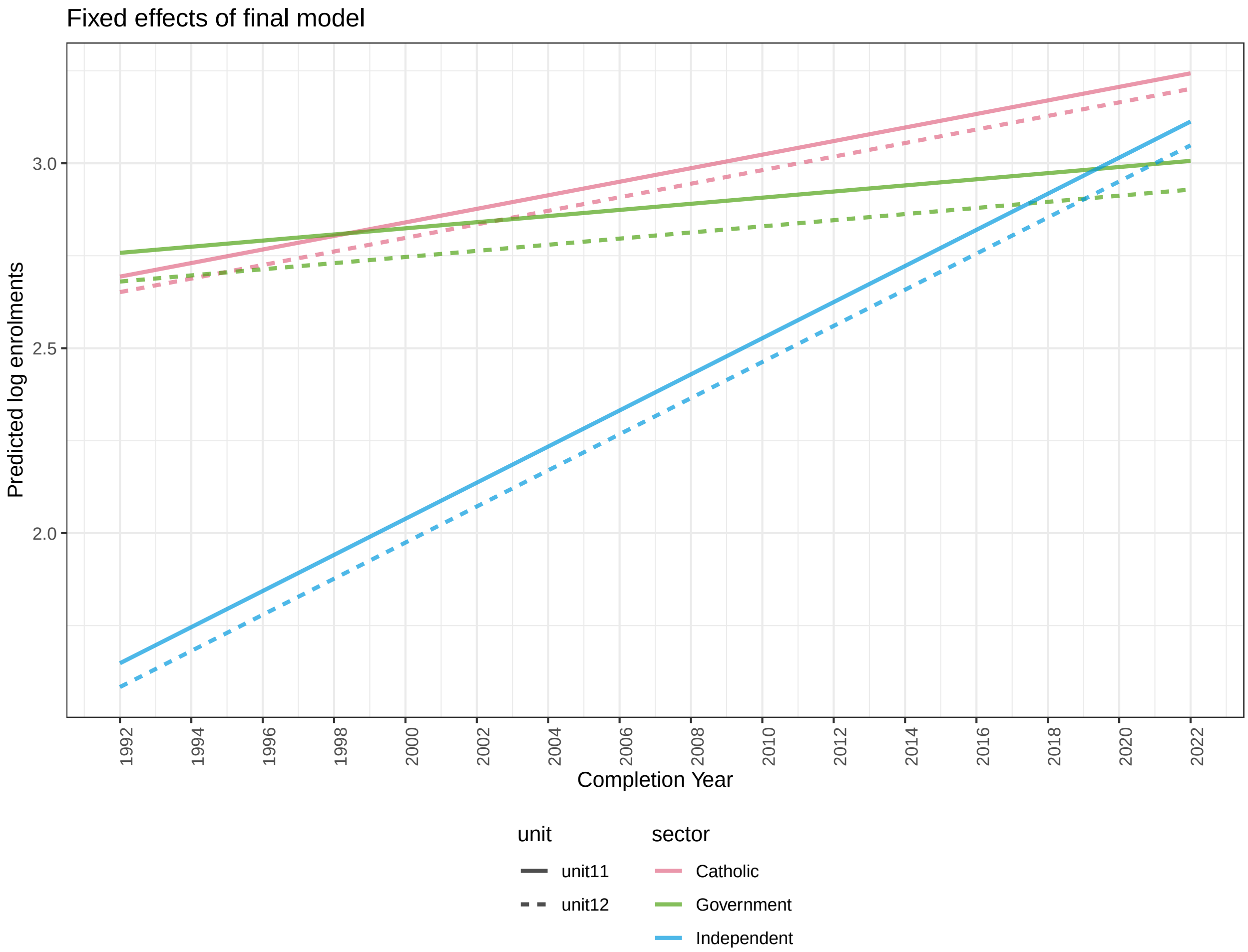

Figure 4: Fixed effects of the final model for Chemistry

The fixed effects are better accentuated in Figure 4. As aforementioned, independent school have the lowest initial status, that is, the least enrolments when the subject was first introduced. However, this low enrolments was matched with a relatively high slope compared to the other sectors. From 2021 onwards, independent schools are expected to have greater enrolments than government schools, on average. Government schools have the highest initial status, but are expected to have the smallest increase in enrolments over the years (as demonstrated by the gentle slope).

Random effects

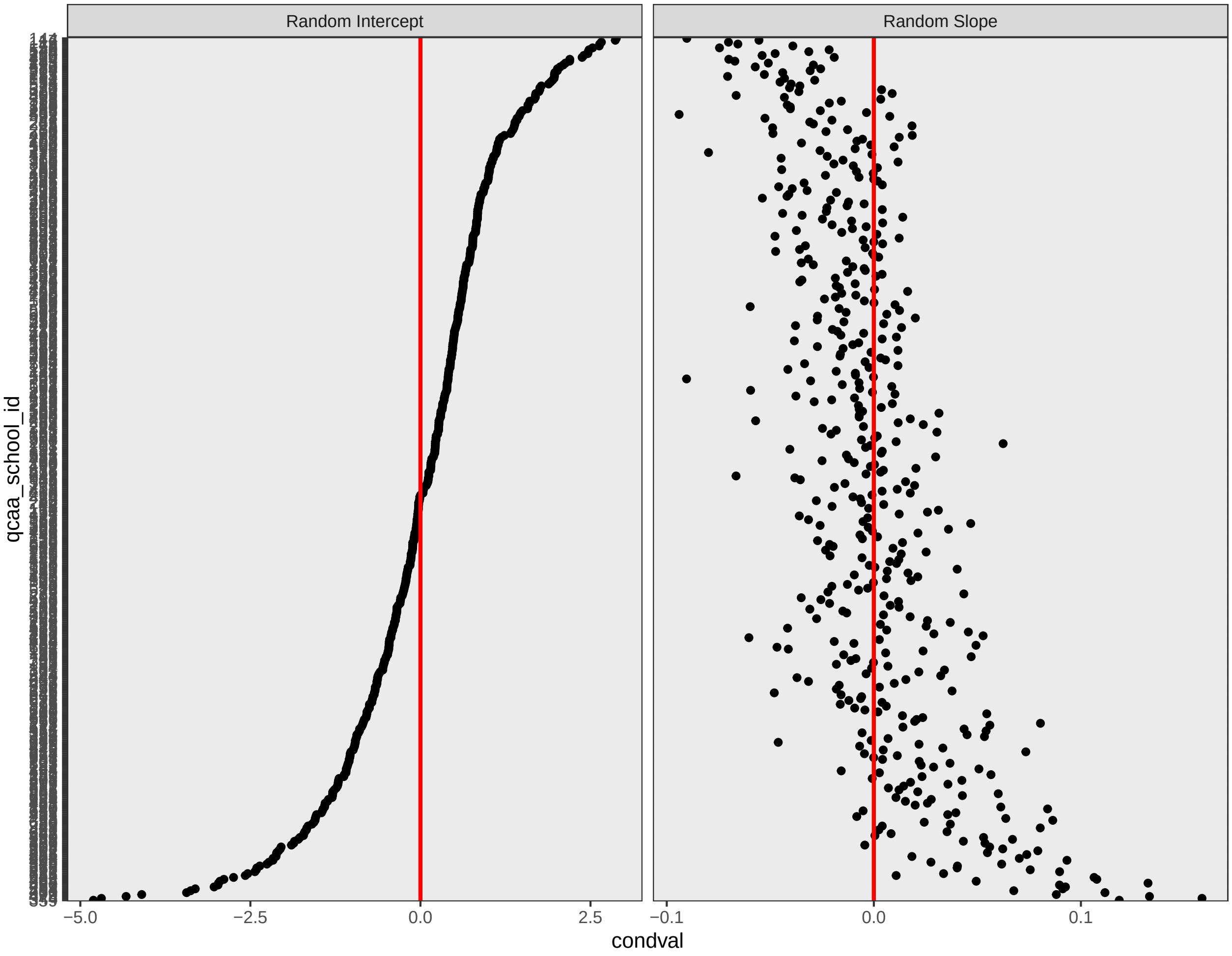

Figure 5: Random effects for schools

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 (csorrelation = -0.76, as shown in the model output). 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.

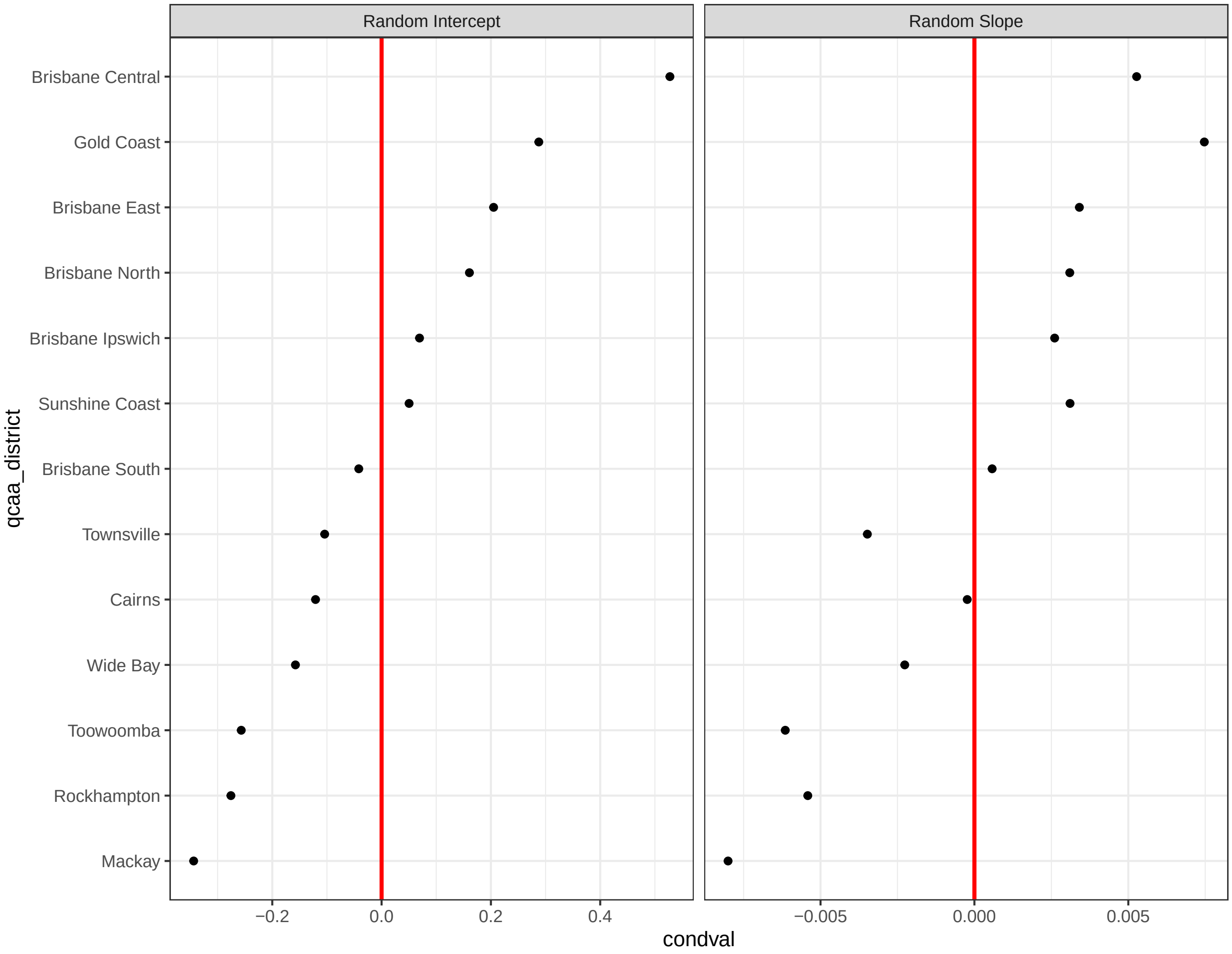

Figure 6: Random intercept for districts

The model includes a random slope at level three, and the effects of the random intercept and slope in level three is shown in Figure 6. As shown in the model output, correlation between the random intercept and slope is 0.62. Loosely speaking, this suggests that districts with large enrolments are going to be larger, while smaller districts are going to increase (decrease) at a slower rate. Based on Figure 6, Gold Coast are estimated to have the highest slope, indicating that the rate of change in enrolment is the greatest compared to the other districts.

Predictions

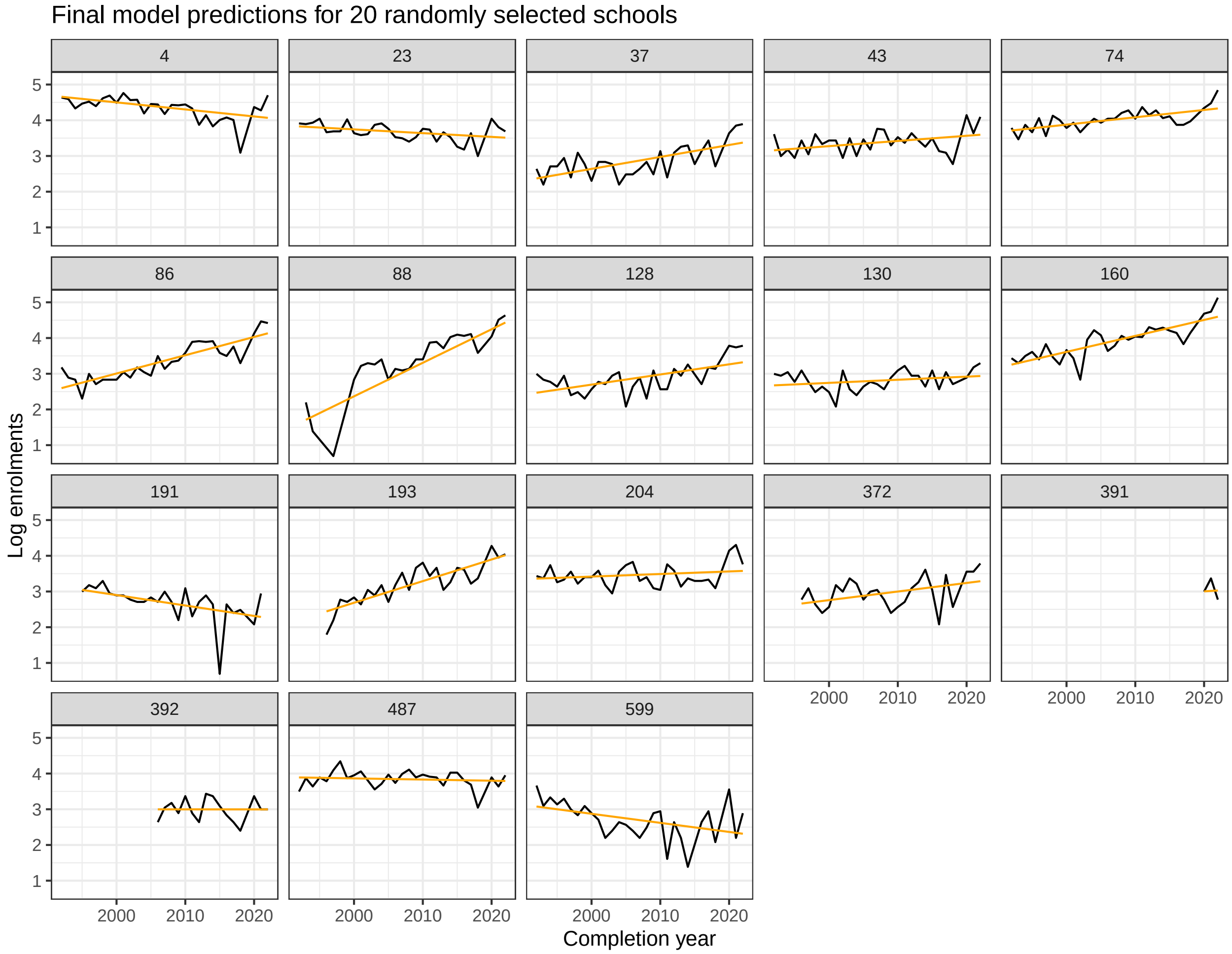

Figure 7: Model predictions for 20 randomly selected schools

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