Multilevel Model for Earth and Environmental Science

Brendi Ang

17/10/2021

Contents

Earth and Environmental Science

Exploring the dataset with basic linear model

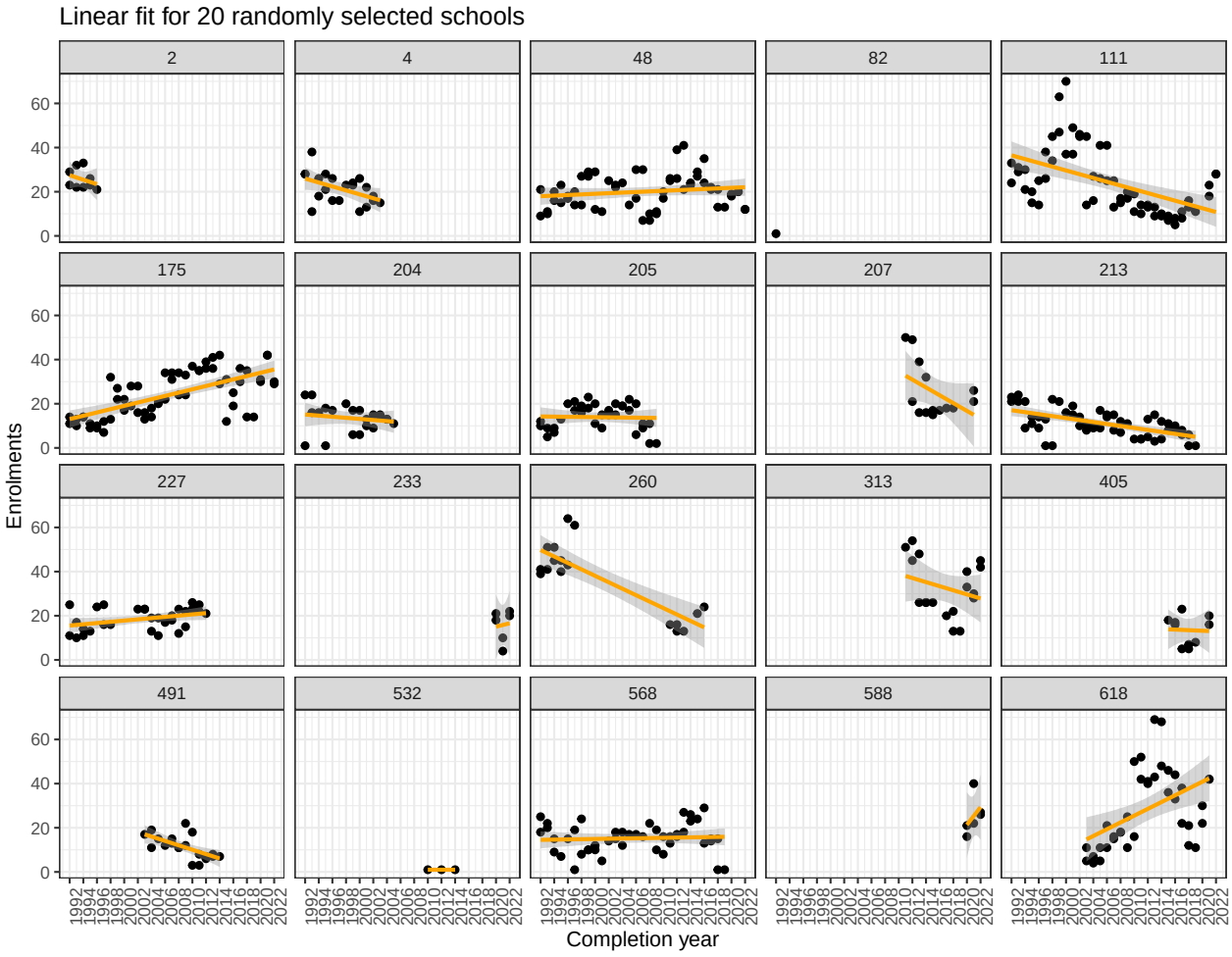

Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Earth and Environmental Science subject

Figure 1 fits a linear model for 20 randomly selected schools. The difference in enrolment numbers across the cohorts between schools is apparent, for example, school 405 and 532 which consistently showed little enrolments (< 20) per year while schools such as school 175 and 618 demonstrates significantly higher enrolment numbers in a single cohort.

Some schools offered the subject in little years, such as school 233 and 588, which only offered the subject in the new QCE system. There were also varying patterns in enrolment trends where school 618 showed a large increase in enrolments over the years while school 111 appears to have a stark decrease in enrolments since it offered the subject. In some cases, there were only a few students enrolments in the school, such as school 532, which only showed 1 enrolment when the school offered the subject.

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

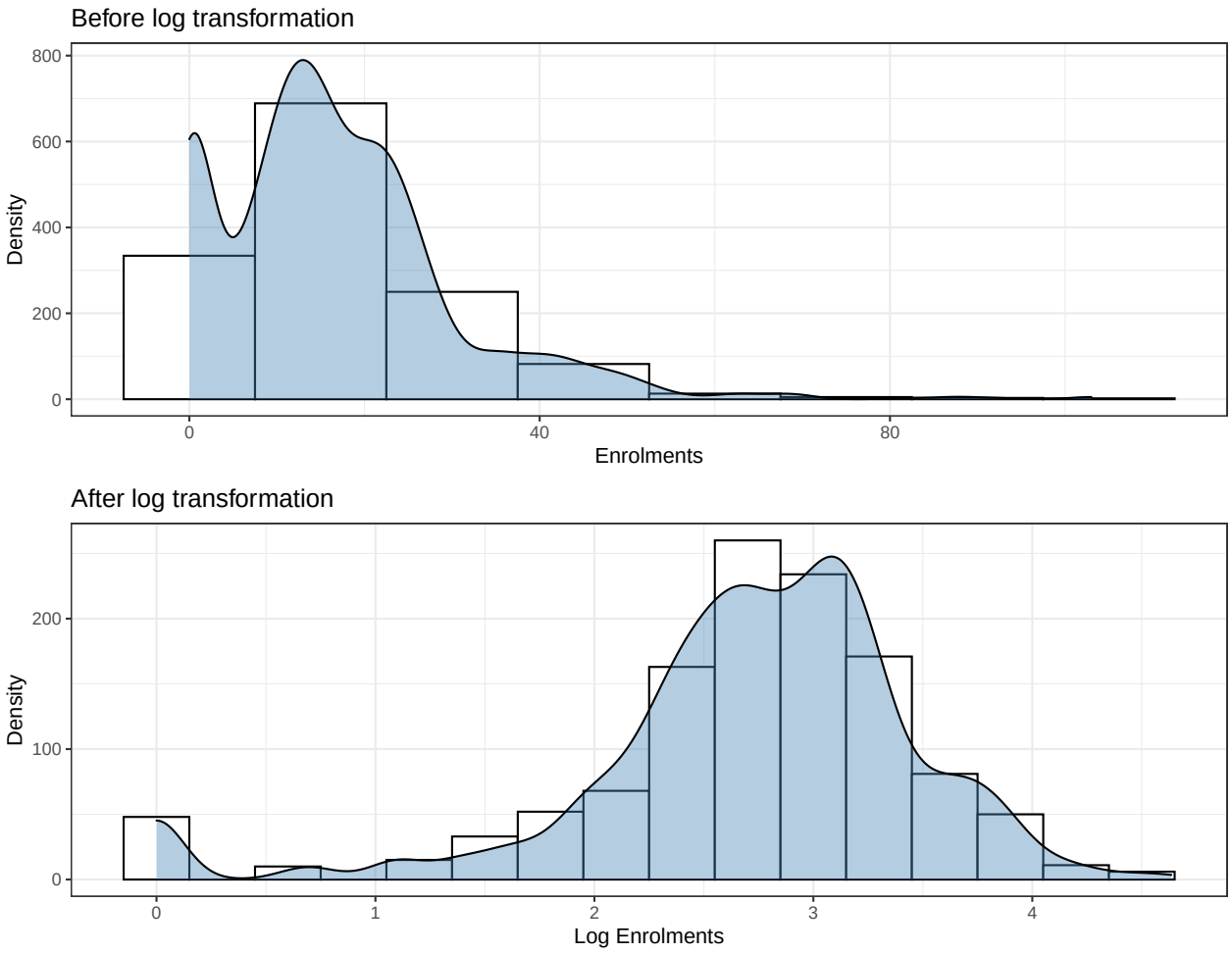

Figure 2: Effects of log transformation for response variable (enrolments) in Earth and Environmental Science subject

The enrolments were right skewed, which is likely to be attributed to the various school sizes (as seen in Figure 1). 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 Earth and Environmental Science

	df	AIC
Model0.0: Within schools	3	2243.65
Model0.1: Schools nested within postcodes	4	2244.80
Model0.2: Schools nested within districts	4	2245.65

As outlined 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 superior 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) 1.51734  1.2318
##  Residual                   0.29507  0.5432

##
##  Fixed effects:

##             Estimate Std. Error  t value
## (Intercept)  2.06781  0.1325348 15.60202

##
##  Number of schools (level-two group) = 92
##  Number of district (level-three group) = NA
```

This model takes into account 92. 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^*:

$$\text{Level-two ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{1.5173}{(1.5173 + 0.2951)} = 0.8372$$

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, 83.72% 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) 2.0601140 1.435310
##                 year92      0.0028797 0.053663 -0.847
##  Residual                   0.2458310 0.495813

##               Estimate  Std. Error  t value
## (Intercept) 1.63148417 0.176855363 9.224963
## year92      0.03527984 0.007508286 4.698787

##  Number of Level Two groups =  92
##  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.

- $\pi_{0ij} = 1.6314$: 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 1992, schools were expected to have an average of 5.1115 ($e^{1.6314842}$) enrolments, which is a relatively low number as compared to the other mathematics and science subjects. Furthermore, Figure **??** demonstrated that the number of schools offering the subject in 1992 was the highest among all years.

On average, the enrolments were expected to increase by 3.59304% ($(e^{0.0353} - 1) \times 100$) per year. The estimated within-school variance decreased by 16.70% (0.2951 to 0.2458), indicating the 16.70% 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
model4.4	11	2083.856	2139.727	-1030.928
model4.5	12	2085.205	2146.155	-1030.603
model4.7	13	2086.236	2152.265	-1030.118
model4.3	10	2086.805	2137.596	-1033.402
model4.1	14	2087.644	2158.753	-1029.822
model4.10	11	2087.848	2143.719	-1032.924
model4.9	11	2087.848	2143.719	-1032.924
model4.2	11	2087.848	2143.719	-1032.924
model4.8	11	2087.848	2143.719	-1032.924
model4.6	12	2089.241	2150.191	-1032.621
model4.0	16	2091.481	2172.748	-1029.740

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` appears to have the optimal (smallest) AIC (Table 2), 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	2223.393	2269.105	-1102.696	2205.393	NA	NA	NA
11	2083.856	2139.727	-1030.928	2061.856	143.537	2	0

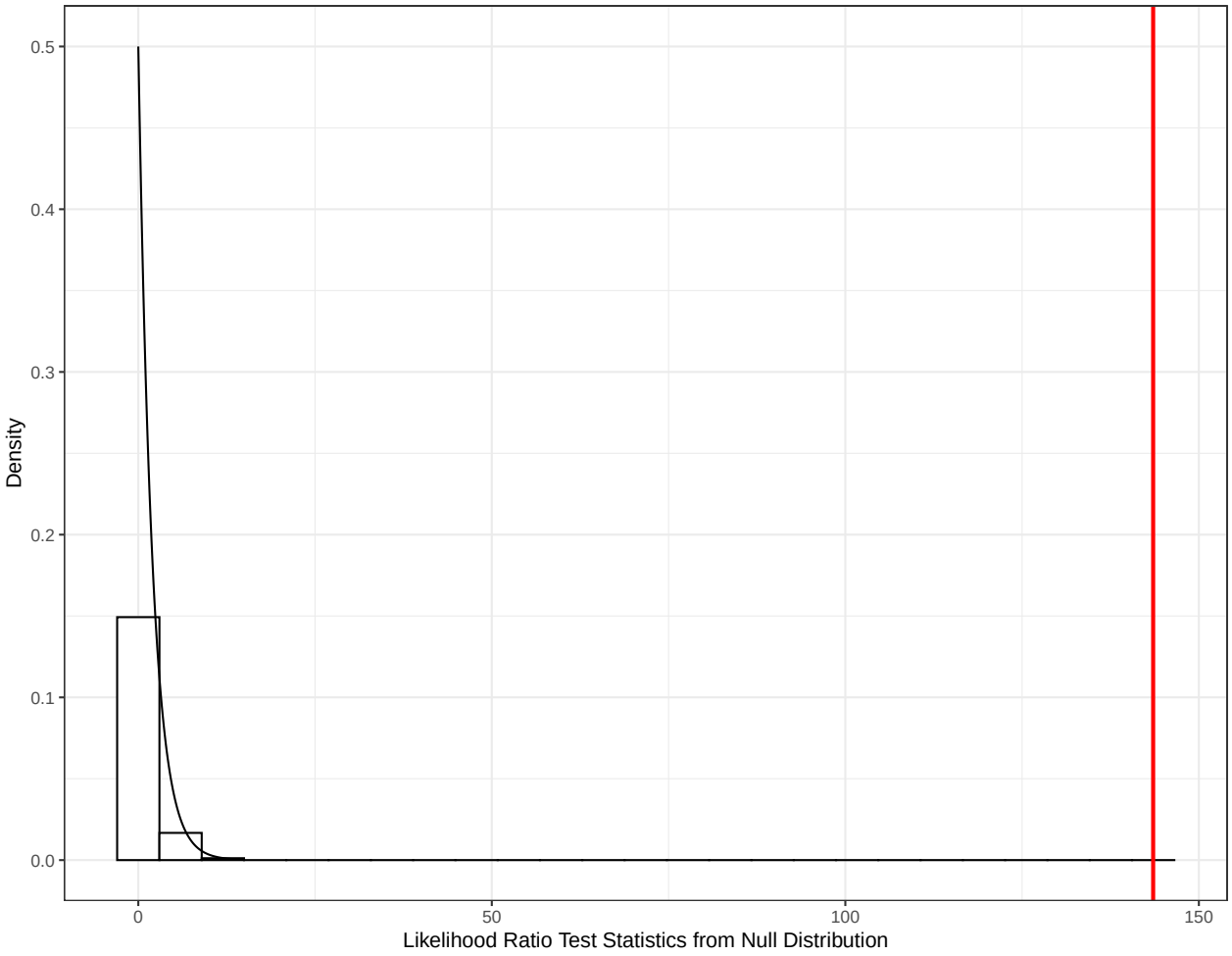

Figure 3: Histogram of likelihood ratio test statistic, with a red vertical line indicating the likelihood ratio test statistic for the actual model

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. Figure 3 displays the likelihood ratio test statistic from the null distribution, with the red line indicates the likelihood ratio test statistic using the actual data.

There is overwhelming statistical evidence ($\chi^2 = 143.537$ and p-value $= 0$ from 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	1.0791392	1.5883744
cor_year92.(Intercept)|qcaa_school_id	-0.9079310	-0.6751860
sd_year92|qcaa_school_id	0.0393192	0.0678080
sigma	0.4746765	0.5167394
(Intercept)	0.5979313	2.9539799
year92	-0.0293459	0.0774094
sectorGovernment	-1.6512267	0.7654638
sectorIndependent	-0.4903149	2.2137154
unityear_12_enrolments	-0.0447976	0.0733969
year92:sectorGovernment	-0.0366734	0.0751267
year92:sectorIndependent	-0.0819655	0.0454768

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) will contain new predictor(`sector`)

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

The composite model can therefore be written as:

$$\begin{aligned} Y_{tij} &= \pi_{0ij} + \pi_{1ij} year92_{tij} + \epsilon_{tij} \\ &= (\beta_{00j} + \beta_{01j} sector_{ij} + \beta_{02j} unit_{ij} + u_{0ij}) + (\beta_{10j} + \beta_{11j} sector_{ij} + u_{1ij}) year92_{tij} + \epsilon_{tij} \\ &= [\beta_{00j} + \beta_{01j} sector_{ij} + \beta_{02} unit_{ij} + \beta_{10j} year92_{tij} + \beta_{11j} sector_{ij} year92_{tij}] + [u_{0ij} + u_{1ij} + \epsilon_{tij}] \end{aligned}$$

Fixed effects

```
summary(model_f)
```

```
##  Groups         Name        Variance  Std.Dev. Corr
##  qcaa_school_id (Intercept) 1.8069480 1.344228
##                 year92      0.0028392 0.053284 -0.824
##  Residual                   0.2457532 0.495735
```

```
##                              Estimate Std. Error    t value
## (Intercept)                1.70508906 0.63997319  2.6643133
## year92                     0.02482993 0.02880730  0.8619320
## sectorGovernment          -0.35142132 0.66866804 -0.5255542
## sectorIndependent          0.96758390 0.74465152  1.2993781
## unityear_12_enrolments     0.01447581 0.02919046  0.4959088
## year92:sectorGovernment    0.01972213 0.03014928  0.6541493
## year92:sectorIndependent -0.02215715 0.03284658 -0.6745648
```

```
##  Number of Level Two groups =  92
##  Number of Level Three groups =  NA
```

Based on the model output, the estimated mean enrolments for government schools are estimated to be 29.63% ($(e^{0.3514213} - 1) \times 100$) less than that of catholic schools when the subject was first introduced in 1992. However, government schools are estimated to have a mean increase of 4.5560% ($(e^{0.0248299+0.0197221} - 1) * \times 100$) per year, which is 1.9918% ($(e^{0.0197221} - 1) * \times 100$) greater than the increase in enrolments in catholic schools.

Independent schools are estimated to have a 163.158% ($(e^{0.9675839} - 1) \times 100$) more than that of catholic schools in 1992. However, independent schools showed a slow increase in enrolments (0.26786%) over per year, on average. This increase in enrolments is -2.1913% ($(exp^{-0.0221572} - 1) * 100$) less than that of catholic schools.

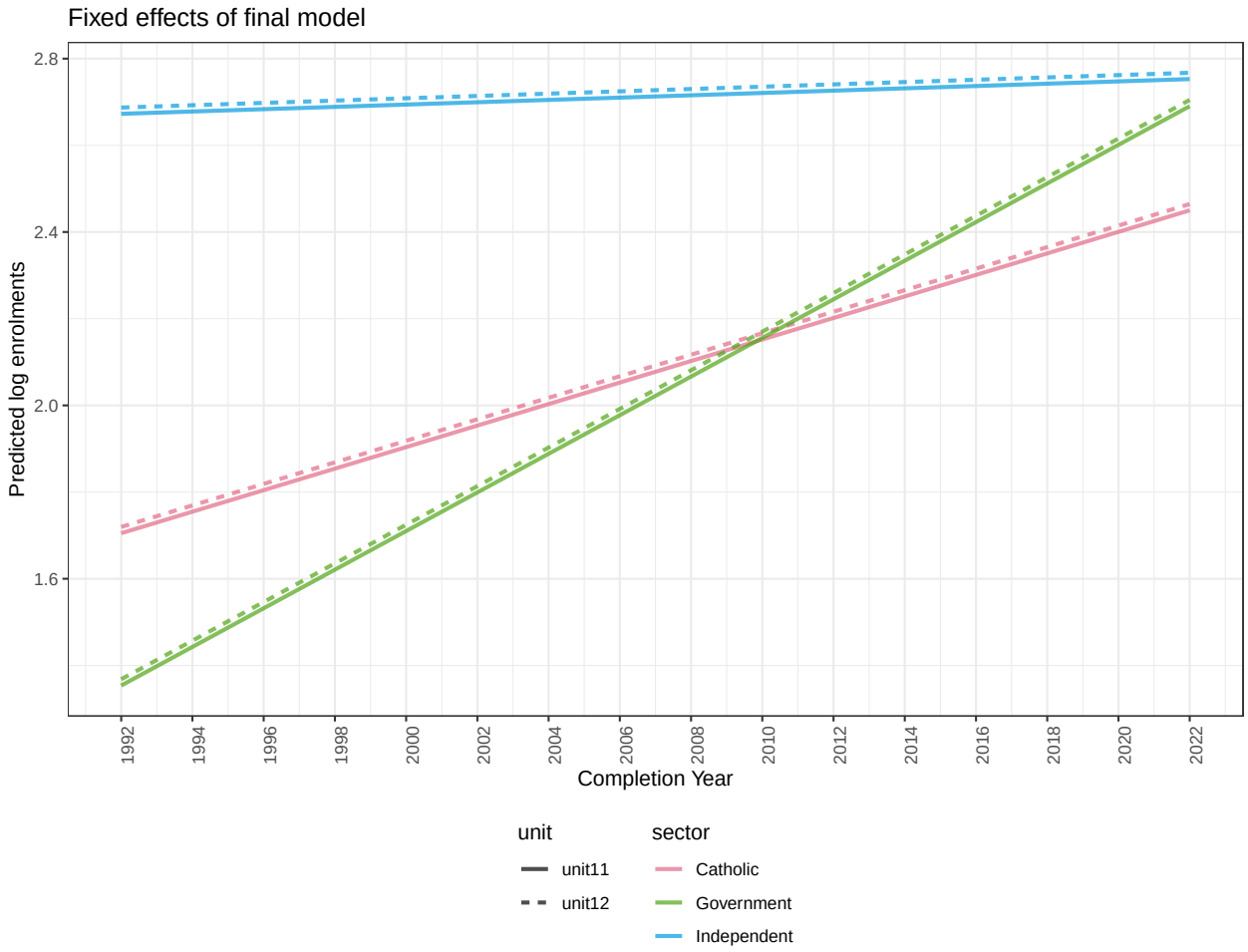

Figure 4: Fixed effects of the final model for Agricultural Practices subject

The results can be better visualised in Figure 4. On average Government schools started off with little enrolments, but showed a stark increase in enrolments over the years. In contrast, Independent schools have relatively high enrolments when the subject was first introduced in 1992, but showed little increase over the years.

Random effects

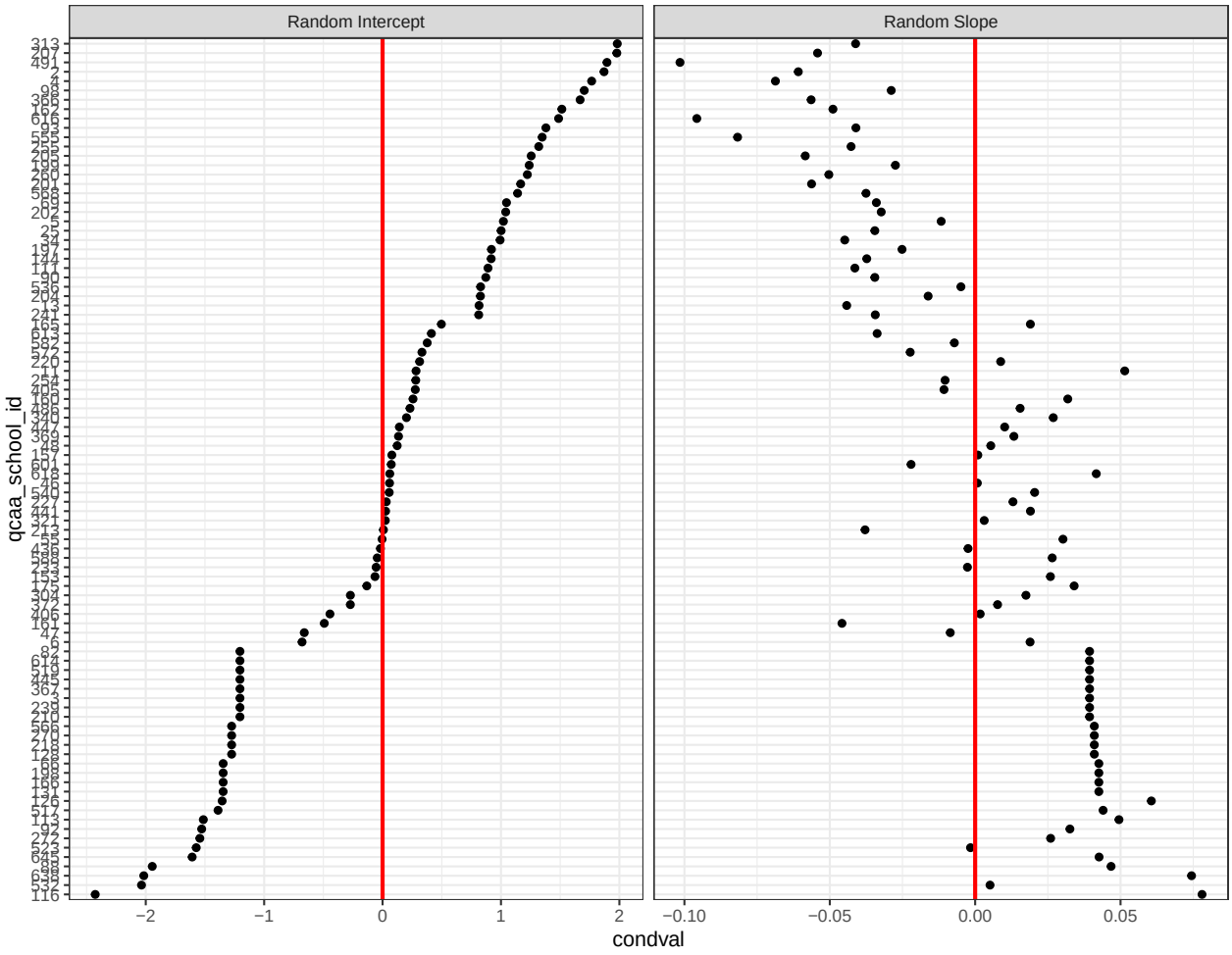

Figure 5: Random effects for all schools

Figure 5 shows the random effects for all 92 schools that offered the subject. There is a clear negative correlation between the random intercept and the random slope, which indicates that in general, schools with lesser enrolments are generally matched with a larger increase in enrolments over the years.

Predictions

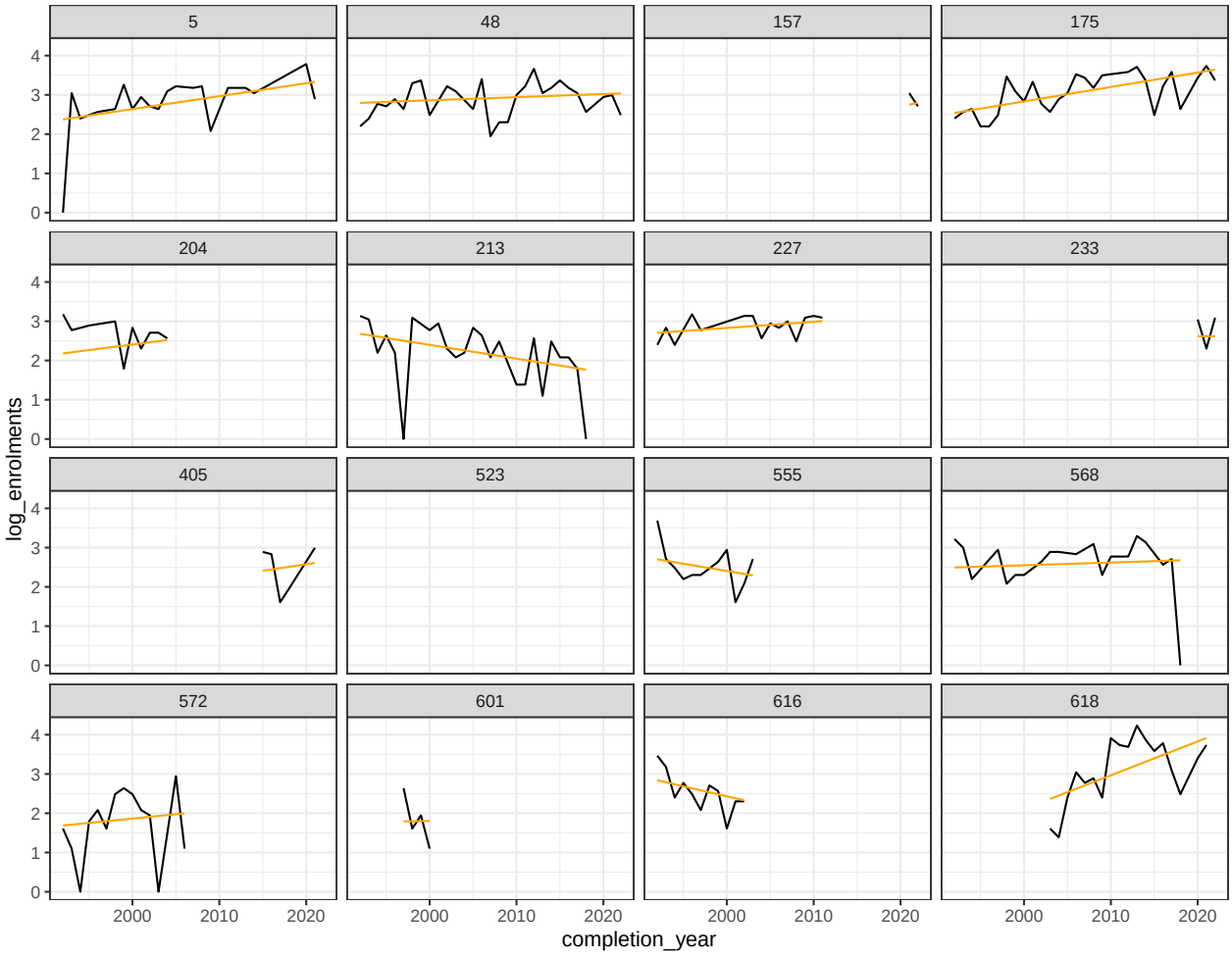

Figure 6: Model predictions for year 11 enrolments for 20 randomly selected schools

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