Git repository for code for this assignment

1 Histograms/Tables: "...Histograms.R"

1.1 Students in each school

To create the histogram for the number of students in each school, I first had to generate the data since it is not directly included. While there are many ways to do this, I settled on an implementation which counted the number of observations which corresponded to each school. This sum is equivalent to the number of students in each school since each observation represents a unique student. To get the histogram, I used R's standard `hist` function with Scott's rule bin breaks. While I would typically use a discrete-oriented histogram function, the number of schools was not large enough to result in a continuous approximation. The result is in Figure 1.

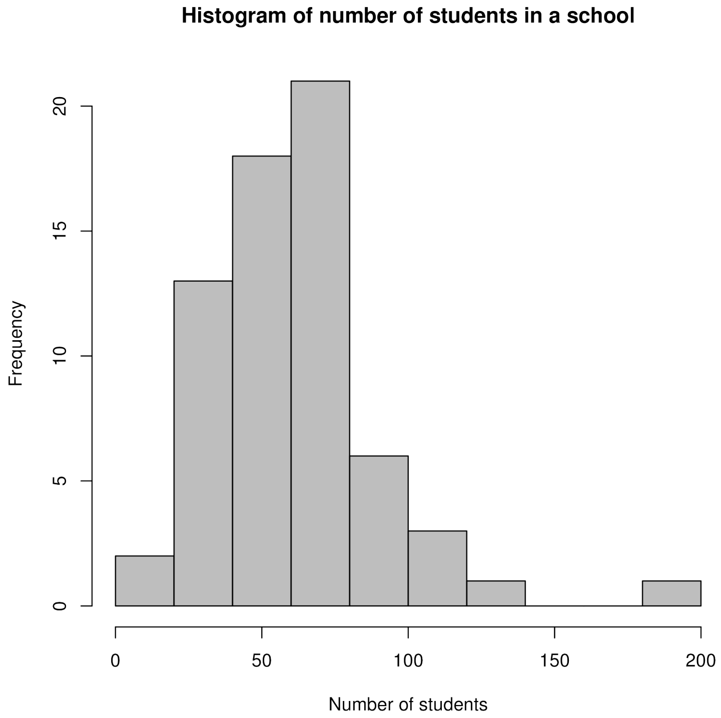

Figure 1

A few interesting features of the histogram/data:

- Center of mass around 60 students
- One school with about 200 (198) students
- One school with 2 students and one school with 8 students

1.2 GCSE exam scores

Creating the normalized exam score histogram was more straightforward. After a preliminary histogram, no outliers appeared, so I again used Scott's rule for the bin widths with the standard `hist()` function:

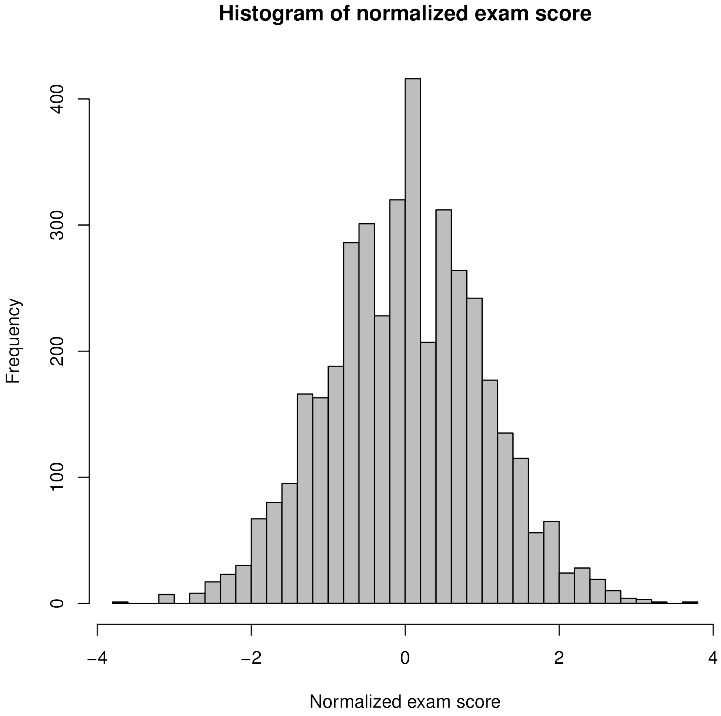

Figure 2

As shown in Figure 2, the normalized exam scores appear mostly normally distributed. The maximum deviation from the mean exam score appears to be about 4 standard deviations in either direction. Another interesting aspect of the histogram are the two dips about 0.5 standard deviations from the mean exam score. These dips could be masked with wider bin widths (using Sturge's rule for the bins accomplishes this).

1.3 Average LRT intake score for schools

In this case again I had to manipulate the given data to cull out the observations for the histogram. To get the average London Reading Test intake scores for the schools, I first confirmed that the number of unique average scores was equal to the number of schools. With this assumption confirmed, I grabbed the unique school averages. To create the histogram, I again used Scott's rule

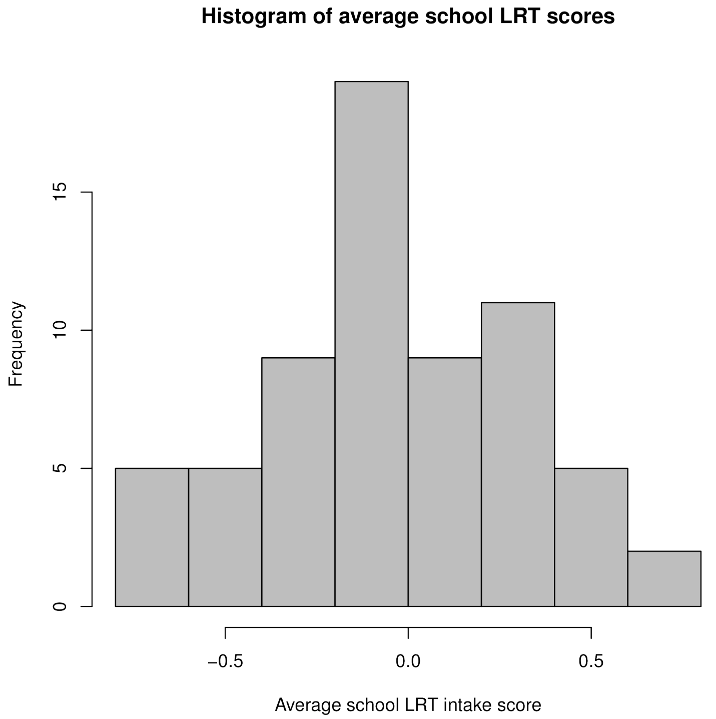

Figure 3

Figure 3 shows that, first, the observations reported are likely normalized test scores (unless students can get negative scores and the average score is magically 0). Next, while the scores are standardized, the distribution is a bit flat overall. The distribution is also fairly compressed. Compared to the standardized GCSE exam scores, the intake scores were at most 0.75 standard deviations away from the mean score. This might be a product of the exam itself. Finally, there appears to be a fairly large bump around the mean LRT intake score.

1.4 Standardized LRT intake scores

Requiring not pre manipulation, I plotted the standardized LRT intake scores for each student. The result, shown in Figure 4 shows that the standardized scores appear normally distributed, with a few dips and a sharp drop over 1 standard deviation below the mean score. While no scores were below three standard deviations from the mean, there appear to be a small bump of scores above three standard deviations from the mean.

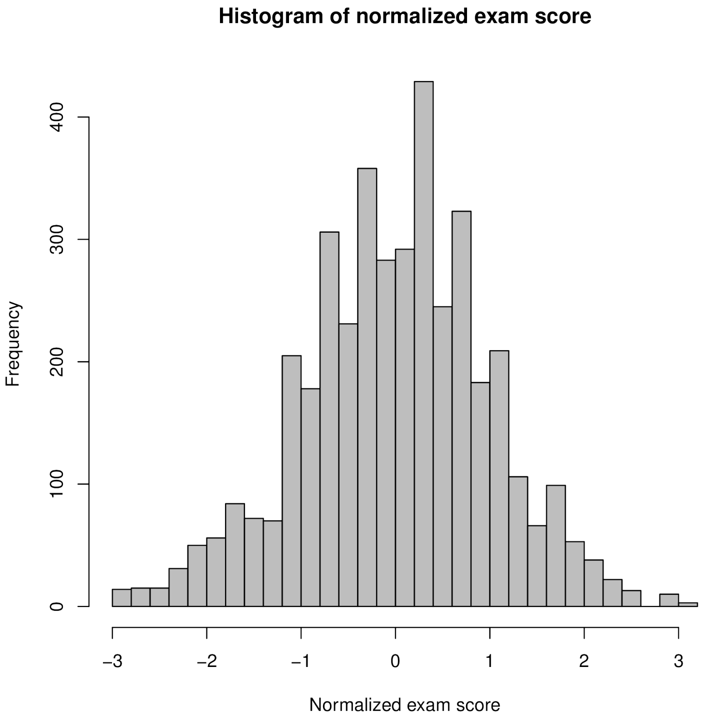

Figure 4

1.5 School gender types

To make a table of the school gender type values, I used the standard R `table()` function. The result in Table 1 shows that over a fourth of the schools are girls-only, over half of the schools are mixed, and only about an eighth of the schools are boys-only. This imbalance could lead to issues using school gender type as a regressor.

Table 1: School gender type

Mixed	**Boys**	**Girls**
2169	513	1377

1.6 Verbal reasoning subscores at intake

As with school gender types, I used the R `table()` function to generate the results in Table 2. Also similar to Table 1, the results are split with over half in one category (mid 50%), over a fourth in another (top 25%), and over an eighth in the last category (bottom 25%). If overlap of the categories is also skewed, we could end up with very few data points for a combination such as "students in an all boys school who performed in the bottom 25% for VR intake".

Table 2: Student verbal reasoning subscores at intake

bottom 25%	**mid 50%**	**top 25%**
640	2263	1156

1.7 Overall LRT scores at intake

Unlike Table 2, the overall LRT scores show a flipped trend in Table 3, where over a fourth of the data are for students in the bottom 25% and just over an eighth are in the top 25%. As these screwed distributions continue to stack up, concern about a lack of data for a cross-section of the data points grows.

Table 3: Student overall London Reading Test scores at intake

bottom 25%	**mid 50%**	**top 25%**
1176	2344	539

1.8 Student's gender

Table 4 shows there are about 50% more female students in the study than male students. While the other factor regressors have had three levels, this factor has two levels. This asserts that the variance in the regression parameter will be more impacted by skew in the category counts.

Table 4: Student gender

Female	**Male**
2436	1623

1.9 Type of schools

Table 5 represents a partial aggregation of the school gender type data. The mixed value is the same between the two factors and the "single" value is the sum of the schools which are boys-only or girls-only. This factor is more balanced than the rest.

Table 5: Student gender

Mixed	**Single**
2169	1890

2 Intercept-only regression (GCSE exam scores): "...Intercepts.R"

2.1 Complete pooling

To find the complete pooling intercept, I conducted a regular linear regression with only a constant independent variable with students GCSE exam scores as the dependent variable (ignoring the clustering of students by schools). The result, as expected, is simply the mean of all the student normalized exam scores which is (about) 0 which is also expected since the exam scores are normalized. To display the result, I plotted the complete pooling mean across each schools data with quartile ranges as shown in Figure 5. As in the lecture notes, the red dashed line is the complete-pooling mean and the solid blue lines represent a 95% confidence interval on the complete-pooling intercept in predicting the school-level intercepts (means). As shown the confidence interval is extremely small relative to the deviations in school-level exam scores.

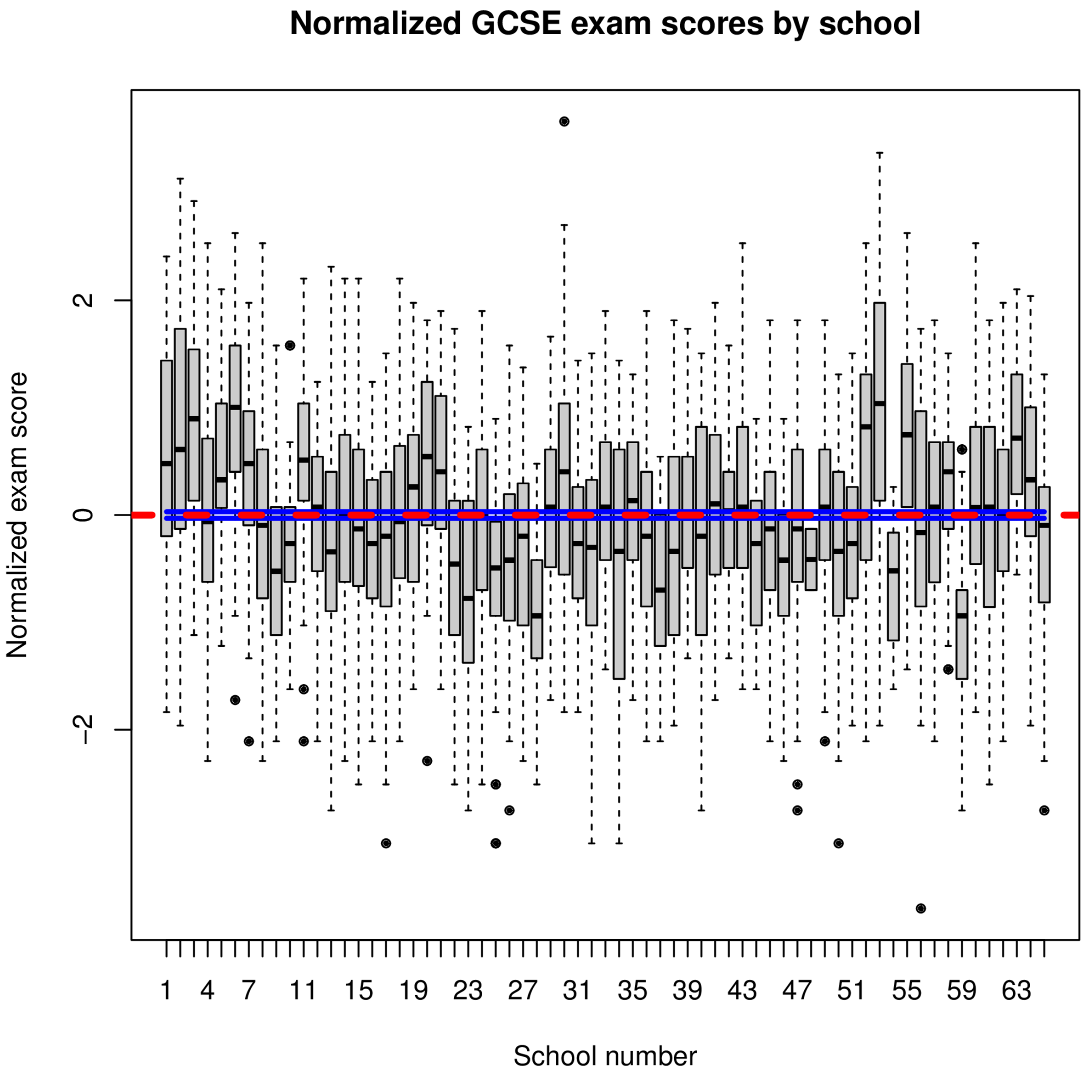

Figure 5: Complete pooling intercept-only regression

2.2 No pooling

To perform the no pooling intercept-only regression on GCSE exam scores, I tried both of the methods used in the notes. The first involved an individual regression on the data from each school using R's `lapply()` function. While this was efficient, it would be difficult to grab the coefficients from the list result. For the sake of easier access to the regression coefficients, I then did a single regression (with a 0 intercept) where the only regressor was a dummy variable for the school number. The coefficients of that single regression are the school-level normalized exam score means. To display the results, I constructed a histogram of the school-level scores and a plot of the standard errors of the means as a function of their value which are both shown in Figure 6. While most of the normalized scores are within 0.5 standard deviations from the mean, about 15 schools are more than 0.5 standard deviations from the mean exam score. There are also two large frequency bumps: one around -0.25 SD from the mean score and another around the mean exam score (though because of the bin placement it appears as though those values are greater than the mean). The standard errors hover between 0.05 and 0.15 standard deviations with a couple of schools which have

SEs larger than 0.25.

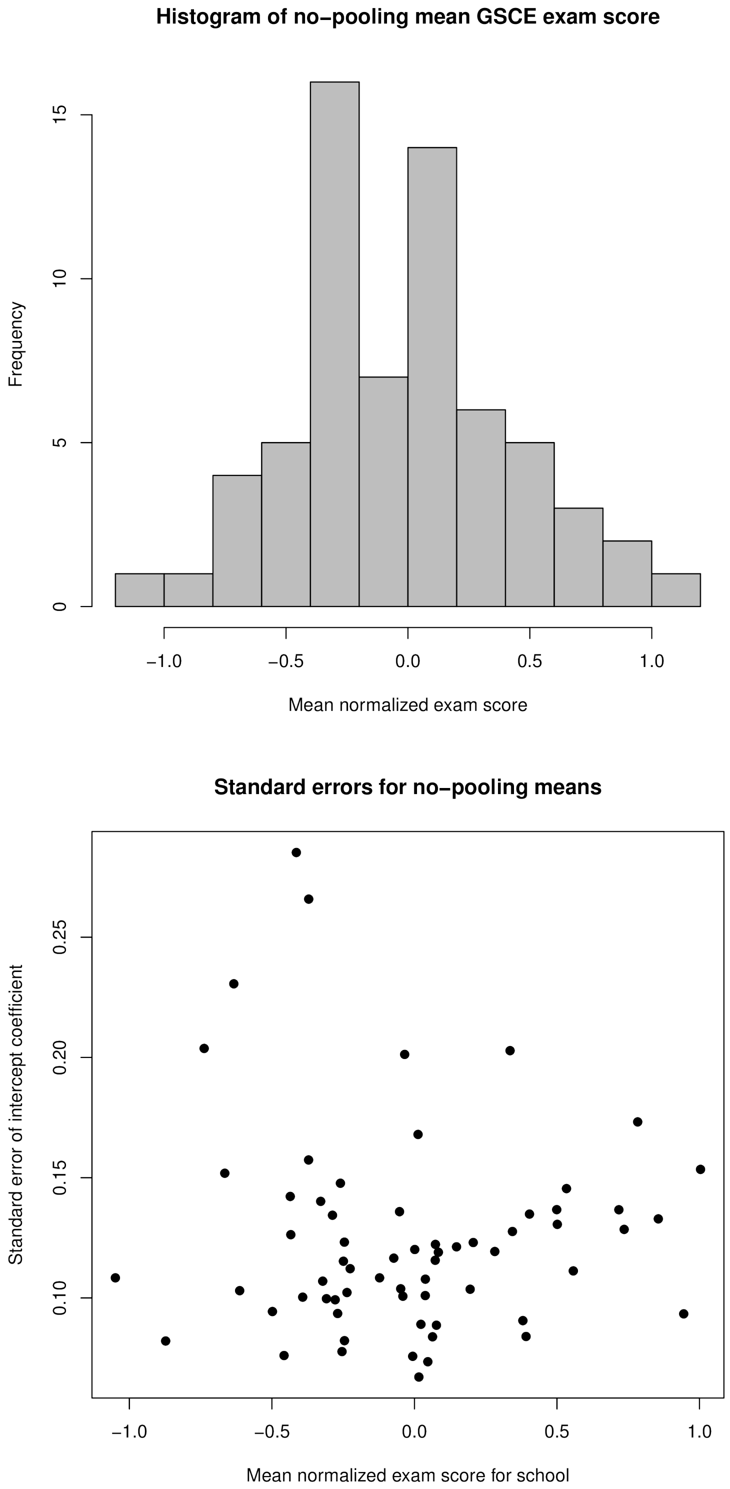

Figure 6: Frequency and standard errors for mean exam scores by school

2.3 Partial pooling

To conduct partial pooling regressions using only an intercept, I used the `lmer()` function of the ”arm” package which allows for mixed effects. I then produced a histogram for the partial pooling normalized mean exam scores by school, shown in Figure 7. The result looks similar to the distribution of the no pooling school means, with deviations above and below the no pooling distribution as the scores are farther from the mean. To better understand the differences in the distributions, I created the QQ plot in Figure 8 which confirms that:

- For schools with mean normalized exam scores lower than the complete pooling mean (0), the partial pooling mean was higher (pulling the values closer to the complete pooling mean)
- For schools with mean normalized exam scores higher than the complete pooling mean, the partial pooling mean was lower (again, pulling the values closer to the complete pooling mean)

This difference is expected because even if the sample size in a given school is large[1], it will still be marginally shifted towards the complete pooling estimate.

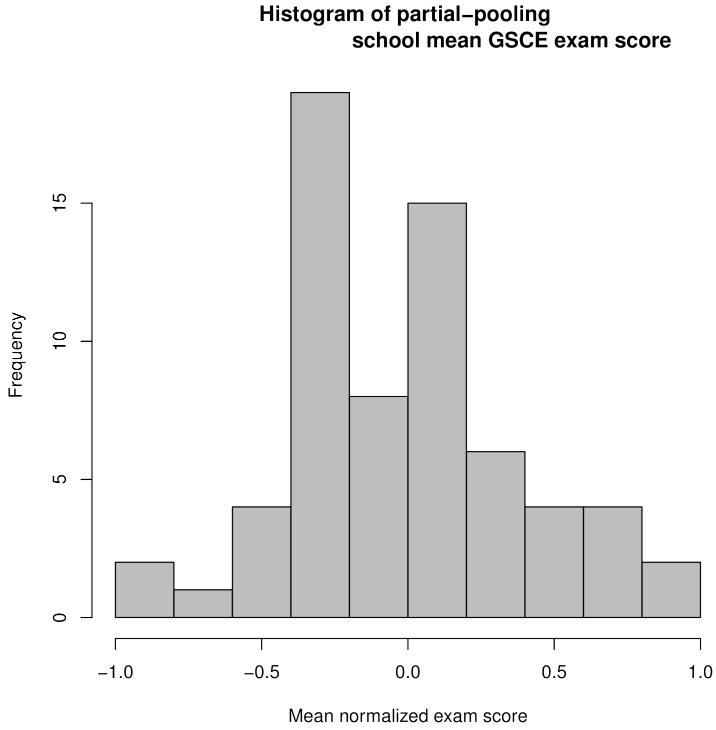

Figure 7

[1] Data set is finite so large $< \infty$

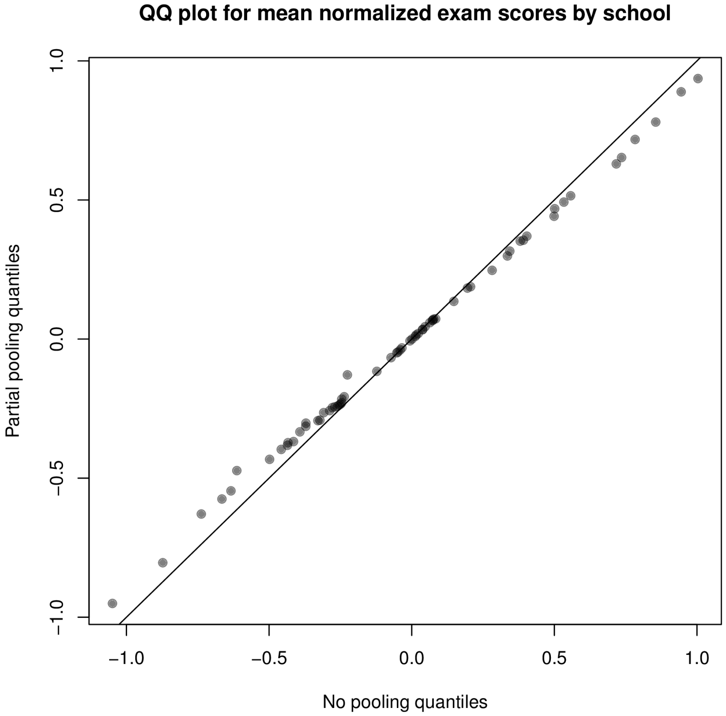

Figure 8

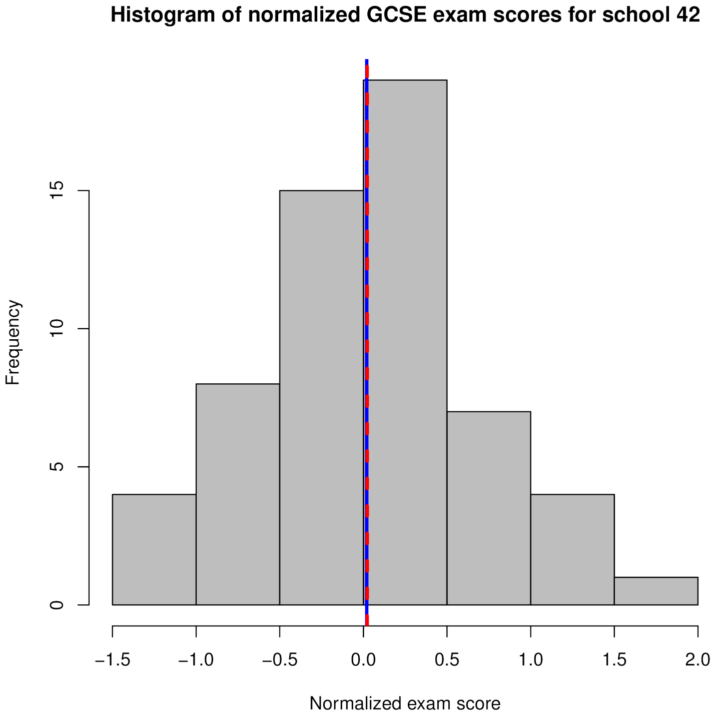

Figure 9: Partial pooling (blue) and no pooling (red) intercepts for school # 42

2.4 Example school

To visualize the difference between the no pooling and partial pooling mean, I pulled the data for school number 42 and created a histogram of the normalized GCSE test scores with vertical lines marking the no pooling and partial pooling intercepts (means). In this case, the fitted intercepts were almost the same, as shown in Figure 9 with the partial pooling slightly closer to 0 than the no pooling (as expected).

3

As indicated in the lecture notes, the high-level purpose of partial pooling, in this case, is to account for three things:

- The variance of exam scores within each school
- The variance of exam scores between schools
- The number of students in each school

Given these figures, a partially pooled school intercept will be closer to complete pooling when the number of students in the school is less than the ratio of the variance of exam scores within the school to the variance of the exam scores between schools. When the number of students in the school is greater than this ratio, the partially polled school intercept will be closer to the no pooling value. This relationship is intuitive because, all else the same:

- A school with a greater number of students gives us more confidence that the mean value for the school is representative (no pooling)
- A school with less variance in the exam scores gives us confidence in the same direction (no pooling)
- When the variation between schools is smaller, we tend to think that there is little connection between the mean score and the school, so we favor the pooled result

4 Adding standardized intake LRT score: “...LRT.R”

4.1 Complete pooling

To add the students standardized intake score on the London Reading Test at age 11 as a predictor to the complete pooling regression, I simply added the variable to my regular linear regression `lm()` model. This resulted in an intercept about 10 times the original intercept without the LRT score (-0.0012 vs -0.00011) but clearly the final value is still very close to 0.

4.2 No pooling

Using the same dummy variable approach, I added `standLRT` to the linear regression model. I then recreated the histogram of intercept values for comparison with the previous no pooling regression. The result in Figure 10 shows a significantly different intercept value distribution. In comparison to the regression without the LRT score, the distribution of intercept values is shifted right substantially.

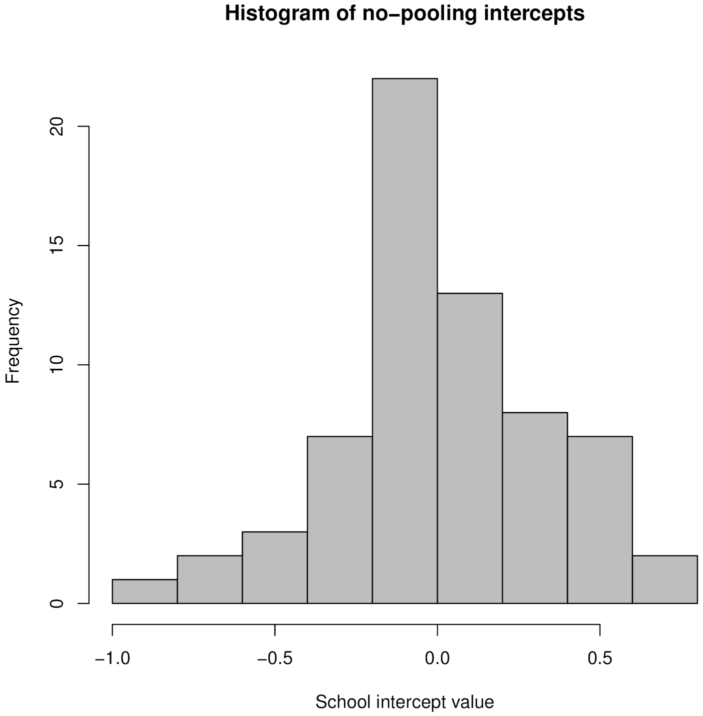

Figure 10

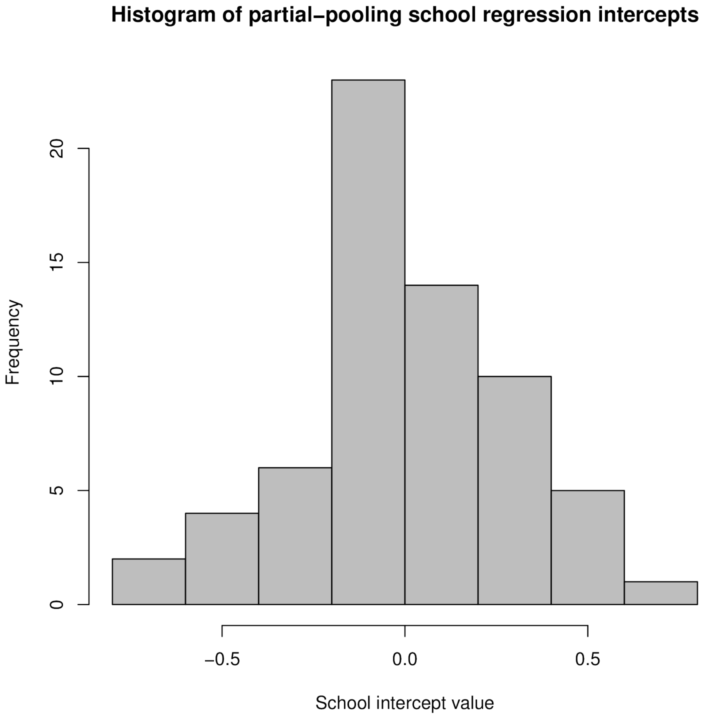

Figure 11

4.3 Partial pooling

For partial pooling, I again utilized the `lmer()` function and added the `standLRT` variable to the regression. To display the changes to the intercepts as this predictor is added, I again constructed a histogram of the intercept values which is shown in Figure 11. The results show a similar change to that of the new no pooling regression (rightward shift).

4.4 Regression comparison

To compare these new regressions, I looked at both the distribution of intercept values and the estimated coefficients for the LRT scores. To compare the distribution of intercept values, I again used a QQ plot, but this time with the complete pooling intercept value added as a single dashed black line. The QQ plot is shown in Figure 12 and indicates that, while the distributions have become more similar for a larger span around the median, the deviations have increased around the tails in the same general trends as before.

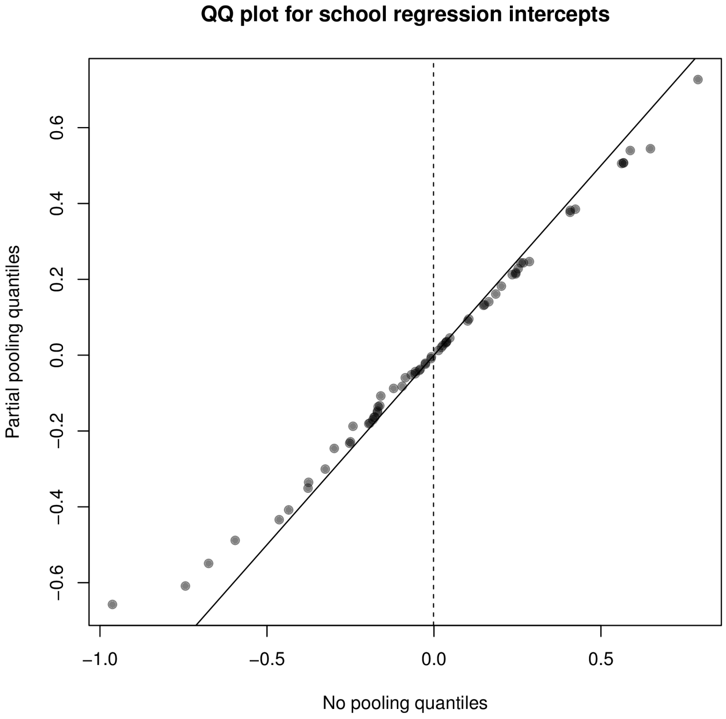

Figure 12

The LRT score coefficients for each regression are shown in Table 6. While the values are similar, the complete pooling coefficient is about 0.03 larger than the no pooling coefficient. As the lecture notes mention, this indicates that there is some correlation between the LRT intake scores and the school-level GCSE scores (intercepts). We should be concerned about this correlation because if it is not accounted for, we may conclude that GCSE scores are more dependent on the school than they actually are.

Table 6

Regression	**standLRT coefficient**
Complete pooling	0.595
Partial pooling	0.563
No pooling	0.559

5 Group-level predictor: "...LRT2.R"

To add the average London Reading Test intake score for the school as a group-level predictor in a multi-level model, I again used the `lmer` function and added the `schavg` variable as a regressor. This produced the distribution of intercept values shown in Figure 13.

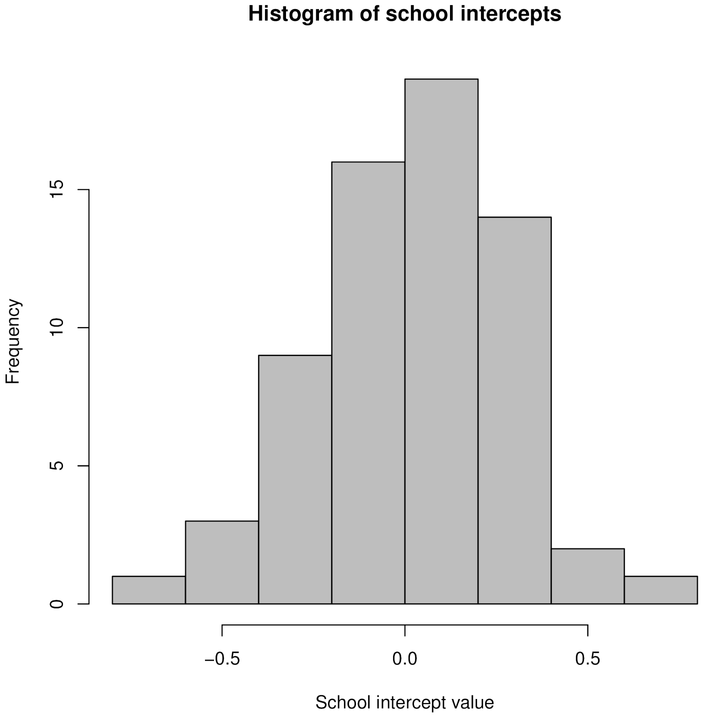

Figure 13

As shown above, the spread of the distribution in the multi-level model is smaller than in any of the previous pooling regressions. Overall, adding in the school average LRT scores reduced the variance of the school intercept values from 0.094 (partially pooled) to 0.079. At this point I realized that if I were to try and complete the rest of the questions on-time, I would be going through the paces without learning anything.[2]

11

Econometricians, statisticians, and machine learning researchers are hungry because when they cook together, the econometrician has to log every step, the ML researcher requires supervision, and the statistician can't

[2] I also re-read the syllabus!

make sense of what either of them are saying.

12

possible issues: - selection bias (students who did not have results for both intake measures were different than students who did) - students who did not take an exam got a score of 0. I wonder how this impacts the regression? 0 is lower than a score of 1, but these results can't be ordered as such?