Multilevel Model for Agricultural Sciences

Brendi Ang

17/10/2021

Contents

Agricultural Science

Exploring the dataset with basic linear model

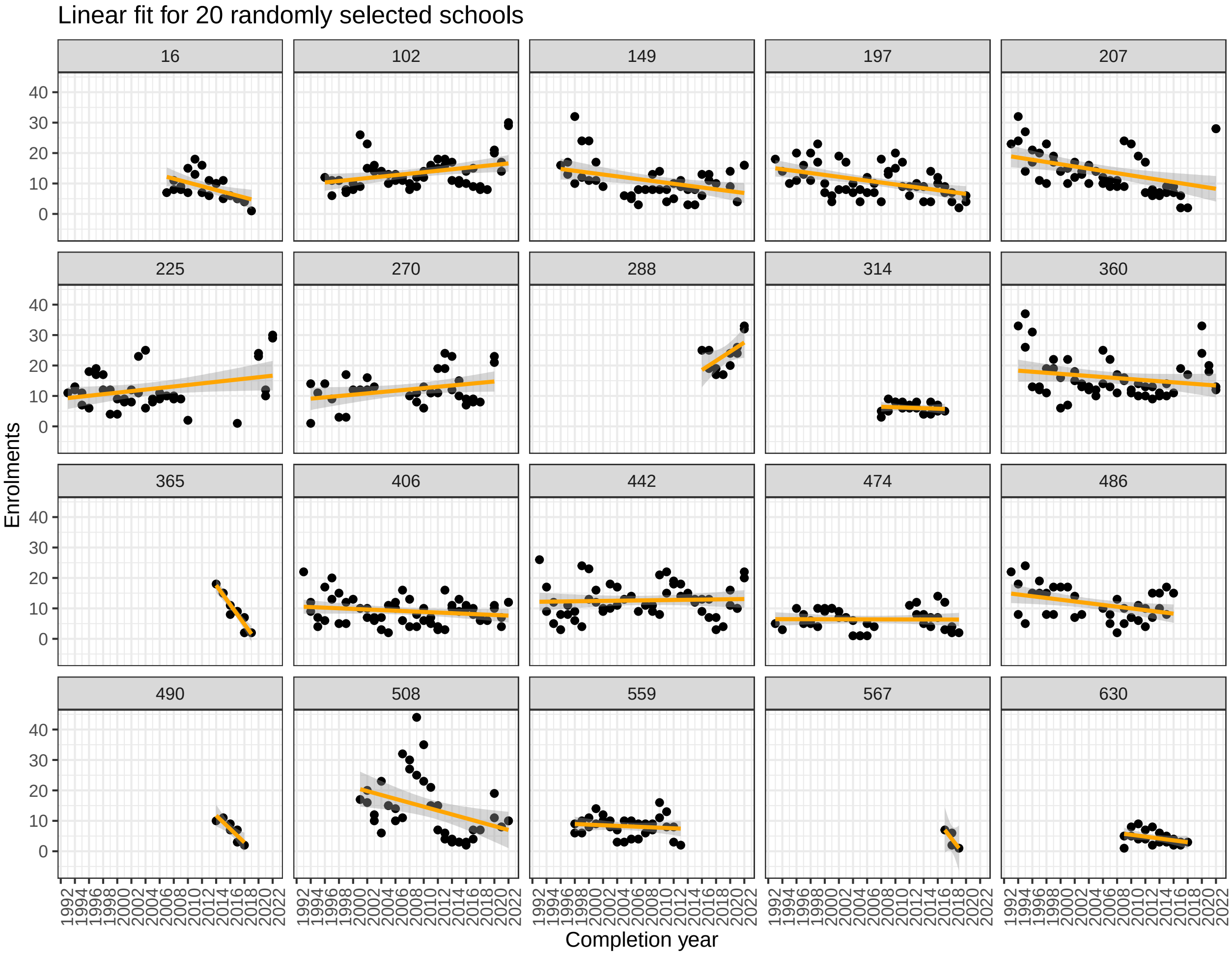


Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Agricultural Science subject

Figure 1 fits a linear model for 20 randomly selected schools. In most of these selected schools, the linear model captured a downward trend in enrolments. Some schools showed a increase, and in particular, school 288 showed a significant increase in enrolments as compared to the other selected schools. It seems that most schools that ceased offering the subject in the later years (*e.g.* schools 365, 490, 567) showed a large downward trend before ceasing the subject. Interestingly, school 508 displayed a rather cubic trend, where enrolments increases exponentially from 2002 before decrease till 2015 and picking up again.

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

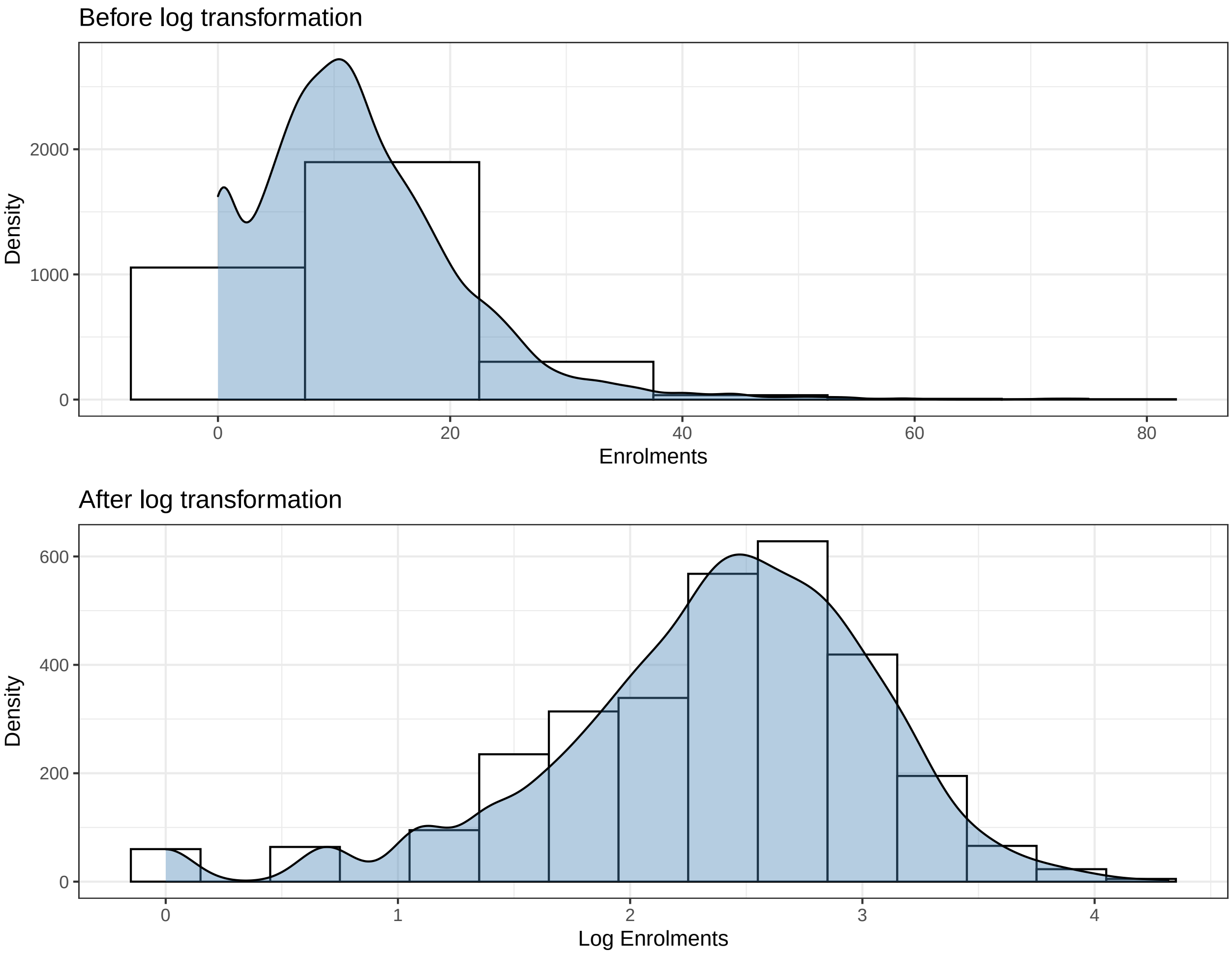


Figure 2: Effects of log transformation for response variable (enrolments) in Agricultural Science subject

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 Agricultural Science

	df	AIC
Model0.1: Schools nested within postcodes	4	5146.230
Model0.0: Within schools	3	5149.113
Model0.2: Schools nested within districts	4	5149.856

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 (`model0.1`) corresponding to schools nested within postcodes is the superior model and will be used in the subsequent analysis.

Intraclass correlation (ICC)

```
summary(model0.1)
```

```
## Random effects:

##  Groups                          Name        Variance Std.Dev.
##  qcaa_school_id:school_postcode (Intercept) 0.40515  0.63651
##  school_postcode                (Intercept) 0.12275  0.35036
##  Residual                                   0.29180  0.54018

##
##  Fixed effects:

##             Estimate Std. Error  t value
## (Intercept) 2.018581 0.07739906 26.08017

##
##  Number of schools (level-two group) = 112
##  Number of school postcodes (level-three group) = 81
```

This model will takes into account 112 schools nested in 81 postcodes. 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 a certain postcode k in time t and in time $t^* \neq t$:

$$\text{Level-two ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.4052}{(0.4052 + 0.1228 + 0.2918)} = 0.4943$$

This can be conceptualised as the correlation between enrolments of two random draws from the same school at two different years. In other words, 49.45% of the total variability is attributable to the changes over time within schools.

The **level-three ICC** refers to the correlation between different schools i and i^* from a specific postcode j.

$$\text{Level-three ICC} = \frac{\phi^2_{00}}{\tau^2_{00} + \phi^2_{00} + \sigma^2} = \frac{0.1228}{(0.4052 + 0.1228 + 0.2918)} = 0.1498$$

Likewise, this can be loosely translated to be the correlation between enrolments of two random draws from two schools from two different postcode. In this case, 14.98% of the total variability is due to the difference between postcodes.

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 postcodes, and schools from different postcodes.

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

```
##  Groups                         Name        Variance  Std.Dev. Corr
##  school_postcode:qcaa_school_id (Intercept) 1.0049220 1.002458
##                                 year94      0.0013985 0.037397 -0.906
##  school_postcode                (Intercept) 0.1565287 0.395637
##                                 year94      0.0002277 0.015090 -0.669
##  Residual                                   0.2539595 0.503944

##               Estimate  Std. Error   t value
## (Intercept) 1.83876433 0.119841404 15.343314
## year94      0.01227337 0.004831677  2.540188

##  Number of Level Two groups =  112
##  Number of Level Three groups =  81
```

- $\pi_{0ij} = 1.8388$: Initial status for school i in postcode j (*i.e.* expected log enrolments when time $= 0$)
- $\pi_{1ij} = 0.0123$: Growth rate for school i in postcode j
- $\epsilon_{tij} = 0.2540$: Variance in within-school residuals after accounting for linear growth overtime

When the subject was first introduced in 1994, schools were expected to have 6.2888 ($e^{1.8388}$) enrolments, on average. As displayed in Figure **??**, the average enrolments for agricultural science is relatively low, where the enrolments during the old QCE system had approximately 10 enrolments while the new QCE system had roughly 15 enrolments per school, on average.

Enrolments were expected to increase by 1.2349% ($(e^{0.01227} - 1) \times 100$) per year. The estimated within-schools variance decreased by 12.968% (0.2918 to 0.2539595), implying that 12.968% 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	npar	AIC	BIC	logLik
model4.3	13	4841.696	4919.622	-2407.848
model4.6	15	4841.897	4931.812	-2405.949
model4.5	15	4843.214	4933.129	-2406.607
model4.1	17	4843.639	4945.542	-2404.819
model4.0	19	4847.303	4961.195	-2404.651
model4.10	14	4848.524	4932.444	-2410.262
model4.9	14	4848.524	4932.444	-2410.262
model4.2	14	4848.524	4932.444	-2410.262
model4.8	14	4848.524	4932.444	-2410.262
model4.4	14	4850.301	4934.221	-2411.151
model4.7	16	4850.361	4946.270	-2409.181

As outlined in step 6, `sector` and `unit` will be added as predictors to the model. The largest possible model (`model4.0`) will then be fitted, before removing the fixed effects one at a time (with `model4.10` being the smallest of all 10 candidate models), while recording the AIC for each model. `model4.3` appears to have the optimal (smallest) AIC (Table 2), and will be used in the next section 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)
11	4837.855	4903.793	-2407.928	4815.855	NA	NA	NA
13	4841.696	4919.622	-2407.848	4815.696	0.1593719	2	0.718

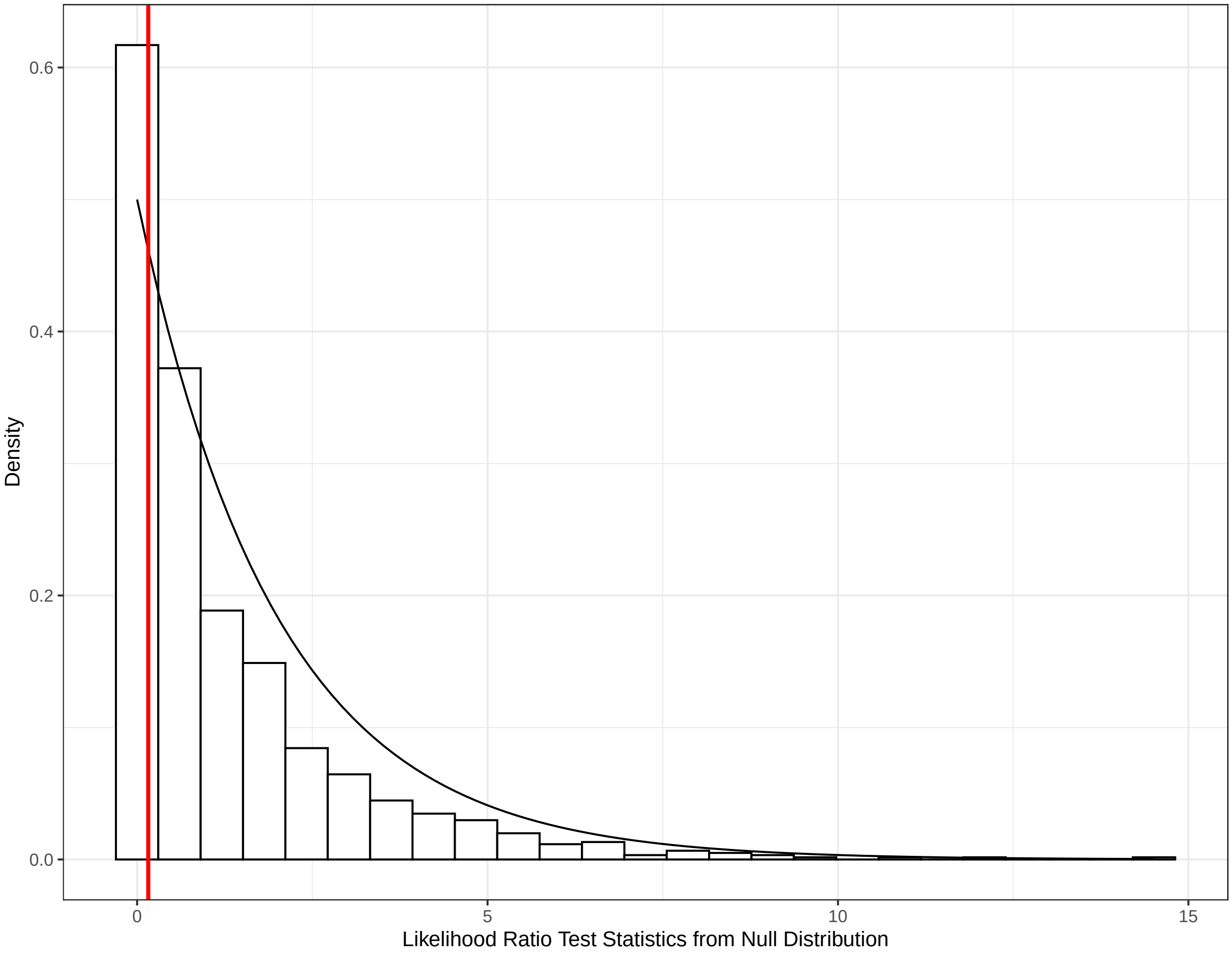


Figure 3: Histogram of likelihood ratio test statistic, with a red vertical line indicating the likelihood ratio test statistic for the actual model

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 indicates the likelihood ratio test statistic using the actual data.

The p-value of 0.718% (Table 3) indicates the proportion of times in which the bootstrap test statistic is greater than the observed test statistic. The large estimated p-value is $0.718 < 0.05$ fails to reject the null hypothesis at the 5% level, indicating that the smaller model (without random slope at level three) is preferred.

Confidence interval

\begin{table}[H]

\caption{95% confidence intervals for fixed and random effects in the final model}

var	2.5 %	97.5 %
sd_(Intercept)|school_postcode:qcaa_school_id	0.8530526	1.2182627
cor_year94.(Intercept)|school_postcode:qcaa_school_id	-0.9690616	-0.8565112
sd_year94|school_postcode:qcaa_school_id	0.0315056	0.0460868
sd_(Intercept)|school_postcode	0.0000009	0.4581153
sigma	0.4876907	0.5140716
(Intercept)	1.4227924	2.2504155
year94	0.0005497	0.0184106
sectorGovernment	-0.3117705	0.4153108
sectorIndependent	-0.0628937	0.7950850
unityear_12_enrolments	-0.2747165	-0.1289871
year94:unityear_12_enrolments	0.0026651	0.0107254

\end{table}

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, and the random effects all exclude 0, further reiterating that they are statistically significant at the 5% level. Some fixed effects such as `unityear_12_enrolments` were insignificant, suggesting that there were no differences between unit 11 and unit 12 units.

Interpreting final model

Composite model

- Level one (measurement variable)

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} year94_{tij} + \epsilon_{tij}$$

- Level two (schools within postcodes)

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

- Level three (postcodes)

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

Therefore, the composite model can be written as:

$$
\begin{aligned}
Y_{tij} &= \pi_{0ij} + \pi_{1ij} year94_{tij} + \epsilon_{tij} \\
&= (\beta_{00j} + \beta_{01j} sector_{ij} + \beta_{02j} unit_{ij} + u_{0ij}) + (\beta_{10j} + \beta_{11j} unit_{ij} + u_{1ij}) year94_{tij} + \epsilon_{tij} \\
&= \ [\gamma_{000} + r_{00j} + (\gamma_{010} + r_{01j}) sector_{ij} + (\gamma_{020} + r_{02j}) unit_{ij} + u_{0ij}] + [\gamma_{100} + \gamma_{110} unit_{ij} + u_{1ij}]\, year94_{tij} + \epsilon_{tij} \\
&= \ [\gamma_{0000} + \gamma_{010} sector_{ij} + \gamma_{020} unit_{ij} + \gamma_{100} year94_{tij} + \gamma_{110} unit_{ij} year94_{tij}] + [r_{00j} + r_{01j} sector_{ij} + r_{02j} unit_{ij} + u_{0ij} + u_1
\end{aligned}
$$

### Fixed effects

```
summary(model_f)
```

```
##  Groups                         Name        Variance  Std.Dev. Corr
##  school_postcode:qcaa_school_id (Intercept) 1.0717490 1.035253
##                                 year94      0.0015528 0.039406 -0.916
##  school_postcode                (Intercept) 0.0812513 0.285046
##  Residual                                   0.2507071 0.500707

##                                   Estimate  Std. Error    t value
## (Intercept)                    1.821947999 0.198742787  9.1673667
## year94                         0.009460629 0.004666531  2.0273366
## sectorGovernment               0.048533174 0.176743531  0.2745966
## sectorIndependent              0.364311290 0.196685041  1.8522572
## unityear_12_enrolments        -0.204347826 0.037573894 -5.4385587
## year94:unityear_12_enrolments  0.006683023 0.002232326  2.9937484

##  Number of Level Two groups =  112
##  Number of Level Three groups =  81
```

Notably, the model assumes that enrolments in different postcodes are assumed to increase at the same rate (as justified by the parametric bootstrap while testing for random effects). Using the model output above (see step 9 for detailed explanation on fixed effects), the estimated increase in mean enrolments for schools between postcodes are estimated to increase at a rate of 0.9506% per year.

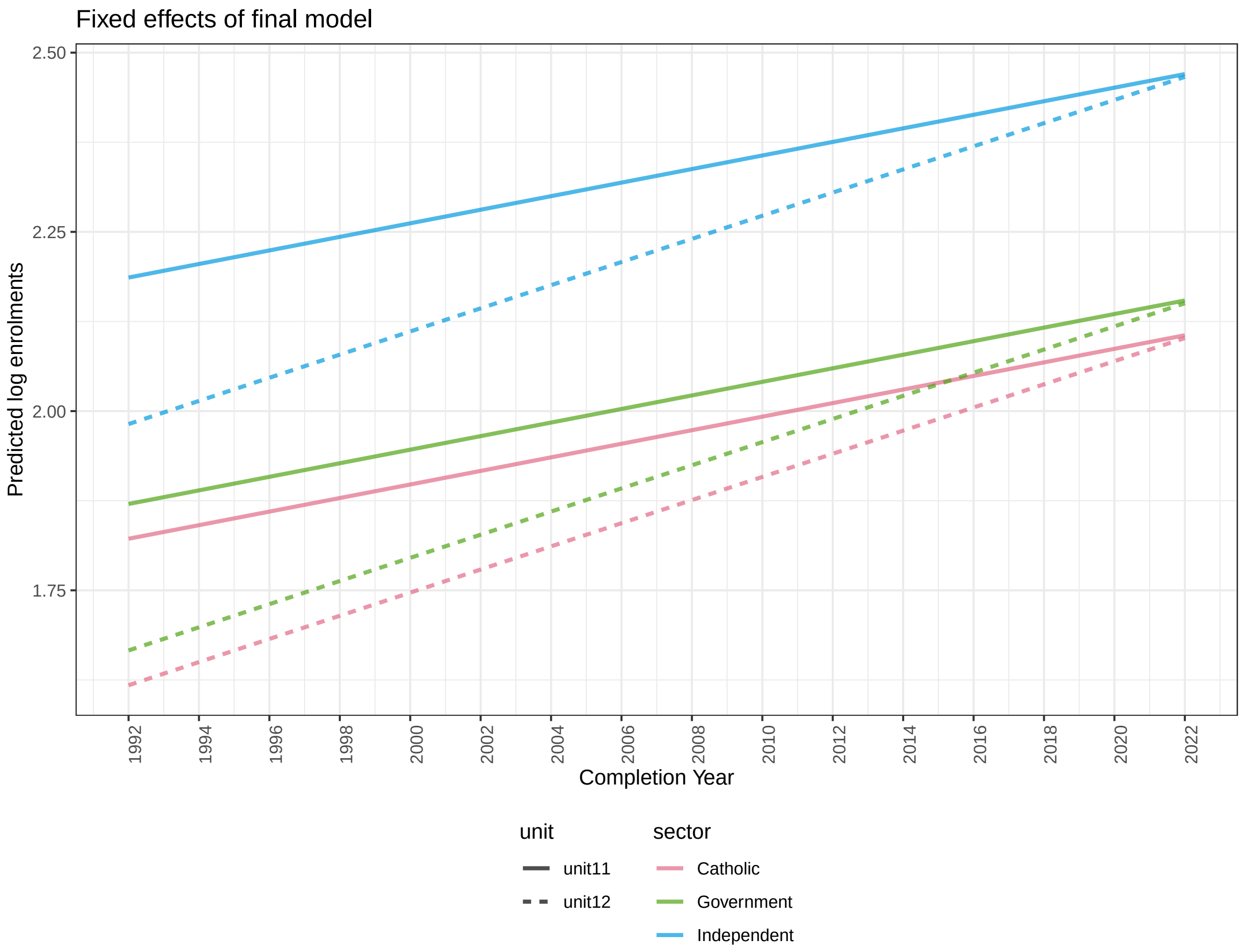
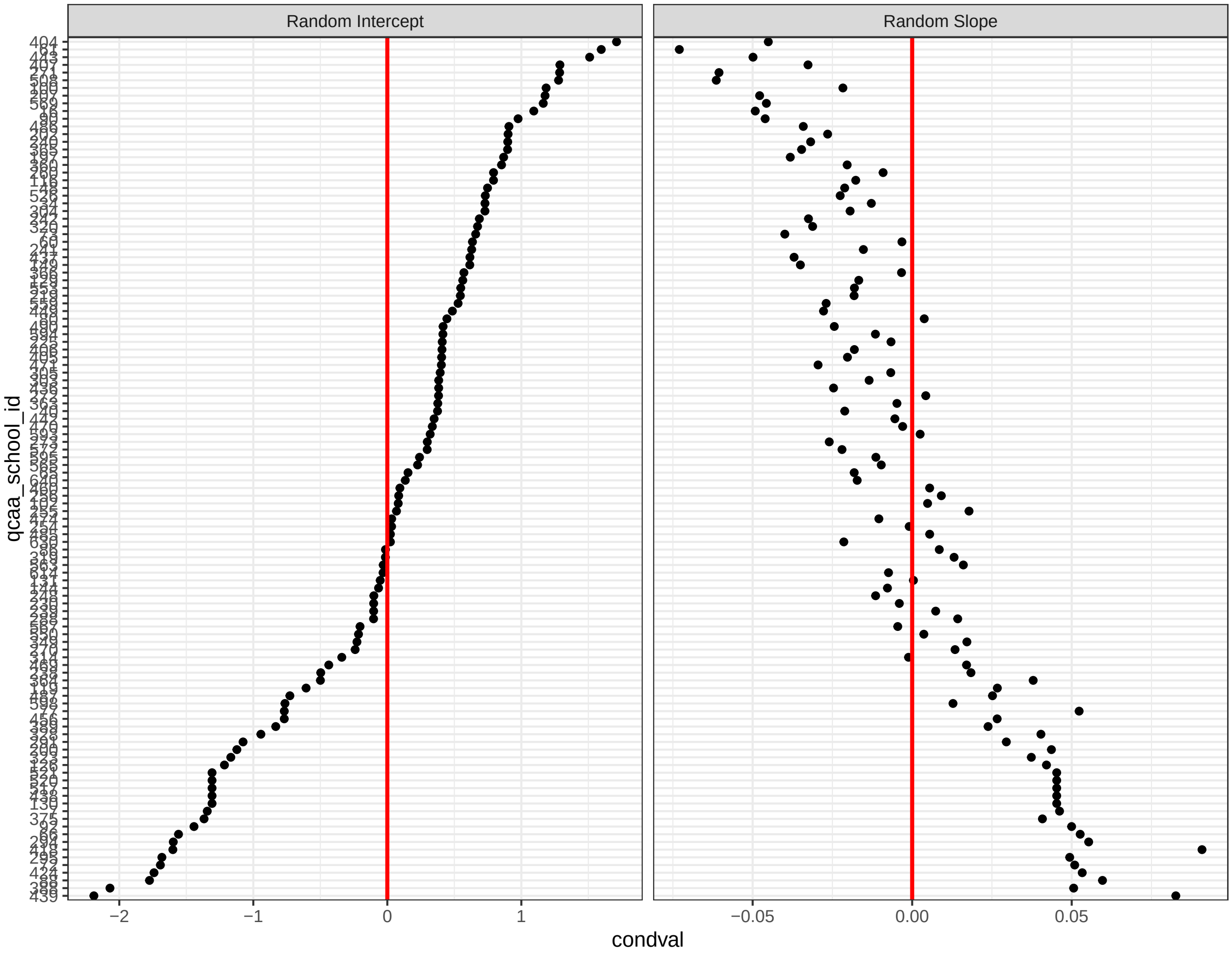


Figure 4: Fixed effects of the final model for Agricultural Science

Based on the fixed effects, independent schools are expected to have the highest enrolments over the years. In all sectors, unit 12 have lower initial status (*i.e.* lower enrolments when subject is first introduced), however, they are expected to increase at a much higher rate relative to unit 11 units. Intuitively, this suggests that on average, there may be more students taking this subject in year 10, before re-enrolling again in year 12, or in the more extreme case, students are dropping the subject after completing year 11.

Random effects



A clear negative correlation in the random intercept and slope can be distinguished, indicating that schools with larger enrolments when the subject was first introduced are expected to show a smaller increase (decrease) in enrolments over the years. Based on the model output, the correlation between random intercept and slope was -0.92.

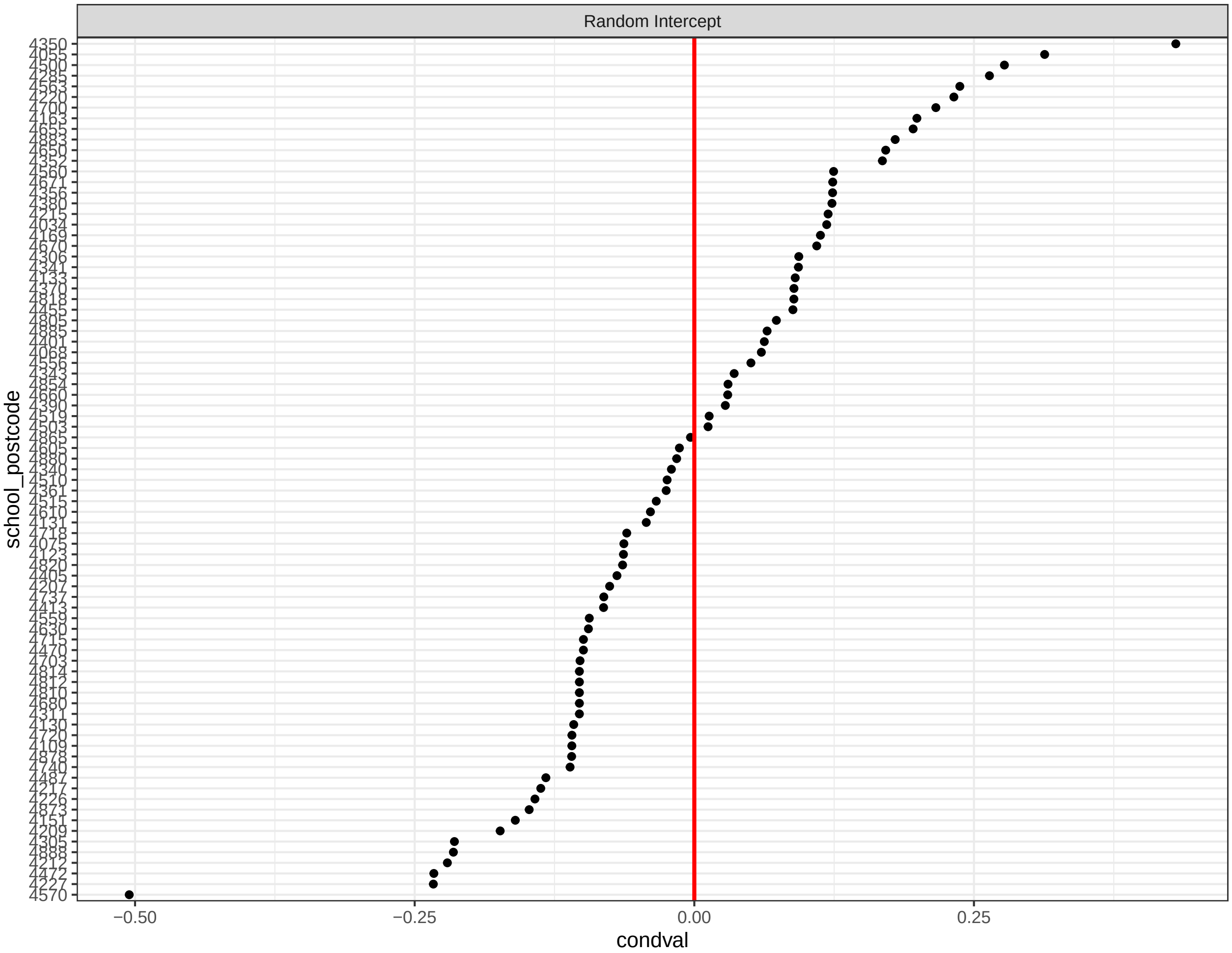


Figure 5: Random intercept for districts

As there is no random slope, enrolments for the across all postcodes are estimated to be the same (justified by the parametric bootstrap). As shown in Figure 5, most of the variations are accounted for in the random intercept, which suggests that some postcodes are associated with larger schools (and enrolments) relative to other postcodes. Schools within postcode 4350 (In Toowoomba district) are predicted to have the highest initial status while schools within postcode 4570 (Wide Bay district) are predicted to have the lowest initial status.

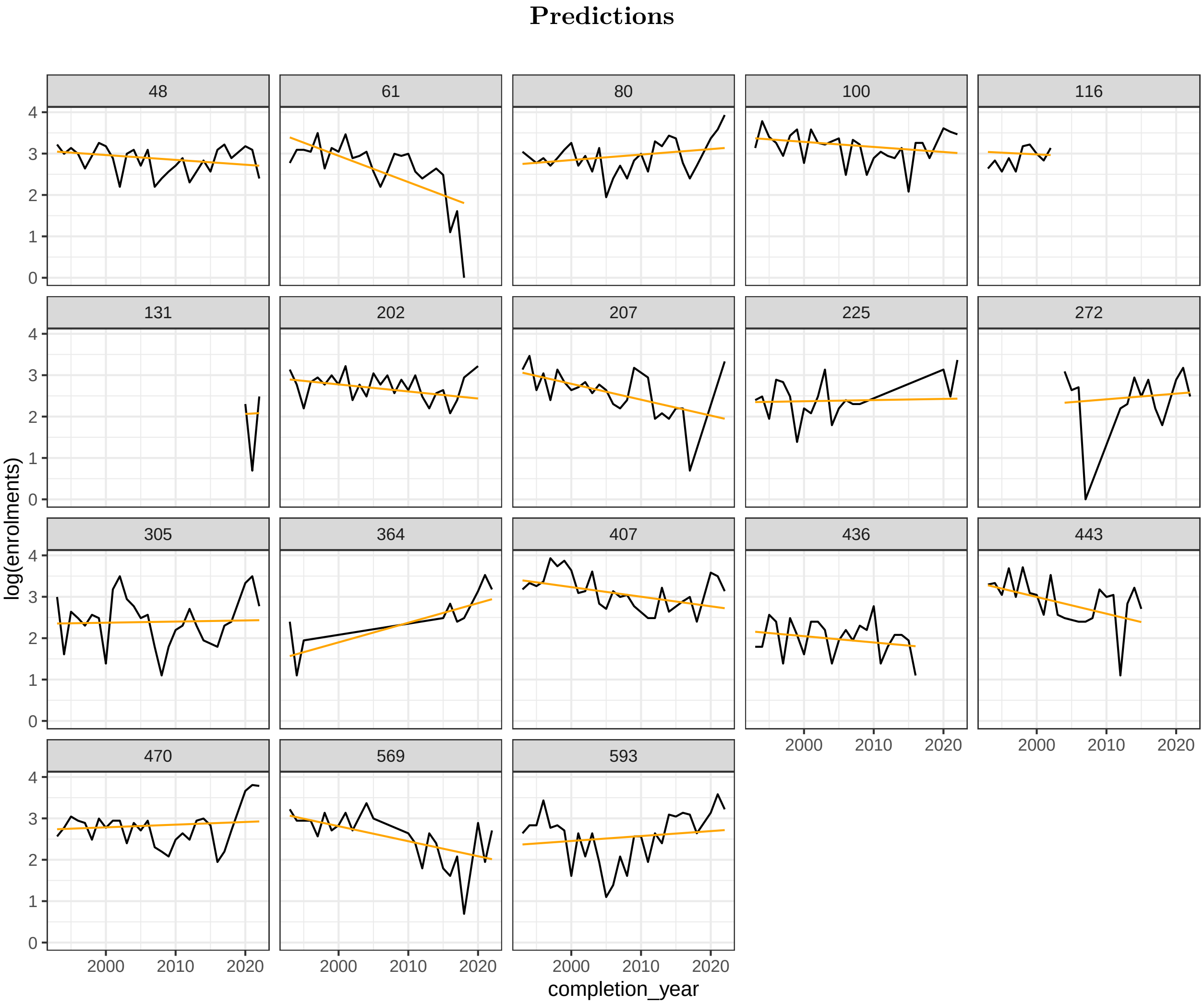


Figure 6: Model predictions for 20 randomly selected schools