Multilevel Model for Psychology

Brendi Ang

17/10/2021

Contents

Psychology

Getting the data ready for modelling

Exploring the dataset with basic linear model for each school

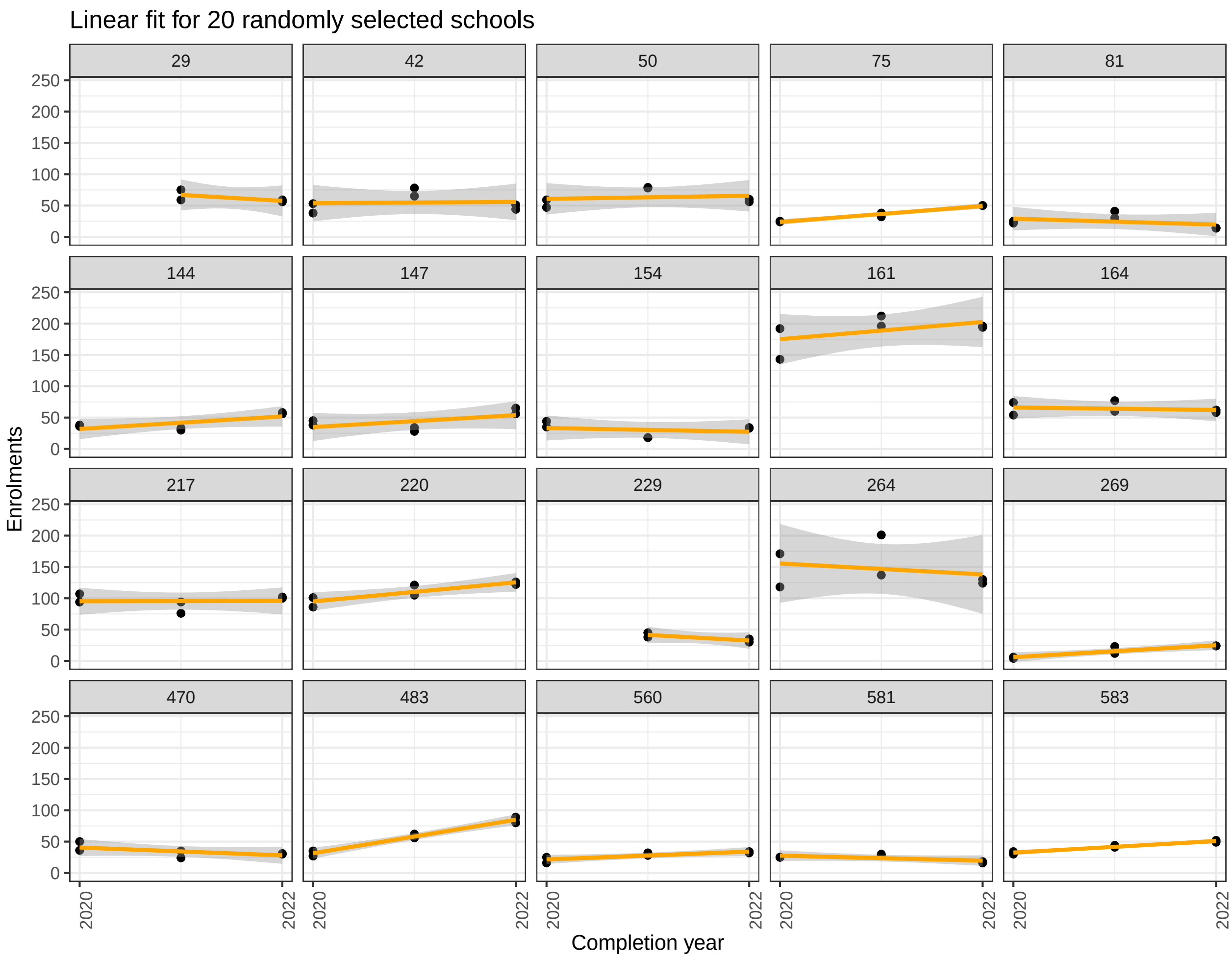
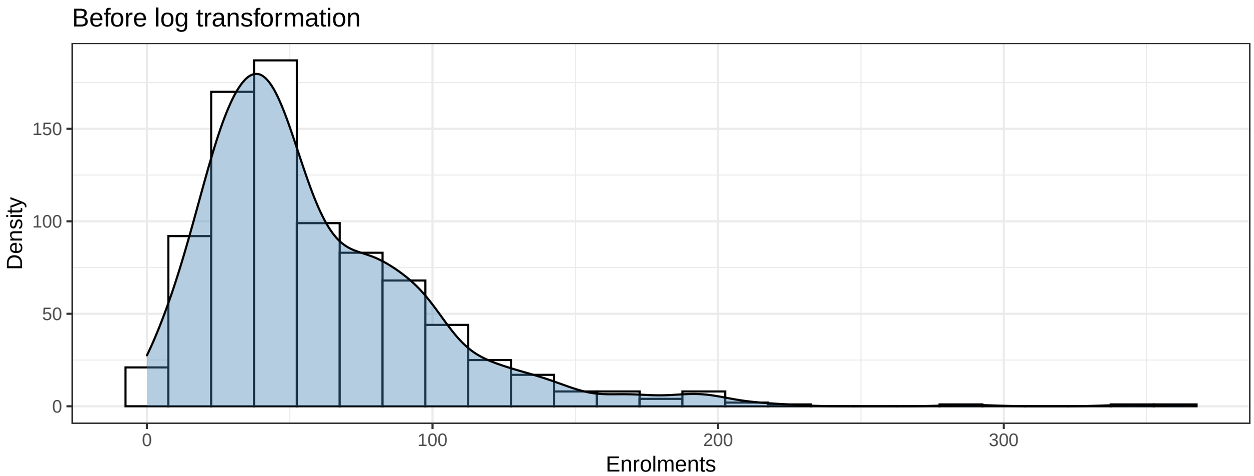
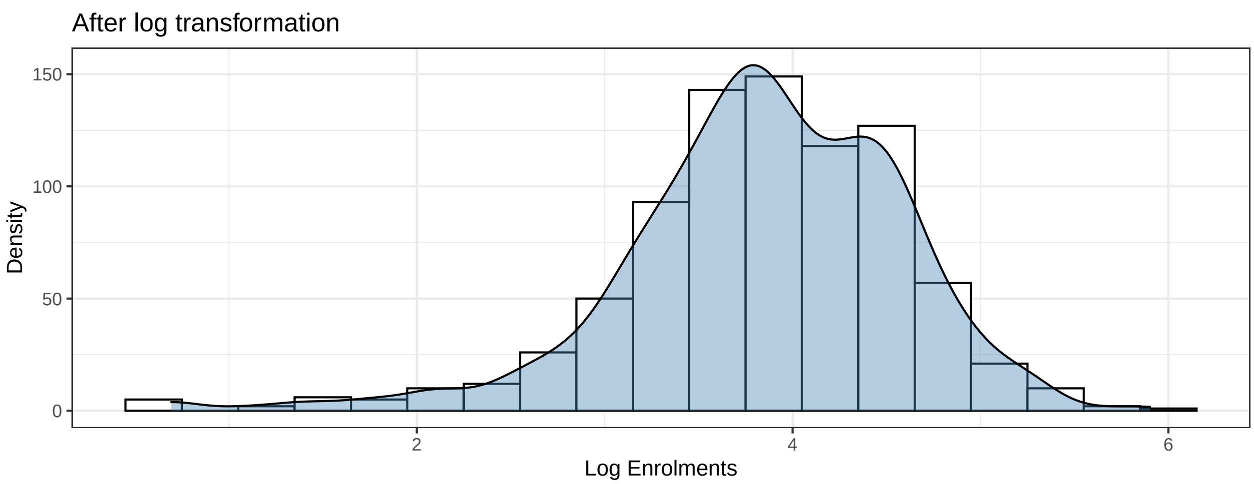


Figure 1: Basic linear model for 20 randomly selected schools to provide an at-a-glance visualisation of enrolment trends within schools for Psychology

As aforementioned, Psychology was one of the few subjects that were introduced in the new QCE system. As seen in Figure **??**, there were only observations for 3 consecutive years (up till cohort graduating in 2022). The linear fit indicates that the enrolment trends for these 20 randomly selected schools were rather constant. Schools differ in cohort size, for instance, schhool 264 and 161 have enrolments above 200 for a single cohort, whereas school 560 and 581 have less than 50 enrolments in a single cohort.

Linearise response variable using log transformation





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 Agricultural Practices

	df	AIC
Model0.2: Schools nested within districts	4	1007.297
Model0.0: Within schools	3	1015.764
Model0.1: Schools nested within postcodes	4	1017.764

Referring back to step 3, three candidate models are fitted, with the AIC shown in Table 1. `Model0.2`, corresponding to having schools nested within districts is the best model, with optimised (lowest) AIC and will be used in the subsequent analysis.

Intraclass correlation (ICC)

```
summary(model)
```

```
## Random effects:

##  Groups                        Name        Variance Std.Dev.
##  qcaa_school_id:qcaa_district (Intercept) 0.482163 0.69438
##  qcaa_district                (Intercept) 0.090152 0.30025
##  Residual                                 0.100377 0.31682

##
##  Fixed effects:

##             Estimate Std. Error  t value
## (Intercept) 3.740694  0.1025459 36.47823

##
##  Number of schools (level-two group) = 169
##  Number of district (level-three group) = 13
```

This model takes into account 169 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 a certain district k in time t and in time t^*:

$$\text{level-two ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.48216}{(0.48216 + 0.09015 + 0.10038)} = 0.7168$$

This can be conceptualised as the correlation between enrolments of two random draws from the same school at two different years. In other words, 71.68% of the total variability is attributable to the differences between schools from the same district rather than changes over time within schools.

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

$$\text{level-two ICC} = \frac{\phi_{00}^2}{\tau_{00}^2 + \phi_{00}^2 + \sigma^2} = \frac{0.09015}{(0.48216 + 0.09015 + 0.10038)} = 0.1340$$

Unconditional Growth model

```
##  Groups                        Name        Variance   Std.Dev. Corr
##  qcaa_district:qcaa_school_id (Intercept) 0.56752267 0.753341
##                                year20      0.06229629 0.249592 -0.345
##  qcaa_district                (Intercept) 0.09609323 0.309989
##                                year20      0.00038629 0.019654 -0.374
##  Residual                                  0.05684574 0.238423
```

```
##              Estimate Std. Error   t value
## (Intercept) 3.63652326 0.10922320 33.294423
## year20      0.08484016 0.02471232  3.433111
```

```
##  Number of Level Two groups =  169
##  Number of Level Three groups =  13
```

The unconditional growth model adds the systematic changes over time, the model specification can be found in step 4. This allows for assessing within-school variability which can be attributed to the linear changes over time. Based on the model output:

- $\pi_{0ij} = 3.6365$: Initial status for school i in district j (*i.e.* expected log enrolments when time $= 0$)
- $\pi_{1ij} = 0.0848$: Growth rate for school i in district j
- $\epsilon_{tij} = 0.0568$: Variance in within-school residuals after accounting for linear growth overtime

Psychology was introduced in 2020, and schools are expected to have 37.9587 ($e^{3.7837}$), on average. Furthermore, enrolments were expected to increase by 8.8499% ($e^{0.0848} - 1) \times 100$) every year. The estimated within-school variance decrease by 43.3695% (0.10038 to 0.0568457), implying that 43.3695% of the within-school variability can be explained by the linear growth over time.

Testing fixed effects

```
## # A tibble: 11 x 5
##    model      npar   AIC   BIC logLik
##    <chr>     <dbl> <dbl> <dbl>  <dbl>
##  1 model4.6     15  770.  841.  -370.
##  2 model4.3     13  771.  832.  -372.
##  3 model4.1     17  773.  853.  -369.
##  4 model4.5     15  774.  845.  -372.
##  5 model4.0     19  776.  866.  -369.
##  6 model4.2     14  786.  852.  -379.
##  7 model4.8     14  786.  852.  -379.
##  8 model4.10    14  786.  852.  -379.
```

```
## 9 model4.9      14  786.  852.  -379.
## 10 model4.7     16  789.  865.  -379.
## 11 model4.4     14  789.  856.  -381.
```

As detailed in step 6, level-two predictors (`sector` and `unit`) are added to the model. The largest possible model will be fitted, before removing each fixed effect one by one whilst recording the AIC for each model. `model4.0` corresponds to the largest model while `model4.10` is the smallest possible model. The model with the optimal (lowest) AIC is the largest possible model `model4.6`, and will be used in subsequent sections.

Parametric bootstrap to test random effects

Table 2: Parametric Bootstrap to compare larger and smaller, nested model

npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr_boot(>Chisq)
13	766.0029	827.4906	-370.0014	740.0029	NA	NA	NA
15	770.0029	840.9502	-370.0014	740.0029	1.77e-05	2	0.829

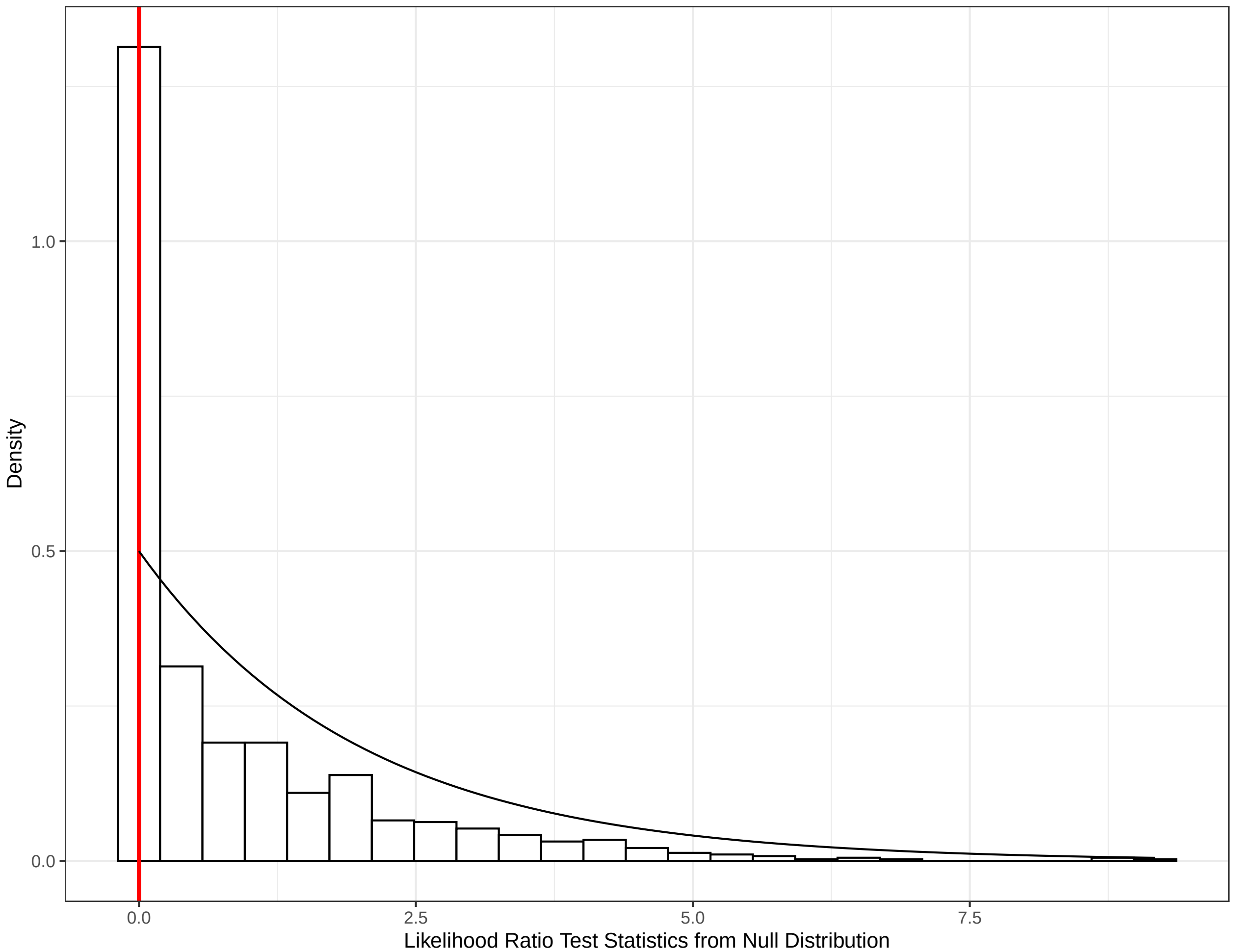


Figure 2: Histogram of likelihood ratio test statistic, with a red vertical line indicating the likelihood ratio test statistic for the actual model

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 2 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 indicates the proportion of times in which the bootstrap test statistic is greater than the observed test statistic. The low estimated p-value is $0.829 < 0.05$ (Table 2) fails to reject the null hypothesis at the 5% level, indicating that the smaller model (excluding random slope at level three) is preferred.

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	0.6379862	0.8190212
cor__year20.(Intercept)|qcaa__district:qcaa__school__id	-0.5087693	-0.1687466
sd__year20|qcaa__district:qcaa__school__id	0.2193296	0.2988031
sd__(Intercept)|qcaa__district	0.0999599	0.4638999
sigma	0.2079300	0.2349048
(Intercept)	3.4710604	4.0367595
year20	-0.0132011	0.0985917
sectorGovernment	-0.0469960	0.5121471
sectorIndependent	-0.6487917	-0.0311876
unityear__12__enrolments	-0.2390228	-0.0974500
year20:unityear__12__enrolments	0.0390586	0.1191998
sectorGovernment:unityear__12__enrolments	-0.1435010	0.0093553
sectorIndependent:unityear__12__enrolments	-0.0911536	0.0578887

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)

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} year20_{tij} + \epsilon_{tij}$$

- Level two (schools within districts)

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

- 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

$$\begin{aligned}
Y_{tij} &= \pi_{0ij} + \pi_{1ij} year20_{tij} + \epsilon_{tij} \\
&= (\beta_{00j} + \beta_{01j} sector_{ij} + \beta_{02j} unit_{ij} + \beta_{03j} sector_{ij} unit_{ij} + u_{0ij}) + \\
&\quad (\beta_{10j} + \beta_{11j} unitij) year20_{tij} + \epsilon_{tij} \\
&= [\gamma_{000} + r_{00j} + (\gamma_{010} + r_{01j}) sector_{ij} + (\gamma_{020} + r_{02j}) unit_{ij} + (\gamma_{030} + r_{03j}) sector_{ij} unit_{ij} + u_{0ij}] + \\
&\quad [\gamma_{100} + \gamma_{110} unit_{ij} + u_{1ij}]\, year20_{tij} + \epsilon_{tij} \\
&= [\gamma_{000} + \gamma_{010} sector_{ij} + \gamma_{020} unit_{ij} + \gamma_{030} sector_{ij} unit_{ij} + \gamma_{100} year20_{tij} + \gamma_{110} unit_{ij} year20_{tij}] \\
&\quad [r_{00j} + r_{01j} sector_{ij} + r_{02j} unit_{ij} + r_{03j} sector_{ij} unit_{ij} + u_{0ij} + u_{1ij} year20_{tij} + \epsilon_{tij}]
\end{aligned}$$

Fixed effects

```
summary(model_f)
```

```
##  Groups                        Name        Variance Std.Dev. Corr
##  qcaa_district:qcaa_school_id (Intercept) 0.528768 0.72716
##                                year20      0.066686 0.25824  -0.363
##  qcaa_district                (Intercept) 0.087848 0.29639
##  Residual                                  0.049415 0.22229
```

```
##                                            Estimate Std. Error    t value
## (Intercept)                              3.74629239 0.14131246 26.5107005
## year20                                   0.04412463 0.02600532  1.6967542
## sectorGovernment                         0.23181923 0.13930364  1.6641291
## sectorIndependent                       -0.33032399 0.14733210 -2.2420368
## unityear_12_enrolments                  -0.17061155 0.03679803 -4.6364314
## year20:unityear_12_enrolments            0.08166456 0.01906433  4.2836322
## sectorGovernment:unityear_12_enrolments  -0.07329452 0.03873036 -1.8924307
## sectorIndependent:unityear_12_enrolments -0.01312105 0.04090607 -0.3207605
```

```
##  Number of Level Two groups =  169
##  Number of Level Three groups =  13
```

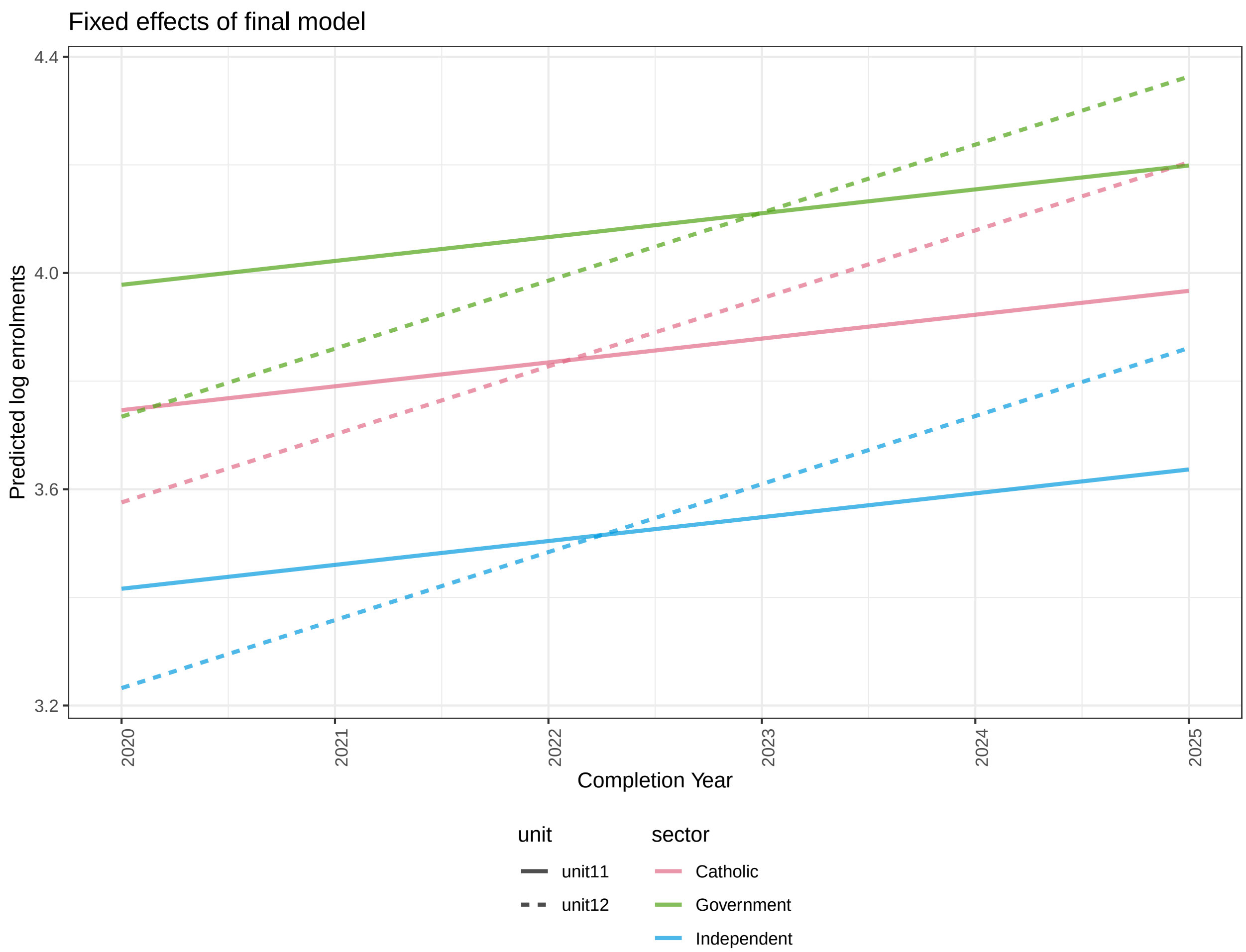
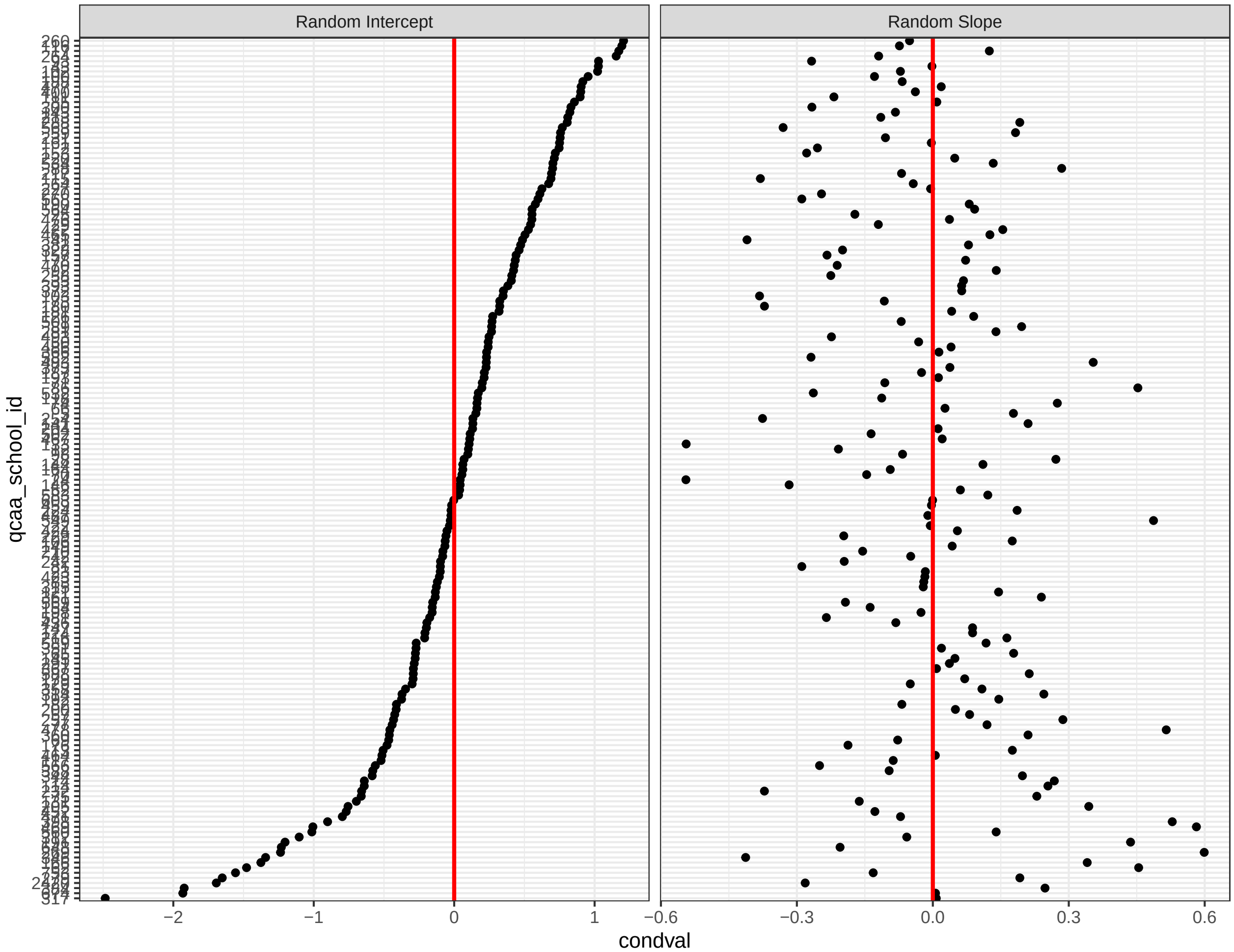


Figure 3: Fixed effects of the final model for Agricultural Practices subject

As the final model does not have a random slope at level 3, schools across sector are assumed to have the same slope in the fixed effects, but at different intercept (baseline). As seen in Figure 3, unit 12 are expected to be more than unit 11 in the later years, suggesting that students may be completing the year 11 syllabus in year 10, before re-enrolling in year 12. It appears that government schools are estimated to have the highest enrolment in Psychology, with independent schools having the least enrolments, on average.

Random effects



As shown in the model output and in Figure **??**, there was little correlation (-0.36) between the random intercept and slope. Schools with appear to decrease or increase in enrolments at different rates; This may be attributed to the fact that the data only consists of observations for three years and there is not enough data to evaluate the enrolment patterns in schools.

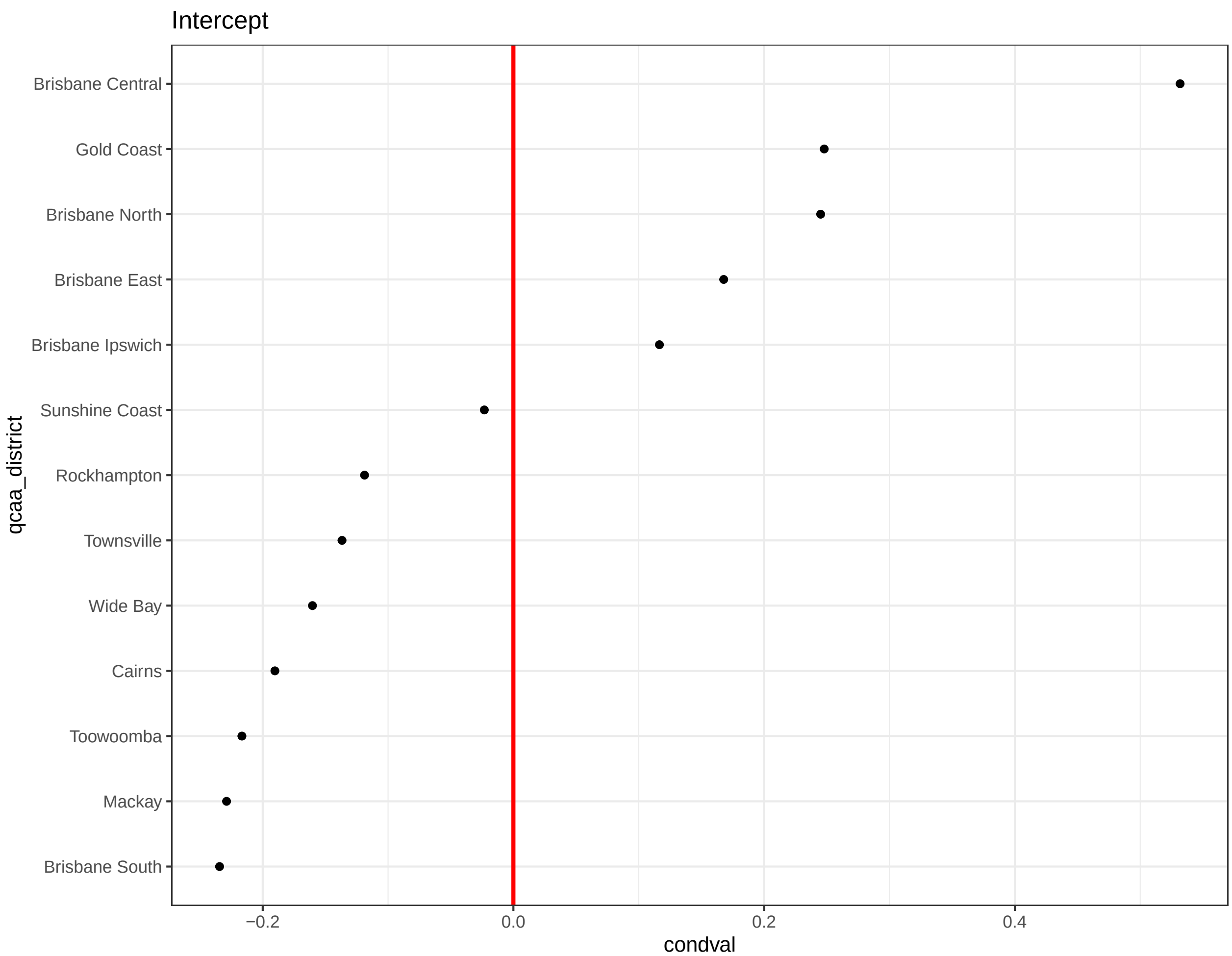


Figure 4: Model predictions for 20 randomly selected schools

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 4 demonstrates that schools in Brisbane Central has the largest enrolments while Brisbane South have the lowest enrolments in this subject, on average.

Predictions

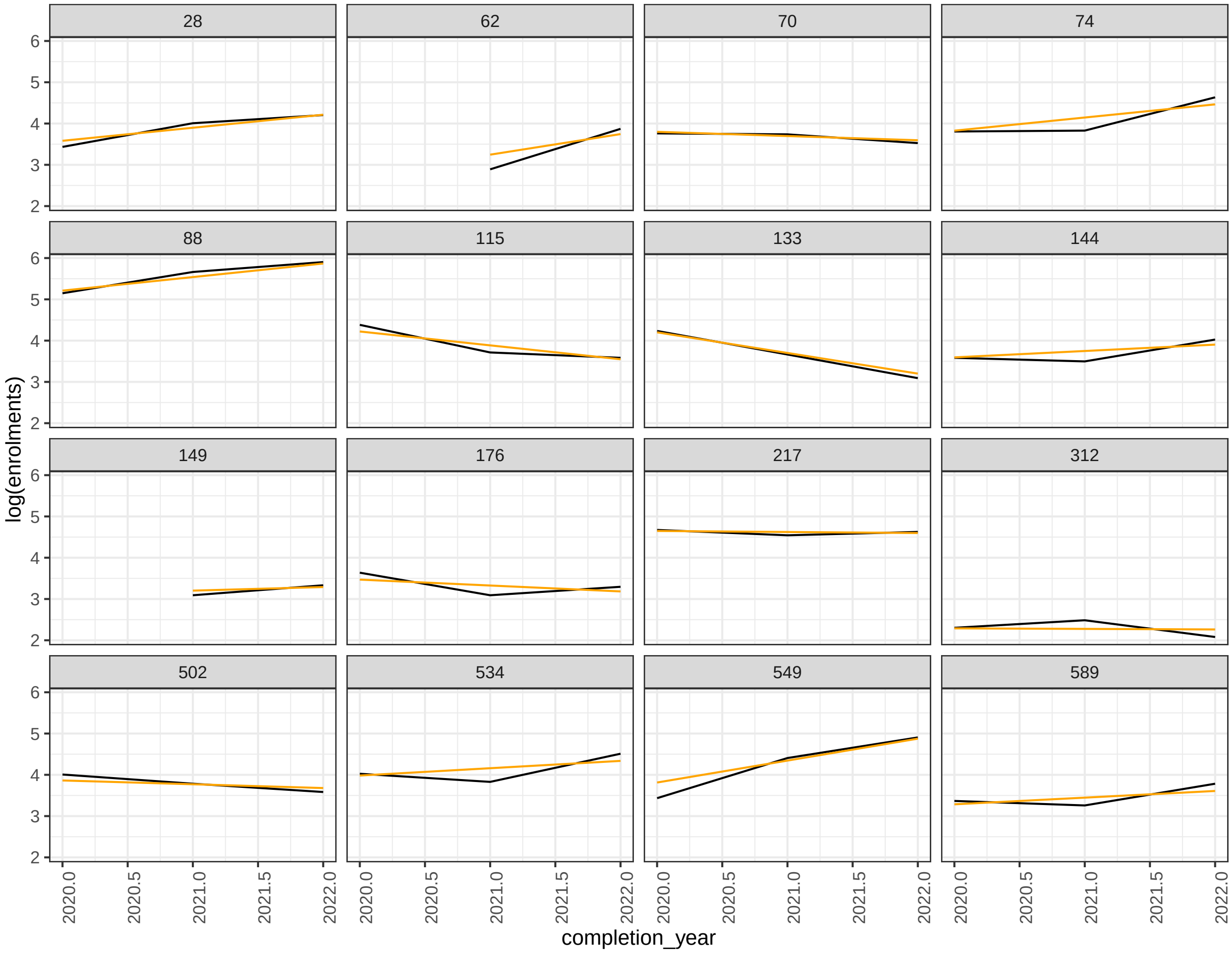


Figure 5: Model predictions for 20 randomly selected schools

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