STA 610 Final Analysis

John Gillen

2024-12-08

Contents

Introduction

The WASL dataset contains information from schools in Washington regarding their academic resources (how experienced their teachers are on average, student:teacher ratio, etc.) and student demographics (percent of white students, proportion of students eligible for free or reduced cost meals), as well as general information (county, school name, etc.), and data on the number of 5th or 10th graders who took and the number who passed the WASL in reading, math, science, and (for 10th graders only) writing during the 2005-2006 school year.

In this analysis, I aimed to build a sensible statistical model to explain the variation in pass rates of schools. In particular, I wanted to find out if staffing or resource-related information (student:teacher ratio, years of experience of teachers, percent of teachers with Master's degree or higher) truly moved the needle when it came to schools' performances on the WASL, compared to demographic information. In short, I found demographic information to be an unequivocally stronger predictor of performance, and further found that experience and education of teachers in particular was of negligible importance (which was somewhat disheartening as someone with an interest in teaching). I was also curious to see if there were any specific characteristics of schools whose pass rates were not well-fit by my model, but none of the explanatory variables were helpful in this regard - the best indicator of a likely outlier was extremely high or extremely low (close to 100% or close to 0%) pass rate.

Exploratory Analysis/Cleaning

There is a hierarchical structure to the data: We have grades nested in schools nested in counties, and then subjects that represent repeated measures across grade levels (except for writing) within schools within counties. There are 39 counties represented in the data and 1481 schools. Of the three subjects taken by both grade levels, on average about 94 students per school took the tests. This is a rather right-tailed distribution, as the median is 67, with the maximum values being close to 800 for each subject. The number of schools at which 10th graders were tested is 416 compared to 1065 for 5th graders. Mean enrollment for the schools is about 522, with a median of 439.

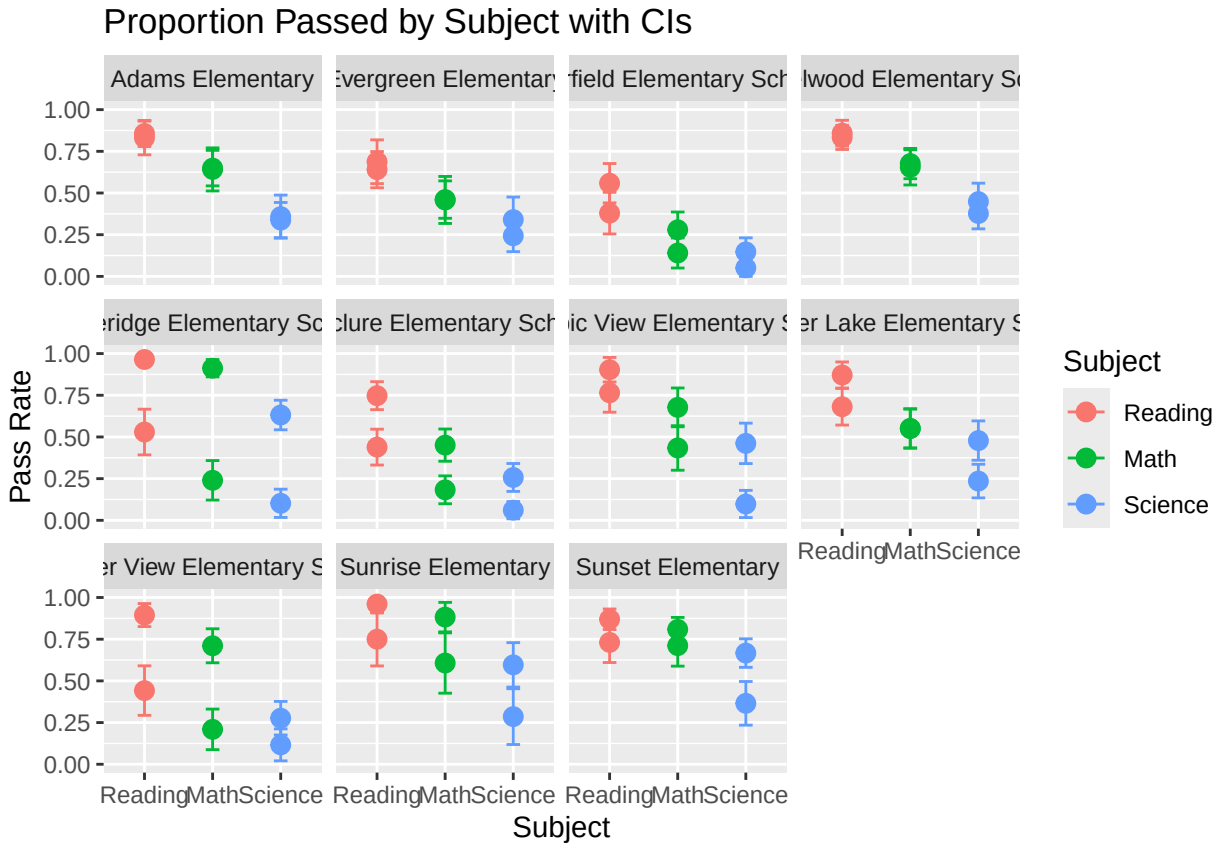
Cleaning

There are 66 rows with missing information in the demographic or resource-related variables. I choose to remove these rows. Rows with missing test-related information can possibly be handled with a different approach. Obviously the 5th grade rows will all be missing information about writing scores, so we can ignore these.

After removing the 66 rows with missing 'school' information, there are 15 rows with missing 'test' information (excluding those who have missing writing information due to being 5th graders). More specifically, these schools reported the number of students who took each (applicable) test, but didn't report the number who passed (e.g. John Marshall High School reported that 7 students took the math test and 8 took the reading and science tests, but do not report how many passed for either of these).

A reasonable assumption may be that schools who didn't report the number of students who passed did so because no students passed. Looking at the 15 rows with missing results, only one of them reported 0 students passing for one subject while not reporting for another subject. In this case, we should consider it unlikely that they would have not reported the other zero, so we eliminate this row (though ironically it looks probable that it should have been a zero). For the rest, we can use our knowledge of the data to try and guess which unreported scores were highly likely to have been zeroes. We know that reading and writing scores are the highest with about 80% passing on average, with math being noticeably lower, and science the lowest. So we will delete the schools who didn't report data for reading or writing (for 10th grade), as it is unlikely that these would be zeroes given that they had nonzero pass rates for math and/or science. Of the 9 rows remaining, there are two who didn't report math scores and none who reported science scores. If we disregard the rows where both aren't reported, the remaining 7 schools all have very low math pass rates (max of 5/15, next best is 2/10), and all had fewer students take the science test than the math test. I believe it is reasonable to assume that these schools all had 0 students pass the science test. We drop the other 8 rows.

Something strange (that I definitely noticed before fitting several models) is going on here: some schools seem to have submitted two sets of score data for the same grade level. It is unclear whether these come from different years, since all other information is the same. Let's look at the outcomes and see if that can give any insight on how to proceed.



This is quite troubling. For some schools, the outcomes are close enough that we can reasonably conclude that the duplicates are a result of separate test dates or different years. However, for some others, the subject-wise scores are drastically different. There are a few ways we can proceed here:

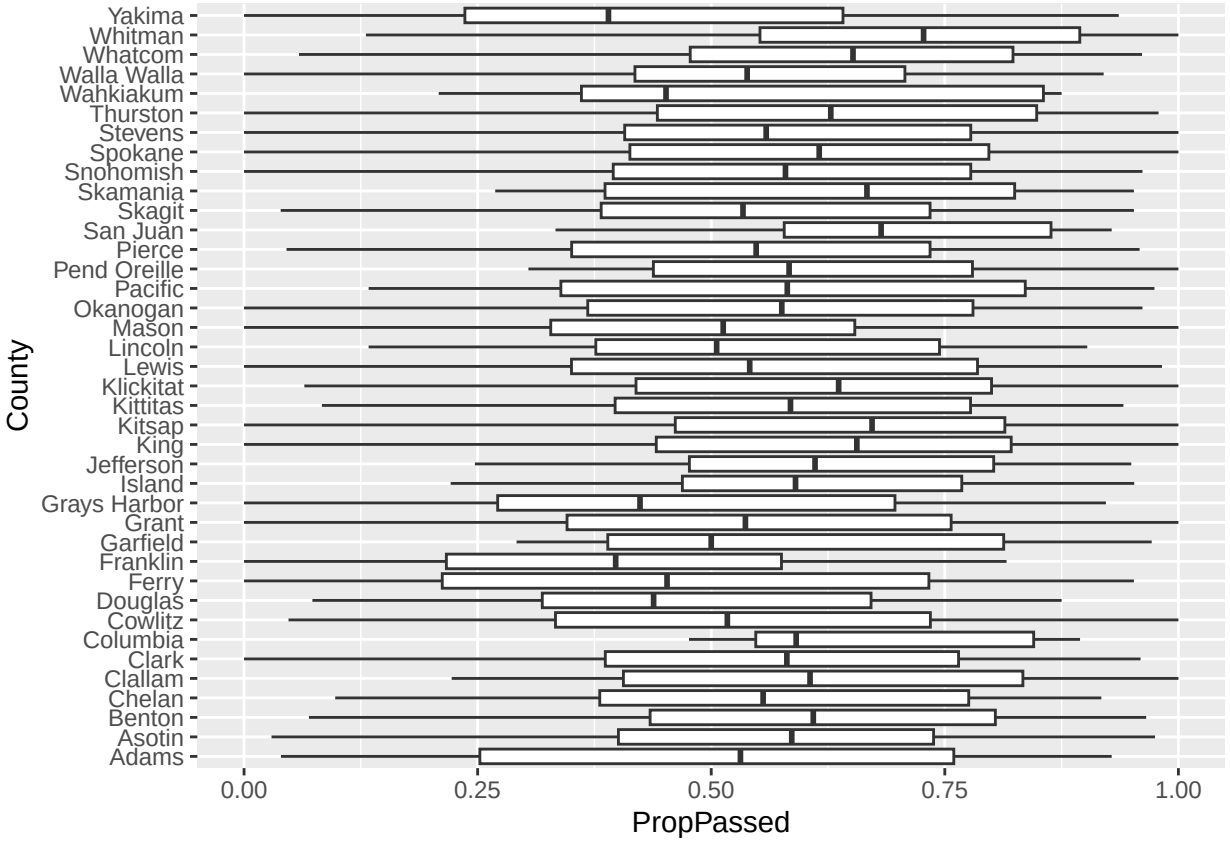
1. Do nothing at all - This can be problematic for the schools with extreme differences when we are fitting models.
2. Combine the duplicate data - This seems reasonable, given that the duplicate scores are reported by the same school. We shouldn't expect year-to-year results to vary too much, so the schools with big differences may have been grouping their reported scores by some other factor, which means combining the duplicates should give us a better idea of their true performance (but again this may not be wise given that we don't know what factor the students are separated by)
3. Drop one or both rows for the duplicates - Luckily, as we can see below, the counties with duplicated data are quite large, and there would only be 11 schools to remove. Dropping one observation is problematic for the duplicates with very different outcomes, so the only viable option here would really be to drop both.

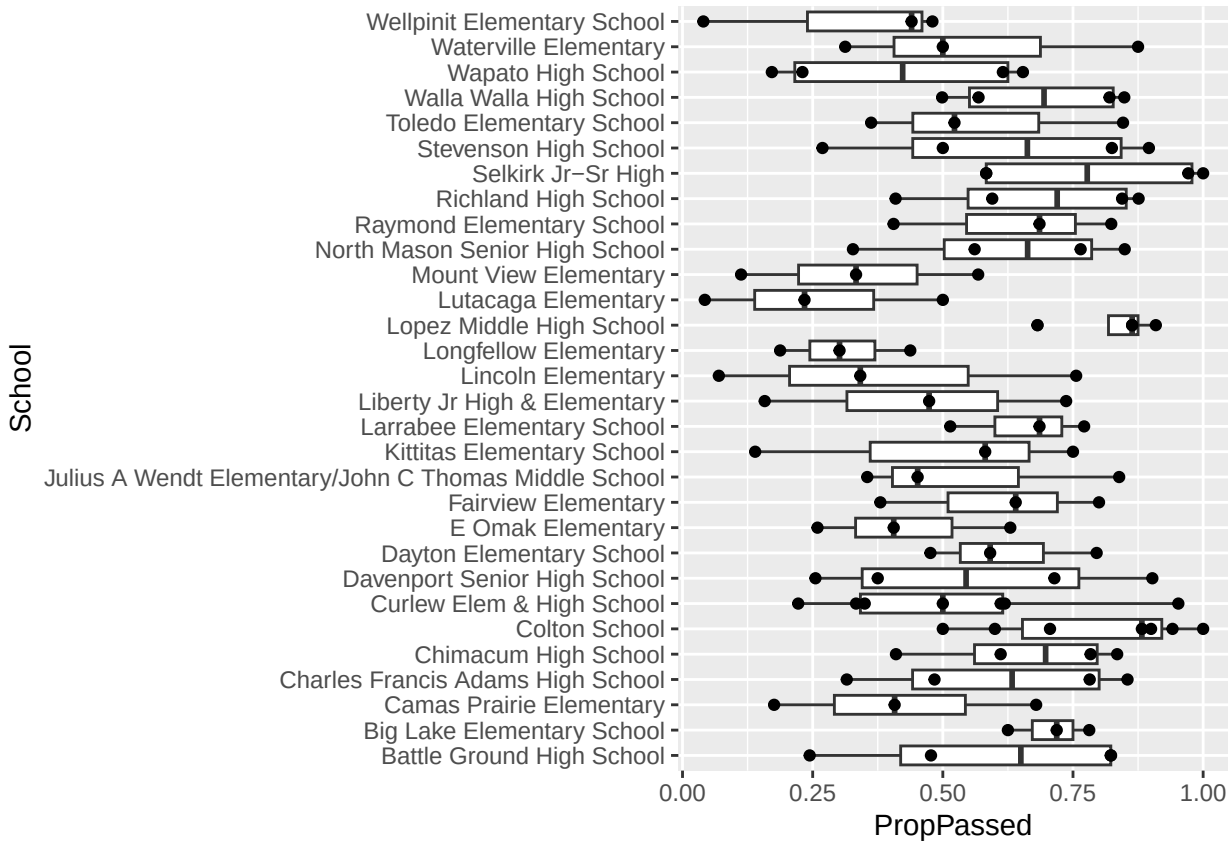
```
## # A tibble: 4 x 2
##   County  Schools
##   <chr>     <int>
## 1 King        326
## 2 Pierce      155
## 3 Spokane     107
## 4 Yakima       56
```

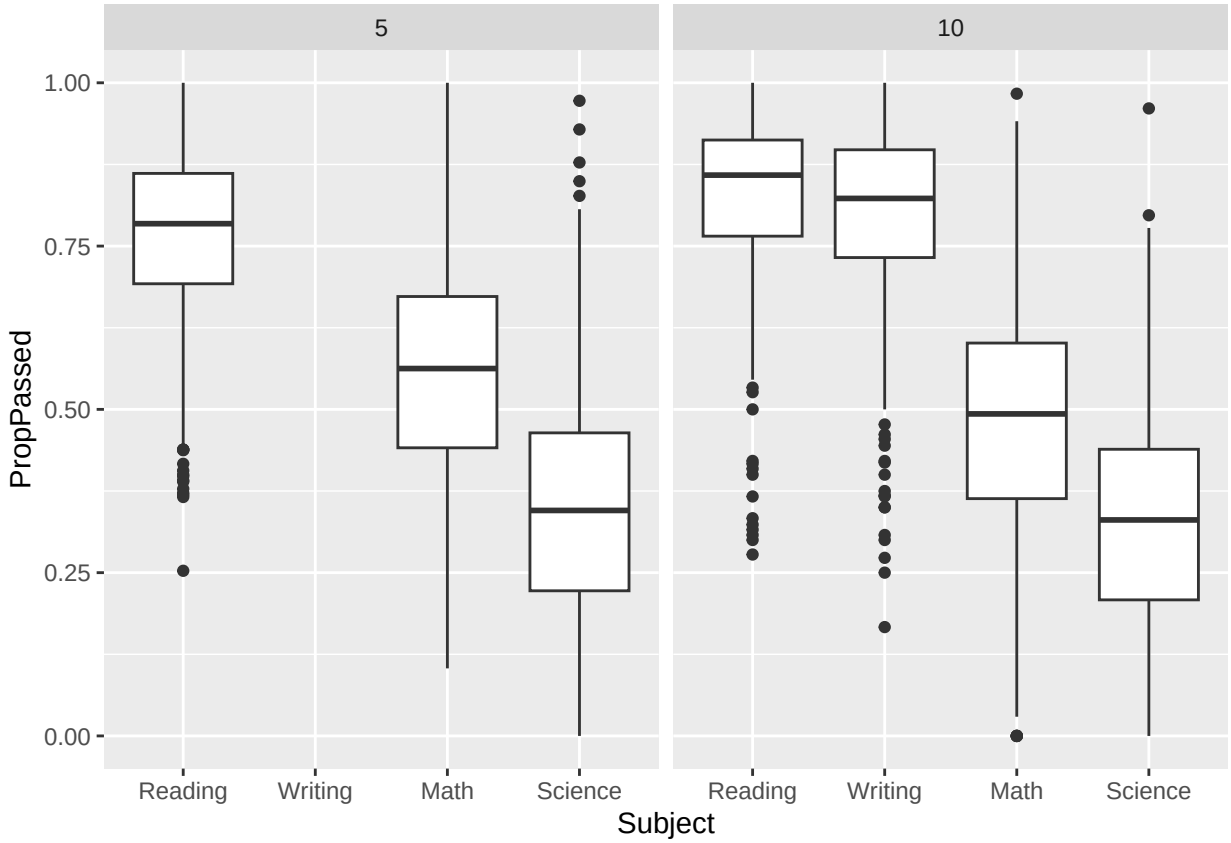
I will choose to remove all schools that were duplicated. It was difficult to decide between this and combining, but I think this is the safer option given that we don't know why the duplicates were separated.

Exploratory Visualization

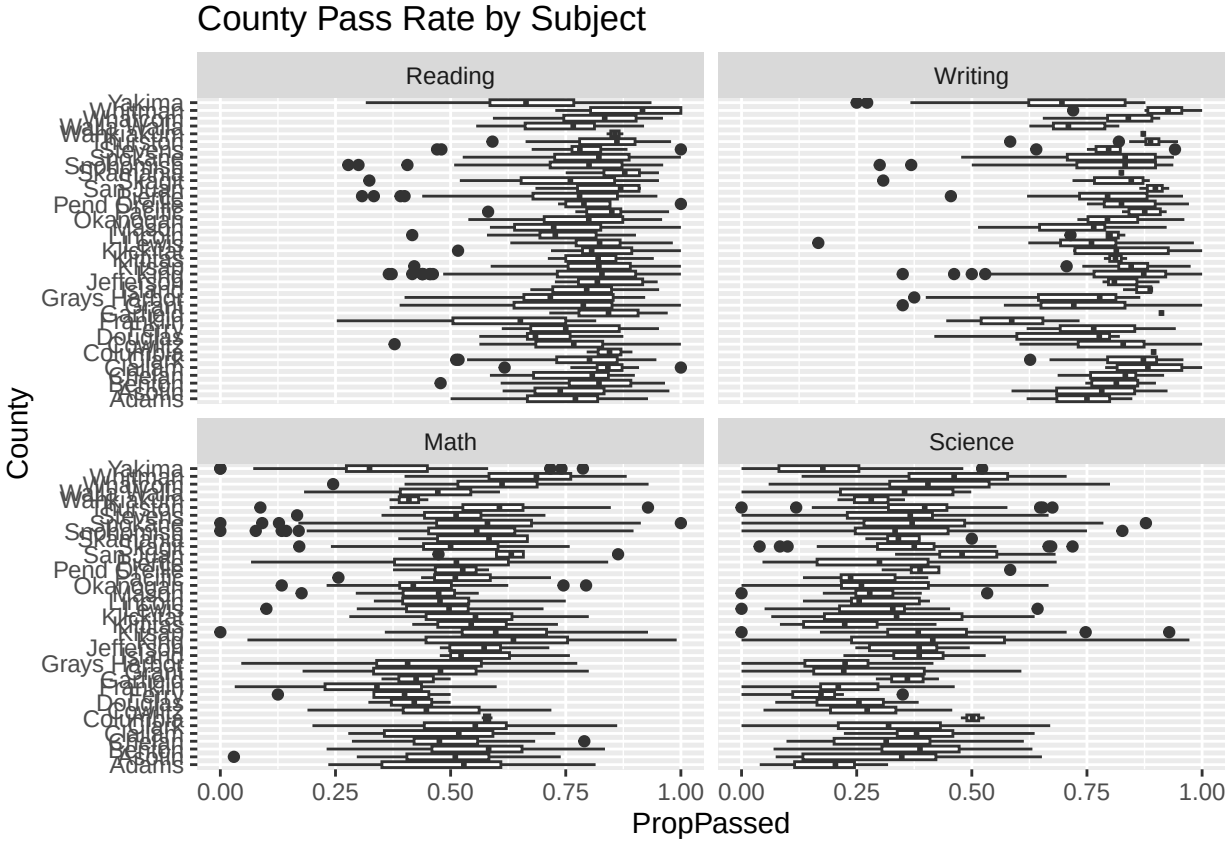
Now for some exploratory visualization. Consider the structure of the data: we have counties that contain schools that contain grade(s) being tested on subjects. We first want to just get a look at each of these variables in isolation (note that we will always have to consider subject paired with grade in some capacity due to the absence of a writing test for 5th graders)

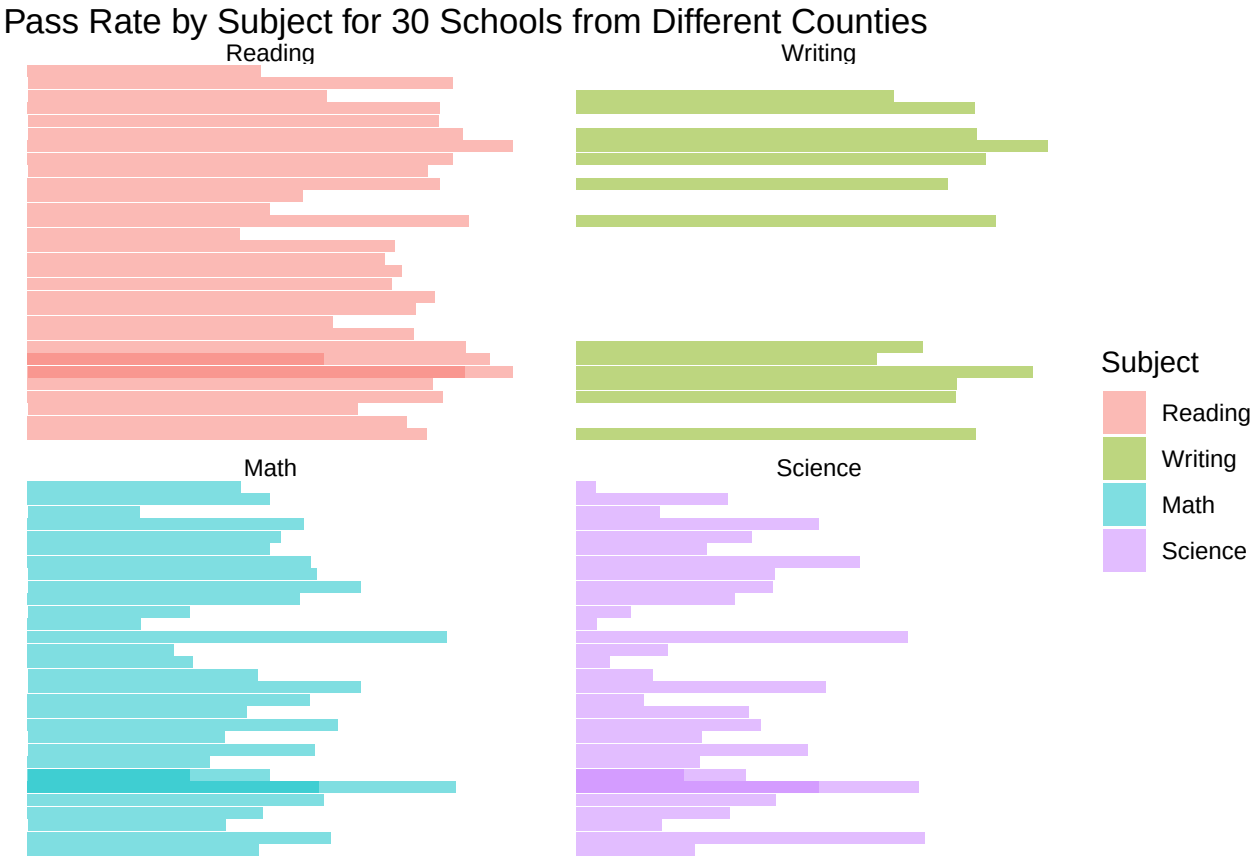


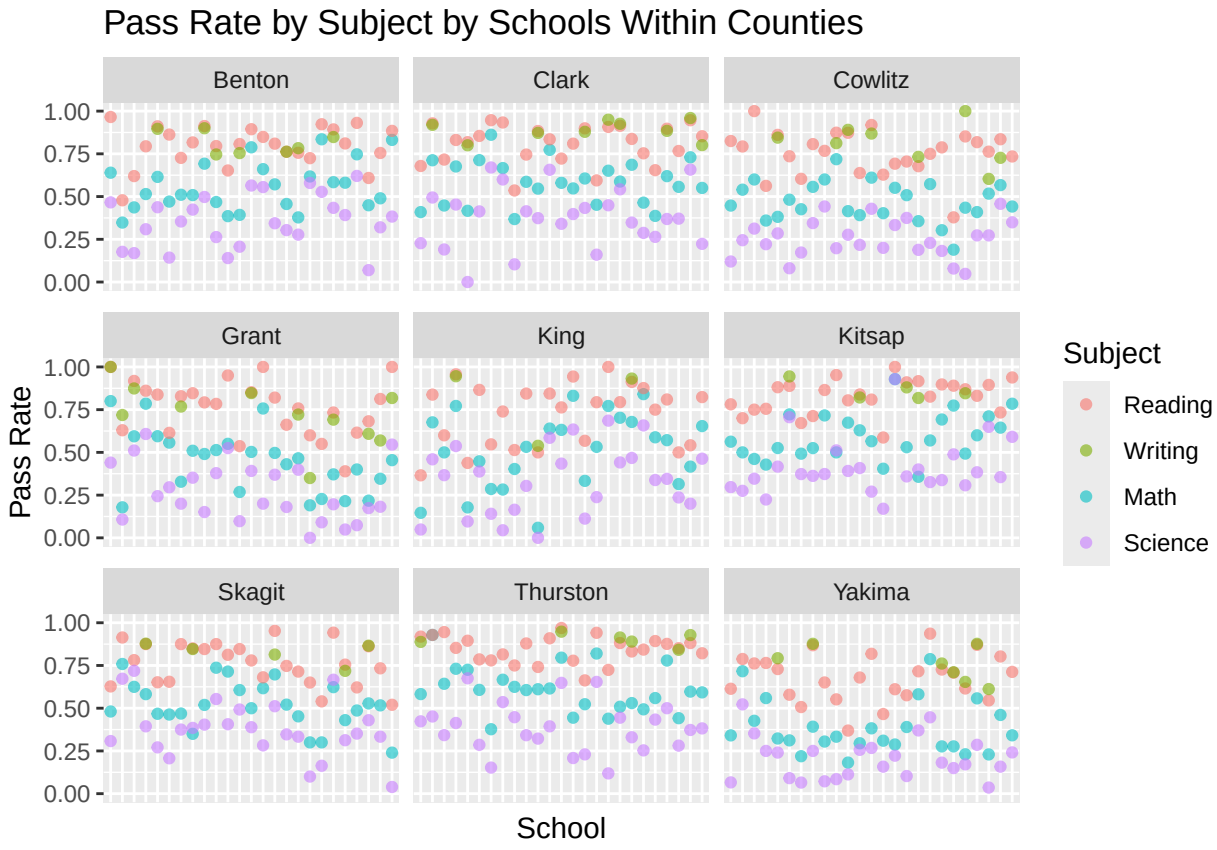

School
Wellpinit Elementary School
Waterville Elementary
Wapato High School
Walla Walla High School
Toledo Elementary School
Stevenson High School
Selkirk Jr−Sr High
Richland High School
Raymond Elementary School
North Mason Senior High School
Mount View Elementary
Lutacaga Elementary
Lopez Middle High School
Longfellow Elementary
Lincoln Elementary
Liberty Jr High & Elementary
Larrabee Elementary School
Kittitas Elementary School
Julius A Wendt Elementary/John C Thomas Middle School
Fairview Elementary
E Omak Elementary
Dayton Elementary School
Davenport Senior High School
Curlew Elem & High School
Colton School
Chimacum High School
Charles Francis Adams High School
Camas Prairie Elementary
Big Lake Elementary School
Battle Ground High School
0.00
0.25
0.50
0.75
1.00
PropPassed

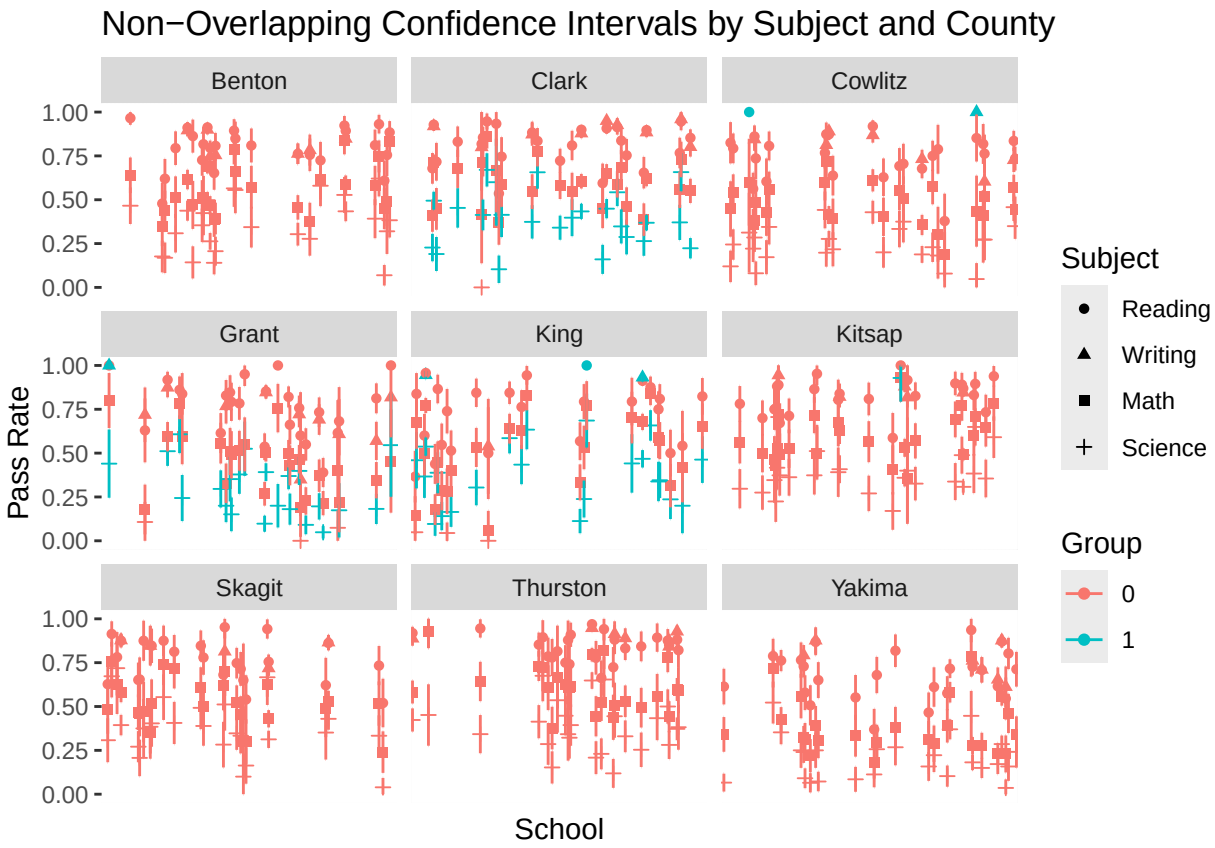

5
10
PropPassed
1.00
0.75
0.50
0.25
0.00
Reading
Writing
Math
Science
Subject

We observe not a ton of apparent between-county variance, but quite a lot between subjects overall and schools across counties. There is also lots of within-county variance and moderate between grade variance by subject. Note that not many schools test both grades, so almost all within-school variance should be explained by subject. We now check for variance in subjects at the county and school levels, and schools within counties.




Pass Rate by Subject for 30 Schools from Different Counties
Reading
Writing
Math
Science
Subject
Reading
Writing
Math
Science

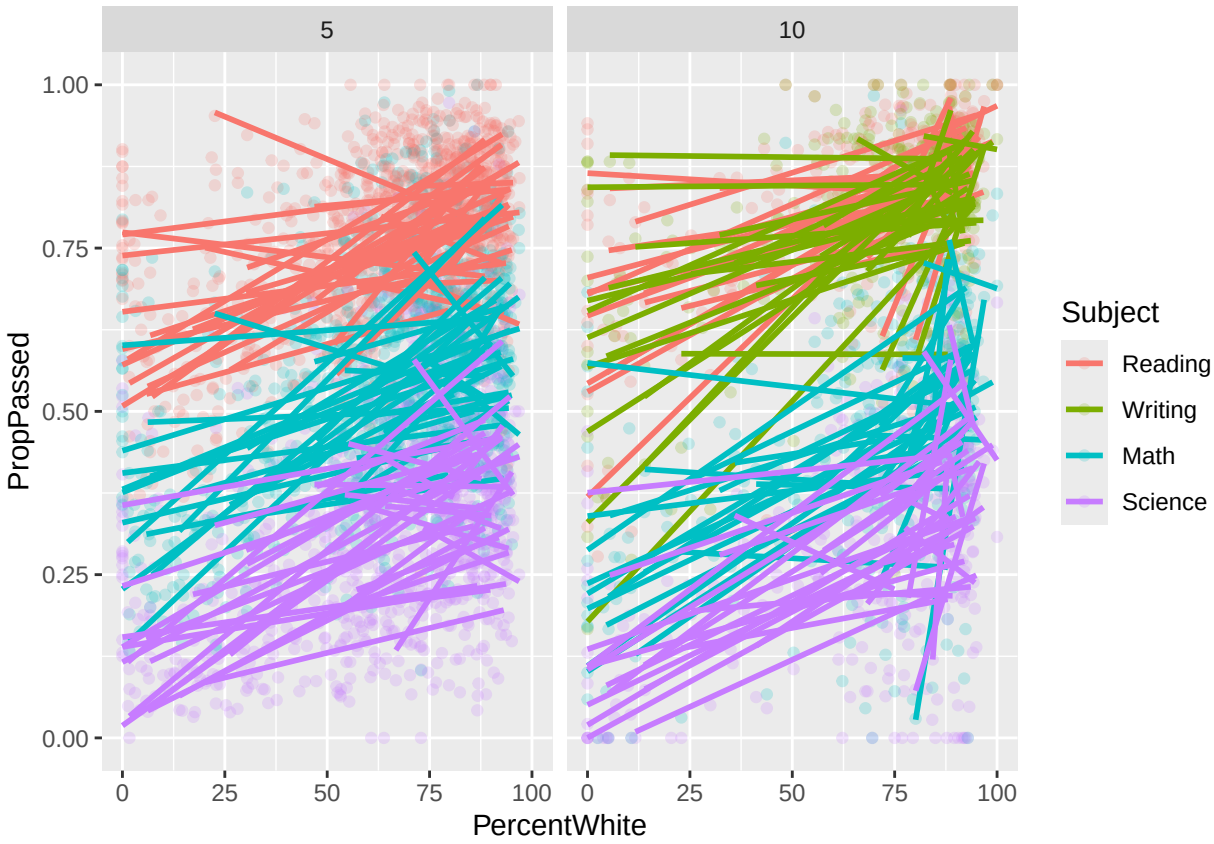

Pass Rate by Subject by Schools Within Counties
Benton
Clark
Cowlitz
Grant
King
Kitsap
Skagit
Thurston
Yakima
1.00
0.75
0.50
0.25
0.00
Pass Rate
School
Subject
Reading
Writing
Math
Science



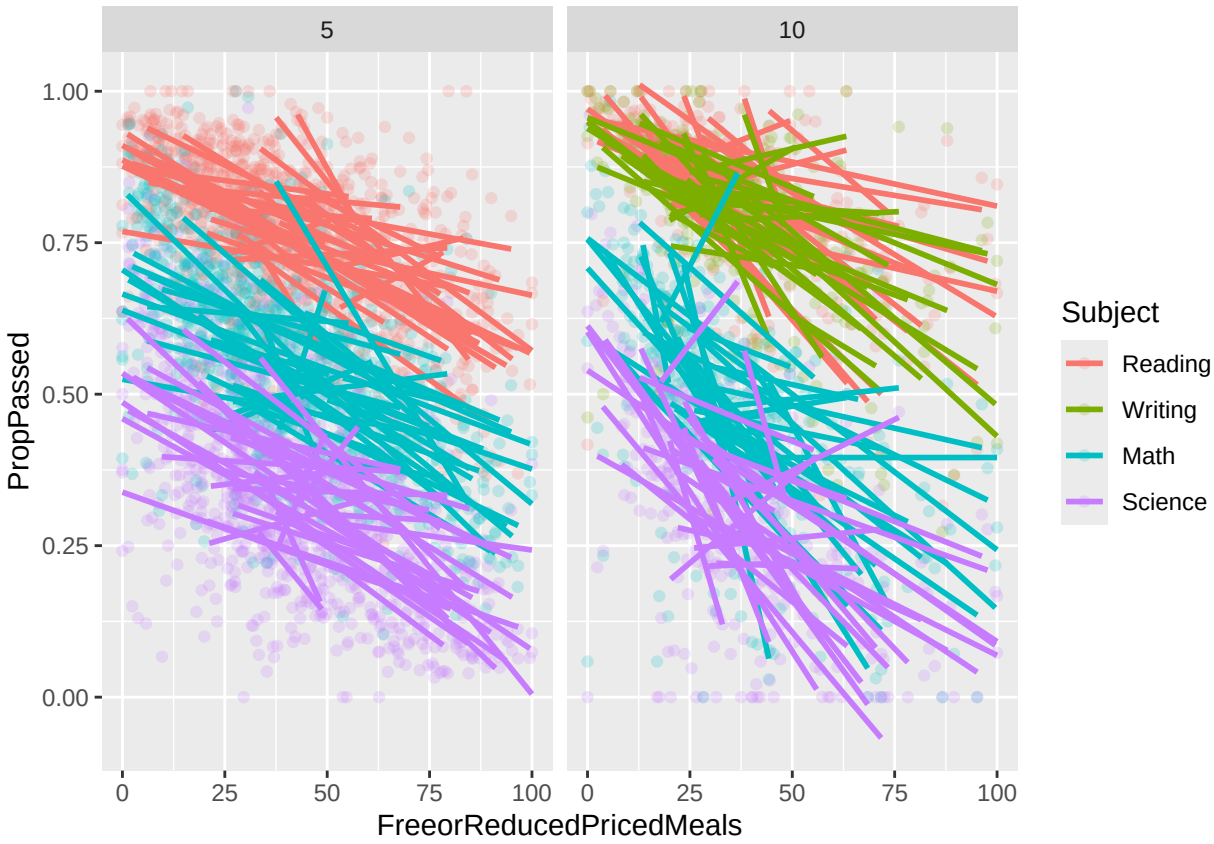
There appears to be some effect within subjects across counties and schools (but this may be explaiend by between-county differences). There also seems to be variability between schools within counties, based on a sample of schools from some of the larger counties.

Though I want to account for the above factors before getting into the demographic and resource variables, let's just have a glance at them for now:

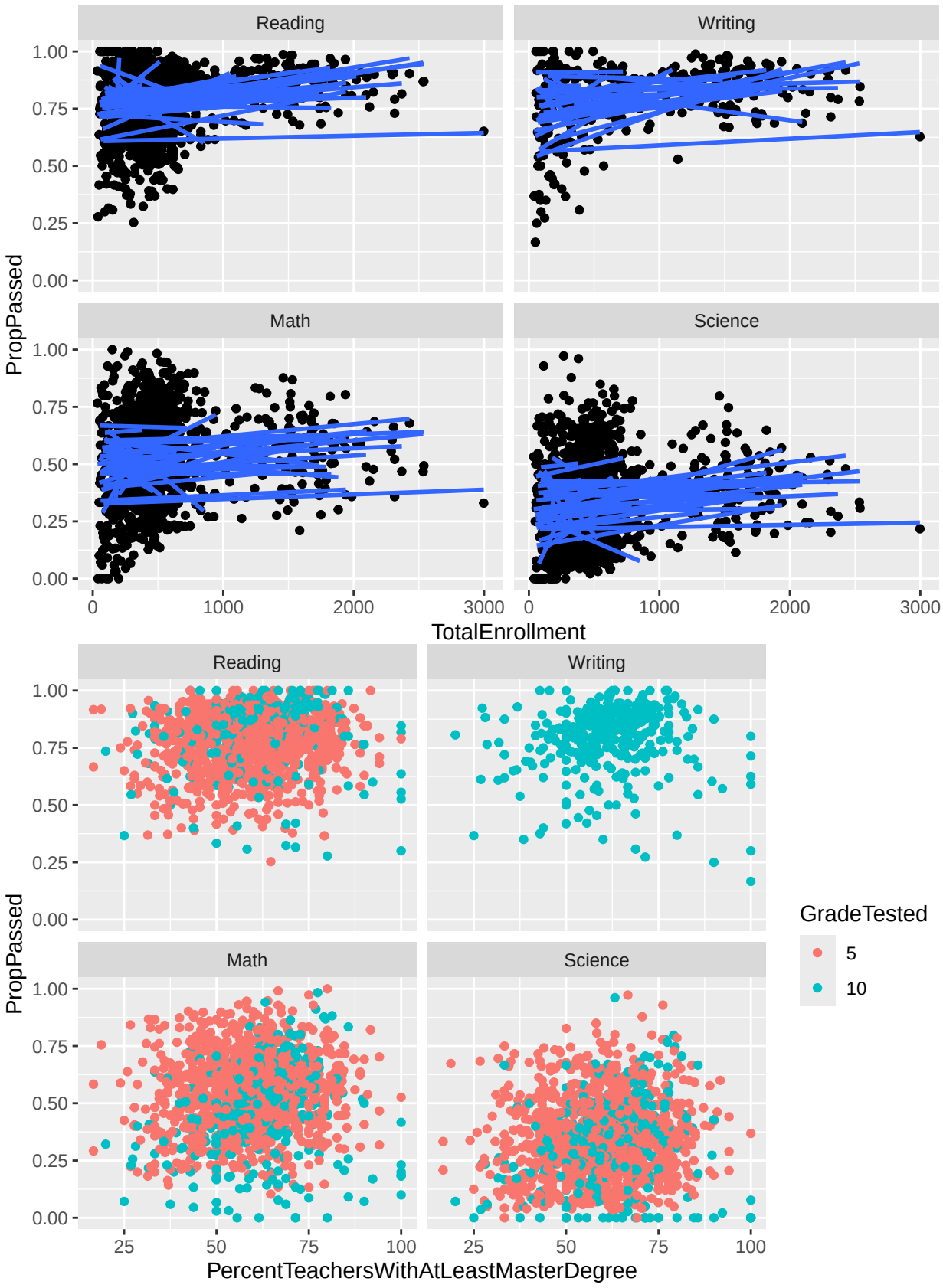
```
## `geom_smooth()` using formula = 'y ~ x'
```

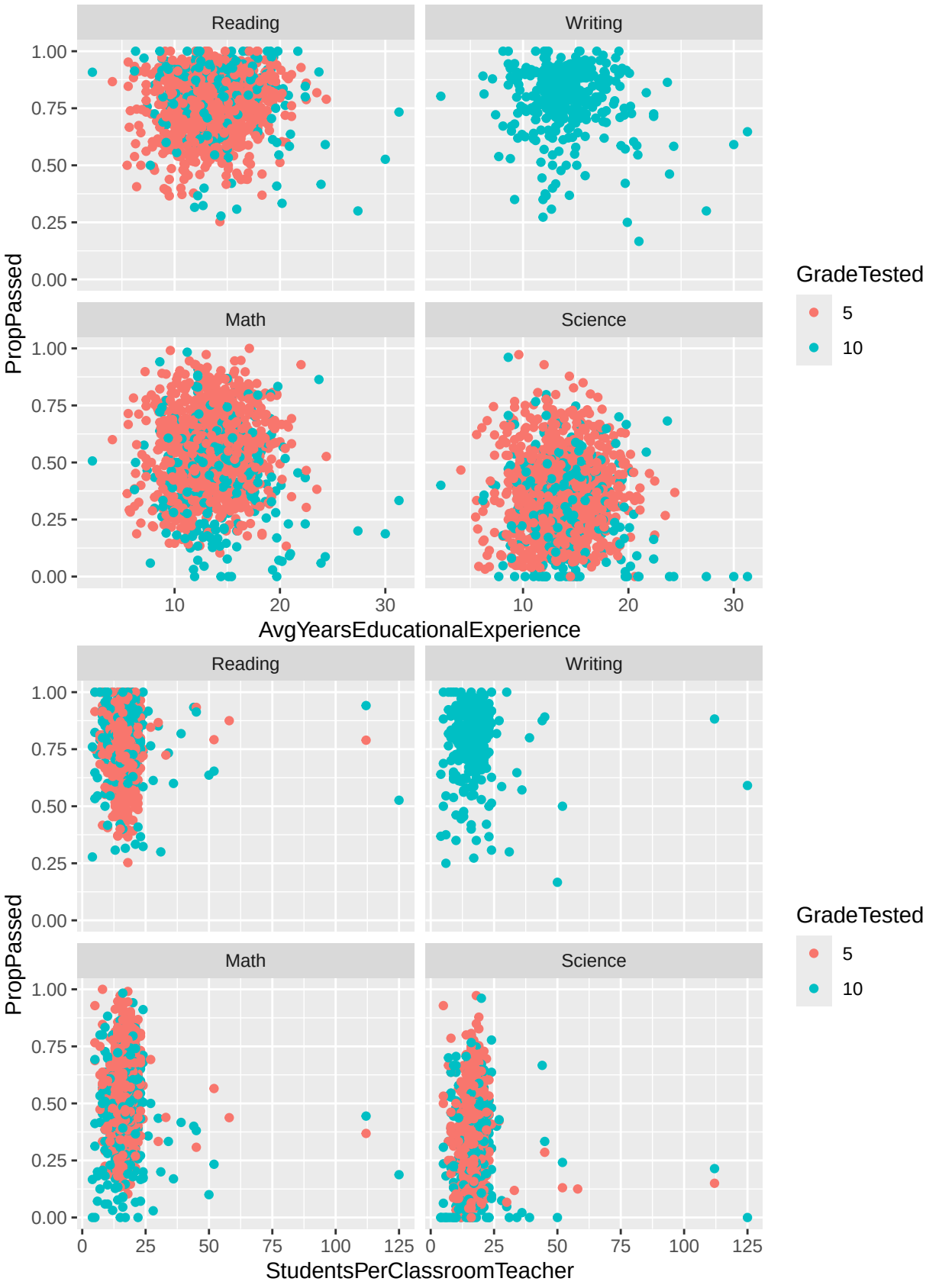



```
## `geom_smooth()` using formula = 'y ~ x'
```



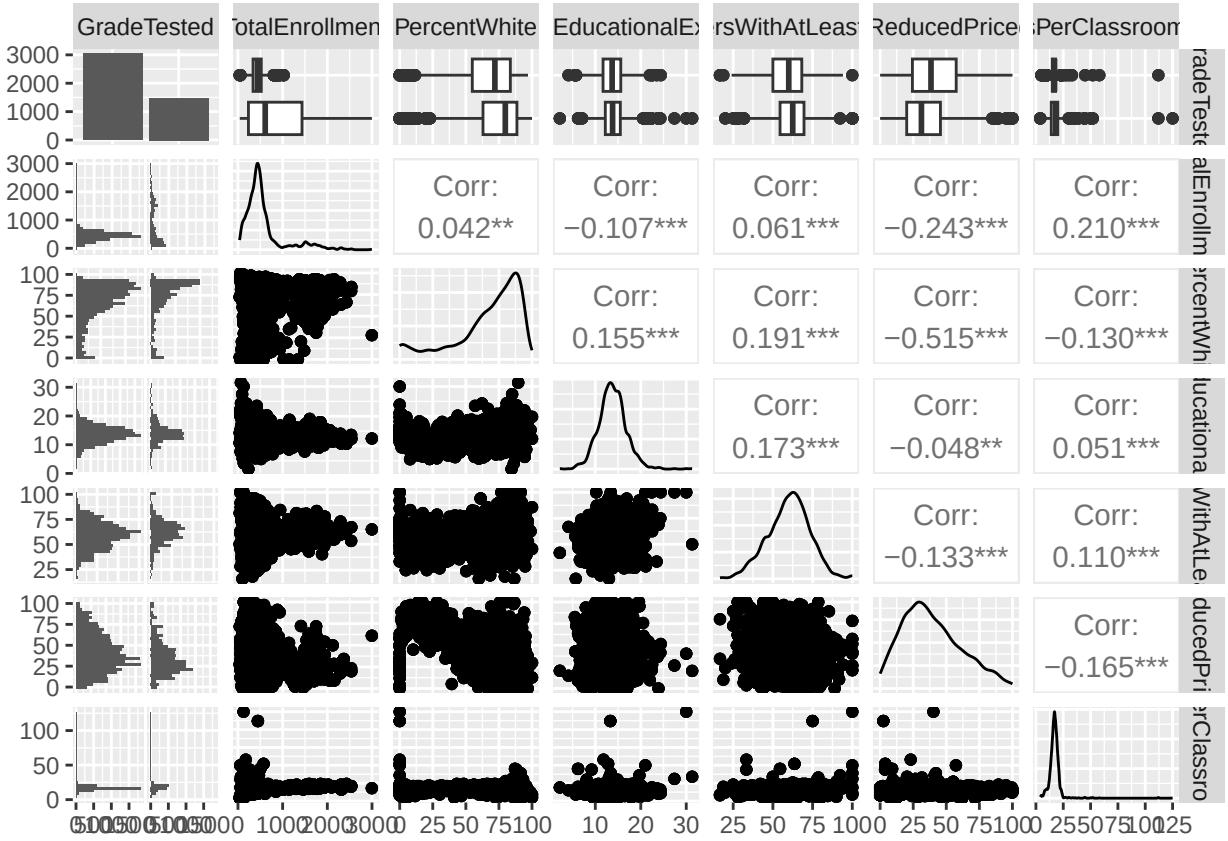

```
## `geom_smooth()` using formula = 'y ~ x'
```


Reading
Writing
Math
Science
PropPassed
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
TotalEnrollment
25
50
75
100
PercentTeachersWithAtLeastMasterDegree
GradeTested
5
10



%White has some correlation with improvement in pass rates, %reduced cost meals has very strong negative

correlation with pass rates for all grade levels and subjects, and most counties. Higher enrollment sees improvement in reading and writing and stability in math and science, and the other three variables don't show any obvious patterns. Let's see how they interact with each other:



There is a somewhat strong negative correlation between %white and %reduced price meals. Other than that, there are some moderate correlations between other variables, but nothing noteworthy.

Model Fitting Process

With this information, we are ready to fit some models. Given the nature of our response, we want to fit a binomial model using `glmer`. The only assumption we can check pre-fitting is that the data includes trials and successes/failures, which we already know it does. Also, it should be noted now that we will be evaluating models by BIC, and using DHARMa for diagnostics. DHARMa calculates standardized residual that make interpretation easier for complicated model structures, and performs several tests to make understanding residual patterns and validating assumptions convenient. It's also much easier to say than sbgcop.

(more info here: https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html)

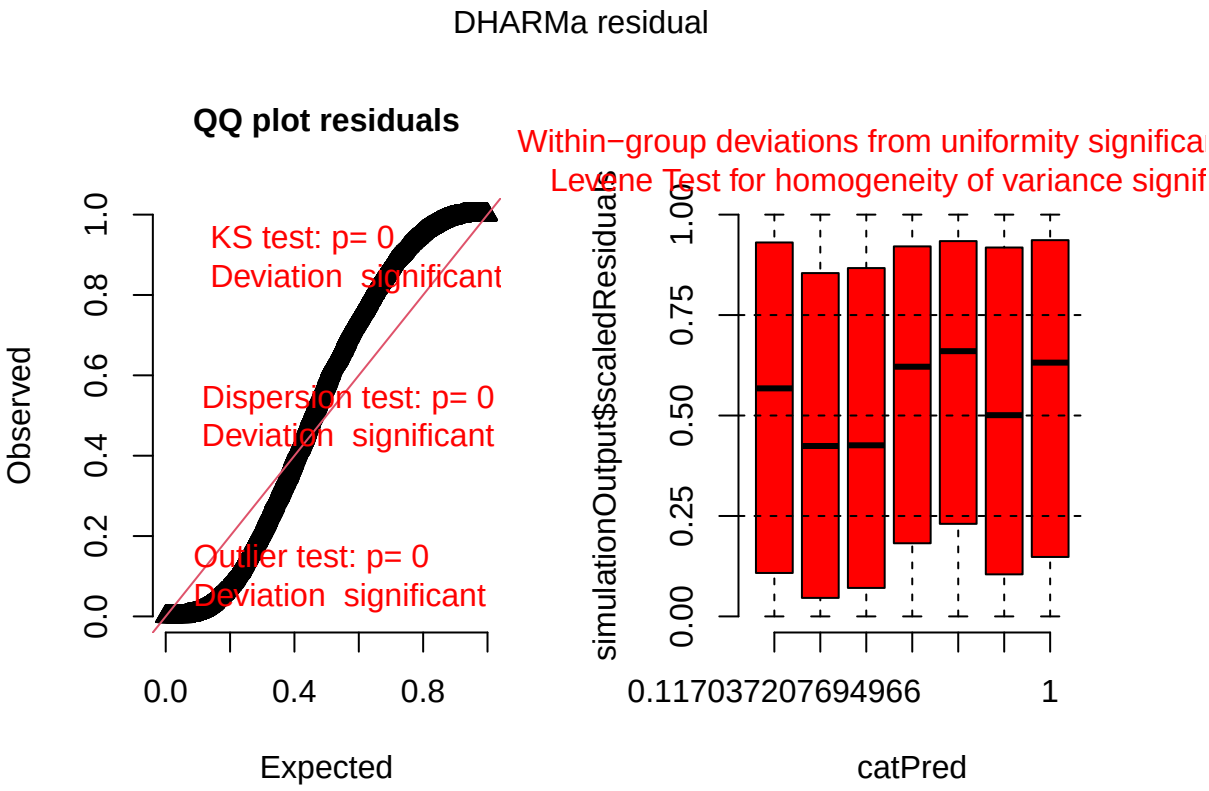
For starters, we will consider only a fixed effect for subject*grade (this will be encoded as a new variable to deal with the redundancy induced by grade and writing being perfectly correlated) and a random intercept for county. We will compare this against a model with only the county intercept, and then try adding a random intercept for school and a random intercept for school within county. We can try Subject_Grade as a random intercept as well to see if there is any difference.

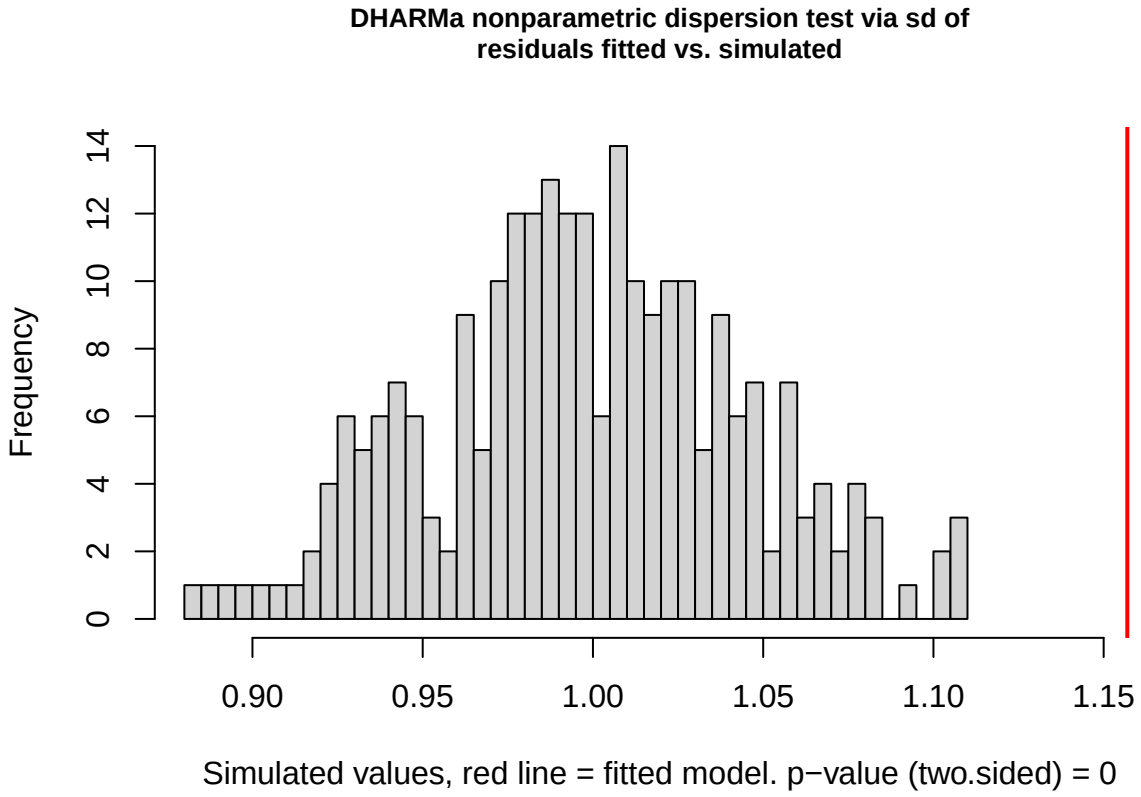
```
## Data: wasl_num
## Models:
## modcounty: cbind(MetStandard, TotalTested - MetStandard) ~ (1 | County)
## modsub_int: cbind(MetStandard, TotalTested - MetStandard) ~ (1 | Subject_Grade)
```

```
## modsub: cbind(MetStandard, TotalTested - MetStandard) ~ (1 | Subject_Grade) + (1 | County)
## modbase: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + (1 | County)
##            npar    AIC    BIC logLik deviance     Chisq Df Pr(>Chisq)
## modcounty     2 134551 134564 -67273   134547
## modsub_int    2  63179  63191 -31587    63175 71372.323  0
## modsub        3  55800  55819 -27897    55794  7380.924  1  < 2.2e-16 ***
## modbase       8  55743  55794 -27863    55727    66.597  5  5.225e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

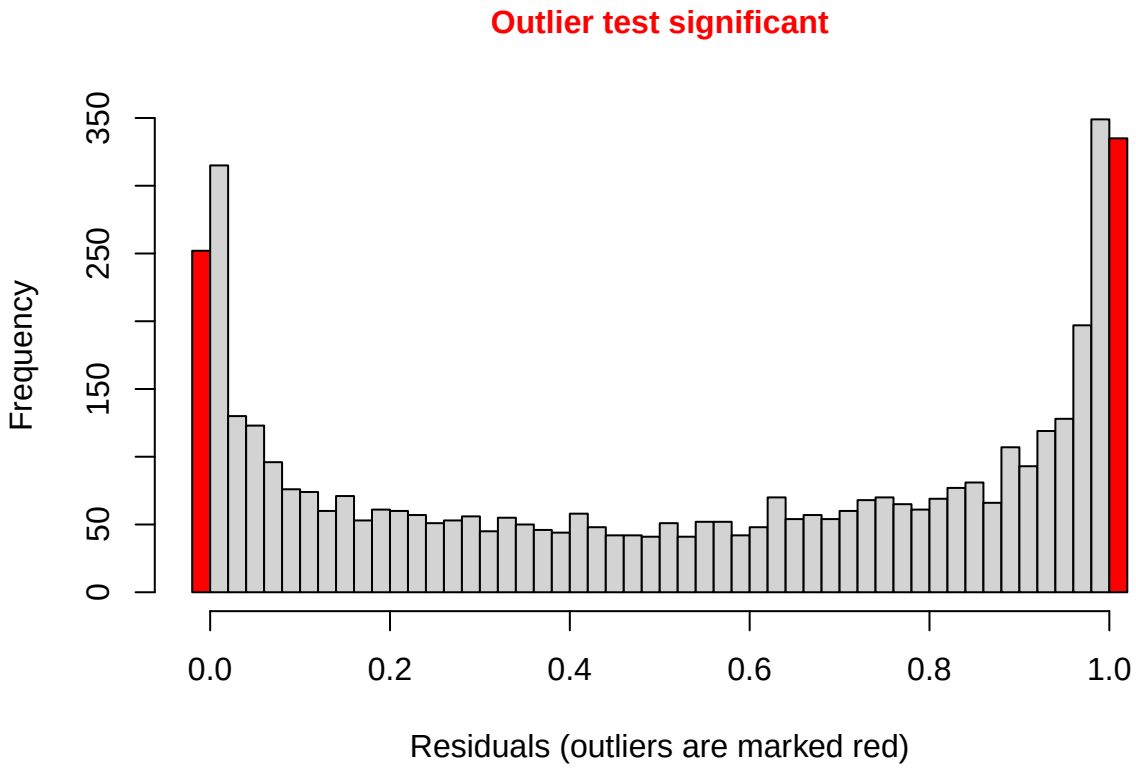
Including the fixed effect for subject*grade more than halves the BIC, and outperforms models treating the interaction as a random intercept, so we proceed with this model. Let's invoke DHARMa:

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```






```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.1596, p-value < 2.2e-16
## alternative hypothesis: two.sided
```



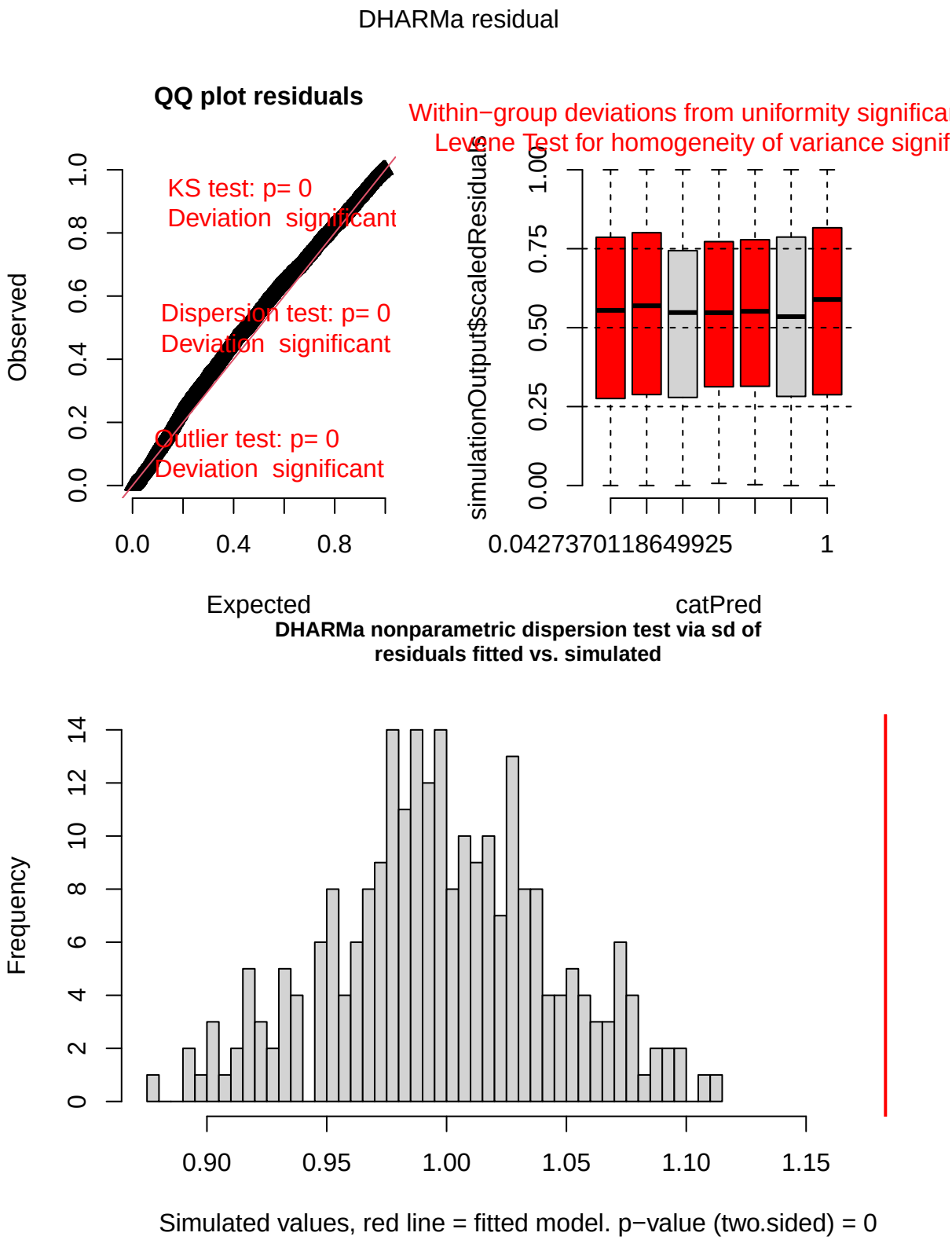

```
##
##  DHARMa bootstrapped outlier test
##
## data:  sim
## outliers at both margin(s) = 587, observations = 4525, p-value <
## 2.2e-16
## alternative hypothesis: two.sided
##  percent confidence interval:
##  0.00198895 0.03271823
## sample estimates:
## outlier frequency (expected: 0.00977900552486188 )
##                                         0.1297238
```

Clearly this model isn't an excellent one for our data. Let's try adding school or County:School-level intercepts:

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0138843 (tol = 0.002, component 1)
```

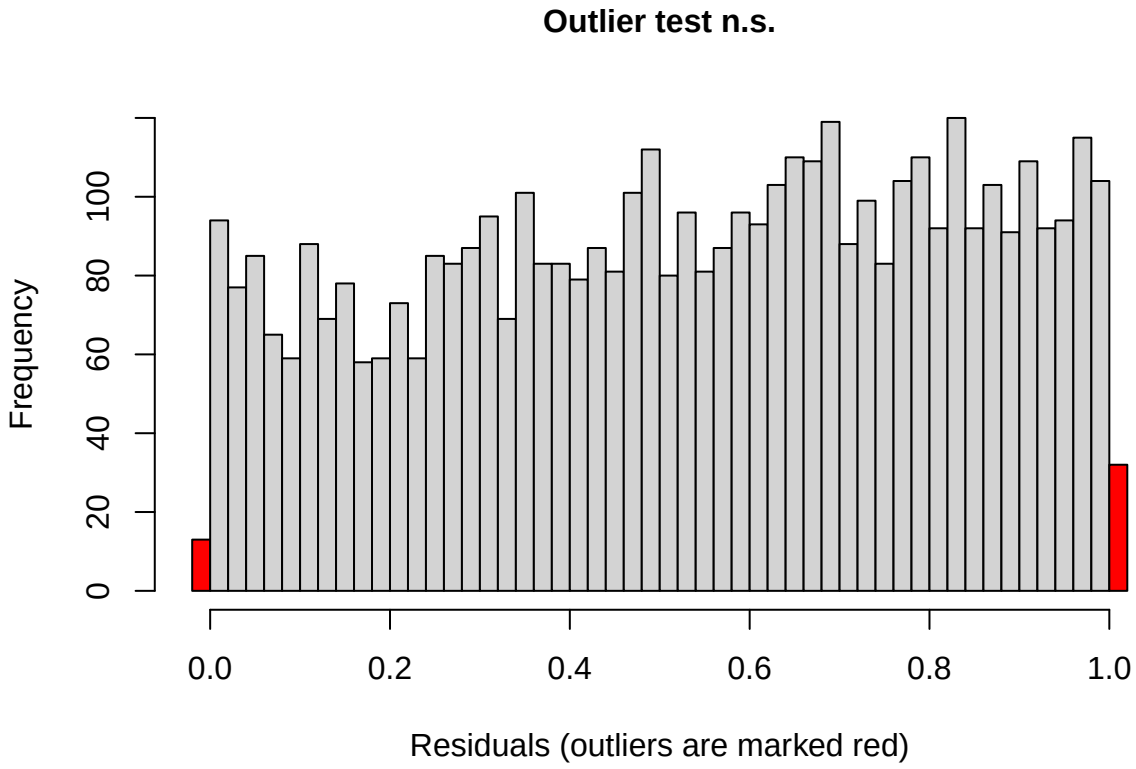
```
## Data: wasl_num
## Models:
## modbase: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + (1 | County)
## modsch: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + (1 | School)
## modboth: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + (1 | County:School)
## modschcounty: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + (1 | School) + (1 | Cou
## modboth2: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + (1 | County/School)
##              npar   AIC   BIC logLik deviance   Chisq Df Pr(>Chisq)
## modbase         8 55743 55794 -27864    55727
## modsch          8 32148 32199 -16066    32132 23595.5  0
## modboth         8 30378 30429 -15181    30362  1769.8  0
## modschcounty    9 31635 31693 -15809    31617     0.0  1          1
## modboth2        9 30249 30307 -15116    30231  1386.1  0
```

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```



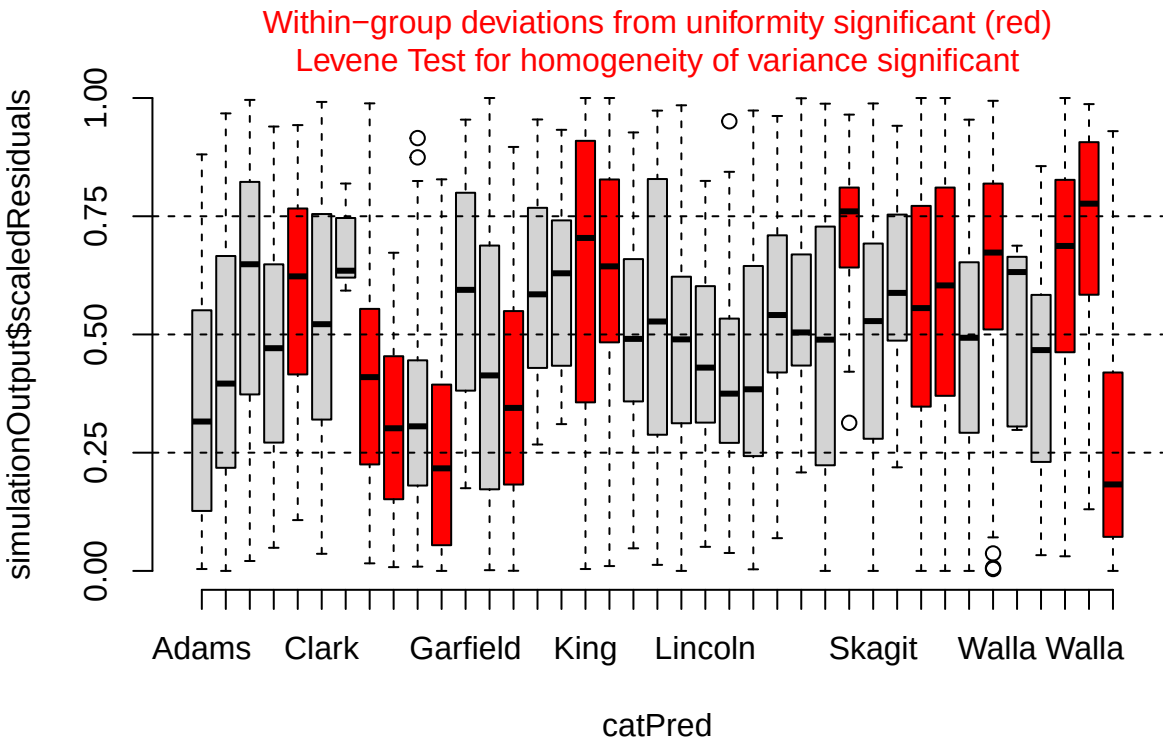

```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
```

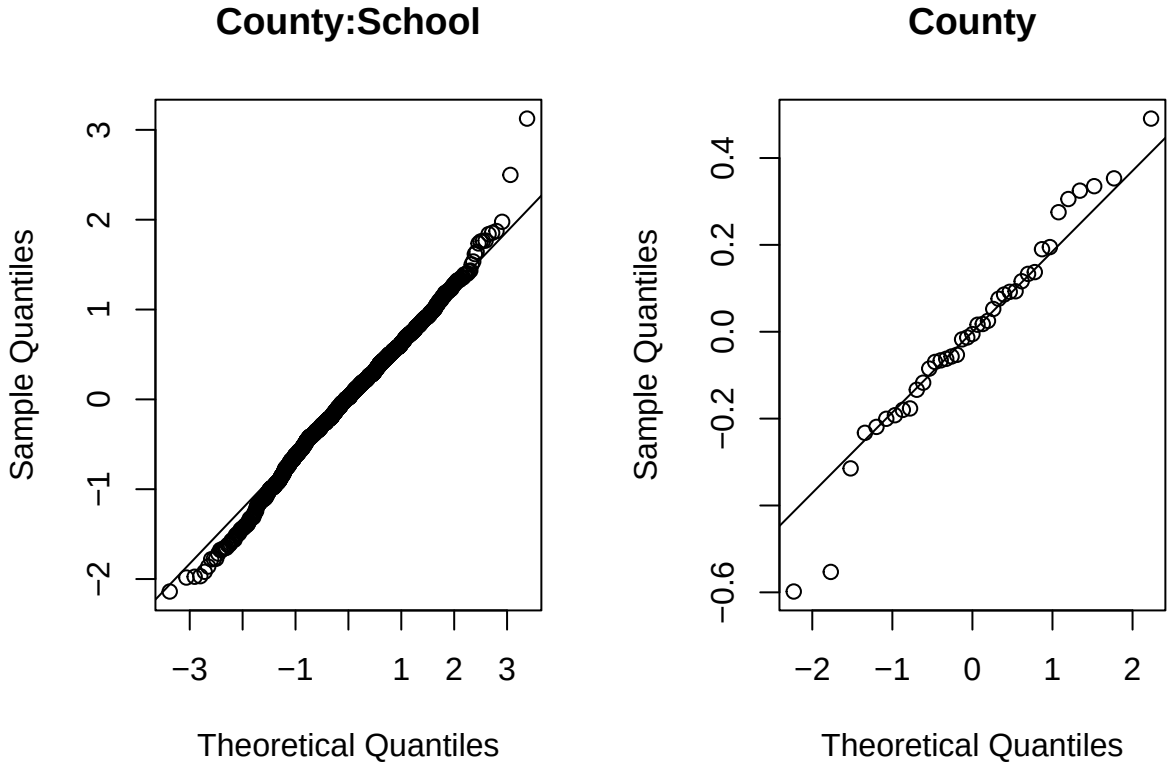
```
##  simulated
##
## data:  simulationOutput
## dispersion = 1.1855, p-value < 2.2e-16
## alternative hypothesis: two.sided
```




```
##
##  DHARMa bootstrapped outlier test
##
## data:  sim
## outliers at both margin(s) = 45, observations = 4525, p-value = 0.66
## alternative hypothesis: two.sided
##  percent confidence interval:
##  0.005071823 0.014922652
## sample estimates:
## outlier frequency (expected: 0.00935027624309392 )
##                                         0.009944751
```

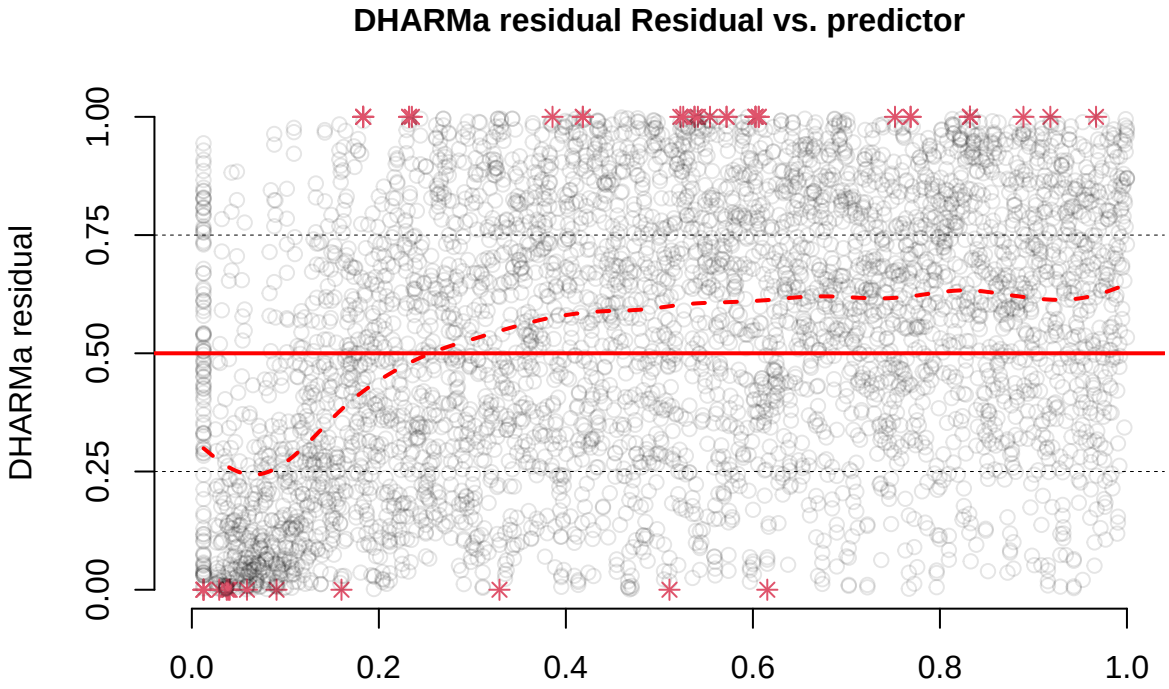
```
## Warning in ensurePredictor(simulationOutput, form): DHARMa:::ensurePredictor:
## character string was provided as predictor. DHARMa has converted to factor
## automatically. To remove this warning, please convert to factor before
## attempting to plot with DHARMa.
```

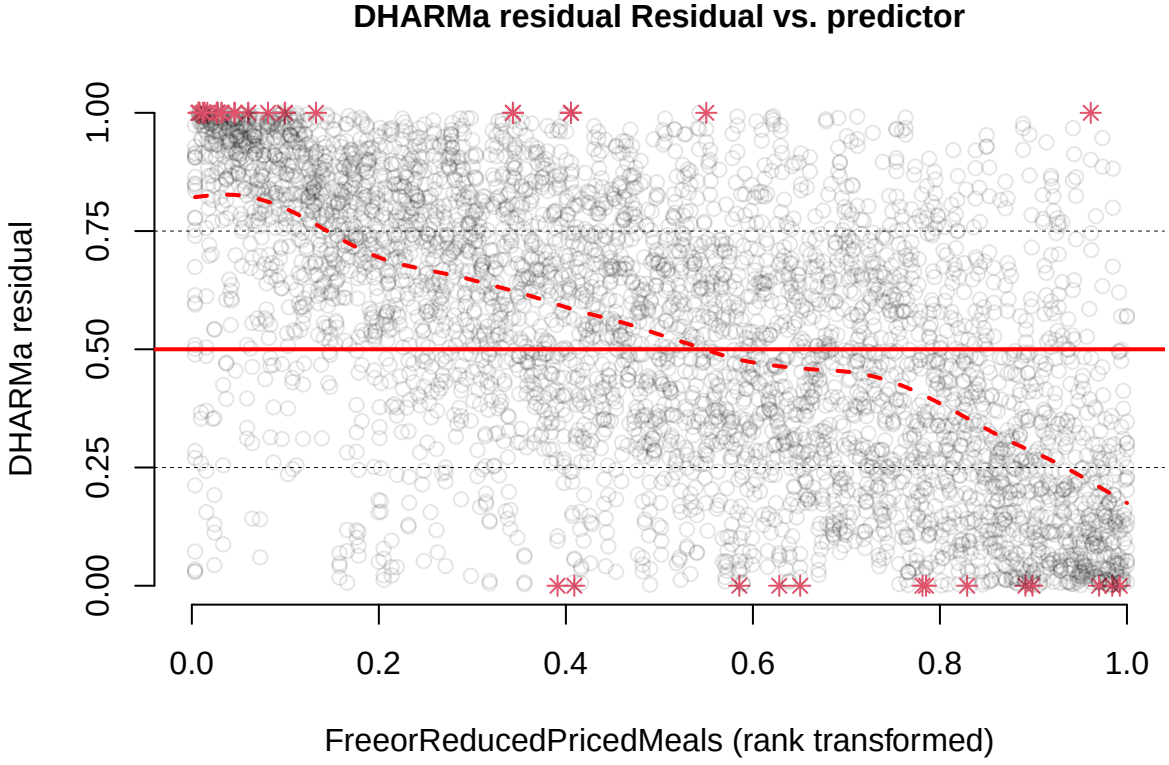




The `1|County:School` intercept reduces our BIC by over 40%! Furthermore, the DHARMa QQ plot shows a better shape and improvements across the board otherwise. Note that, as stated in the message in the code above, there is a tendency of the default outlier test to have inflated type 1 error rates for binomial data. Running the suggested alternative shows that the outliers are under control. Now we are in somewhat uncharted territory, but our 'base' model accounting for the 'global' factors is established. Let's look at the demographic factors. We can actually use DHARMa to help scout for potentially-usable variables in our

model.

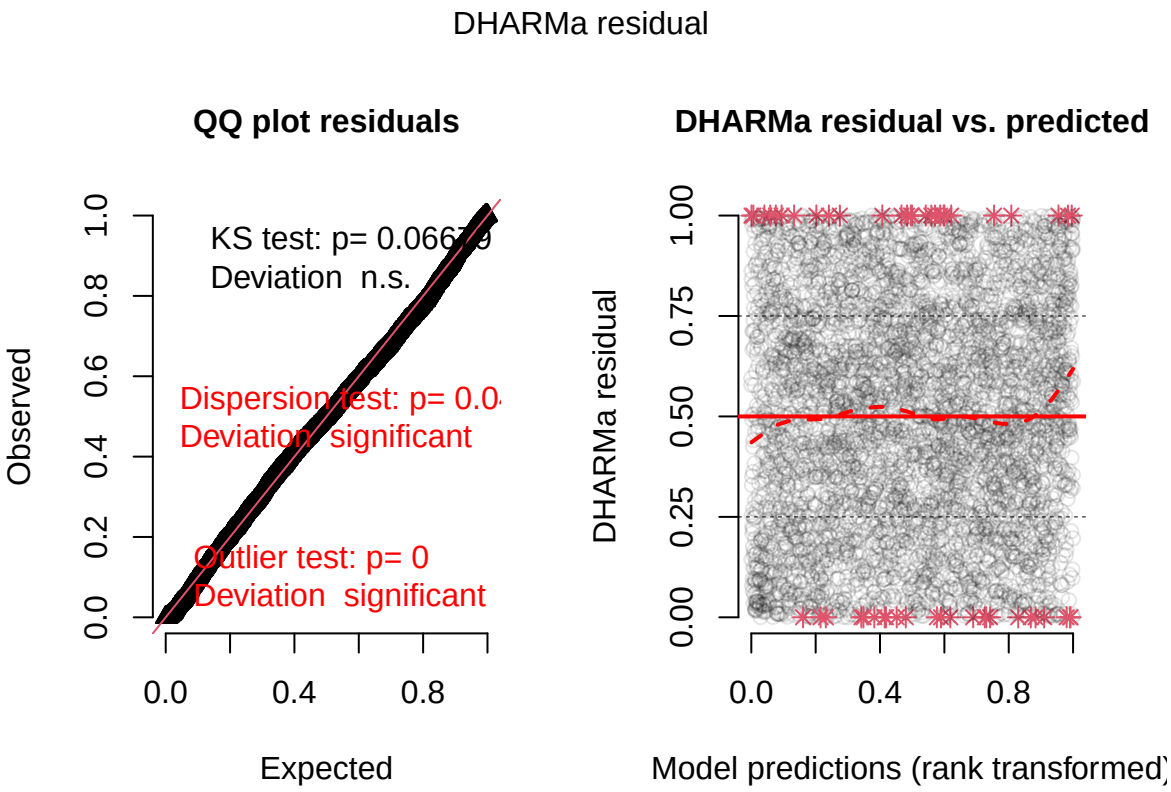


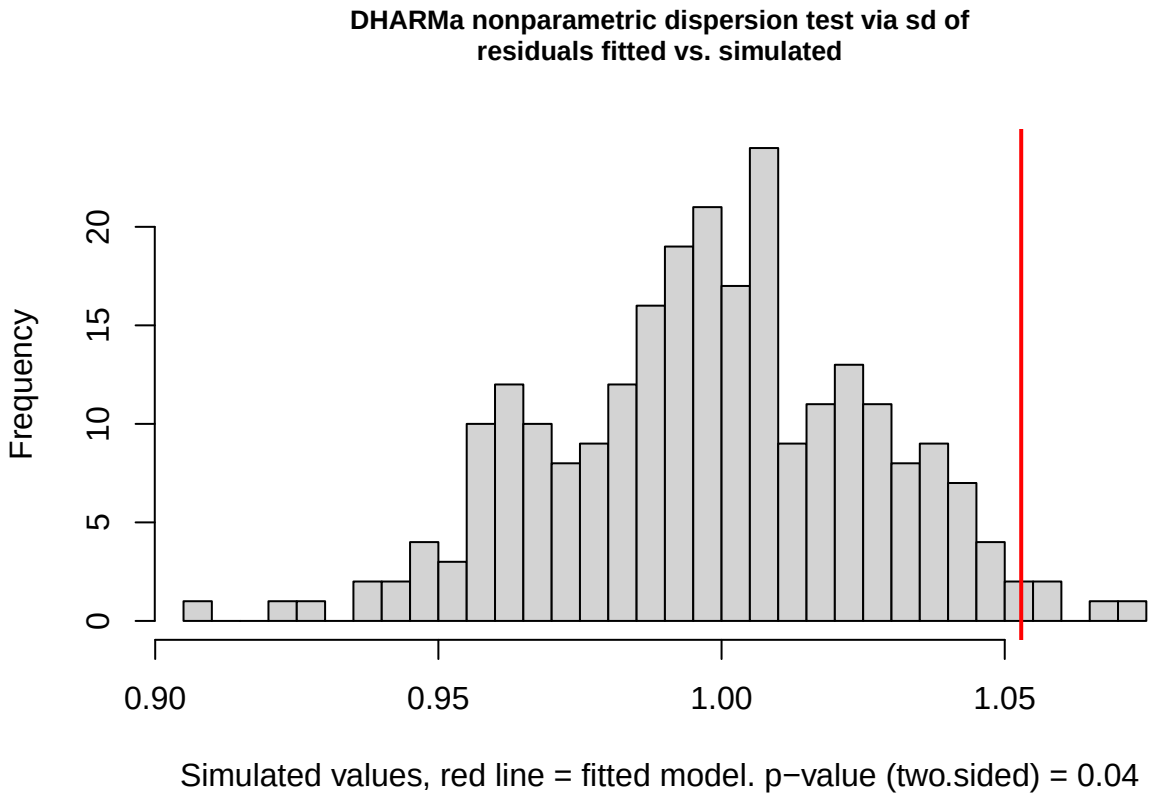


It looks like including effects for both %white and %reduced meal cost could help address some of the outliers. We may want to even try a random slope for %reduced meal cost. To circumvent convergence issues, we first scale our continuous predictors (and do logit transforms on the two proportional ones).

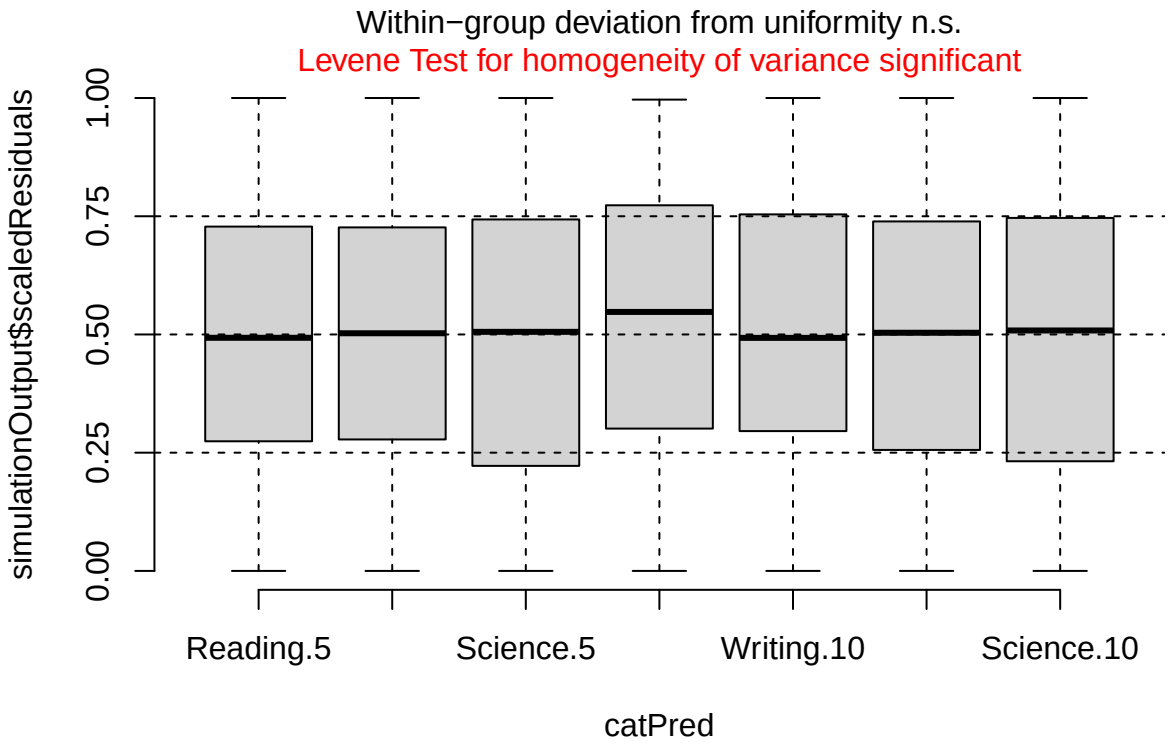
```
## Data: wasl_num
```

```
## Models:
## modboth2: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + (1 | County/School)
## modw: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + PercentWhite_logit + (1 | Coun
## modm: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals_logit
## modwm: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals_logi
##          npar   AIC   BIC logLik deviance   Chisq Df Pr(>Chisq)
## modboth2    9 30249 30307 -15116    30231
## modw       10 29928 29992 -14954    29908 323.665  1  < 2.2e-16 ***
## modm       10 29569 29634 -14775    29549 358.275  0
## modwm      11 29514 29585 -14746    29492  57.042  1  4.266e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```

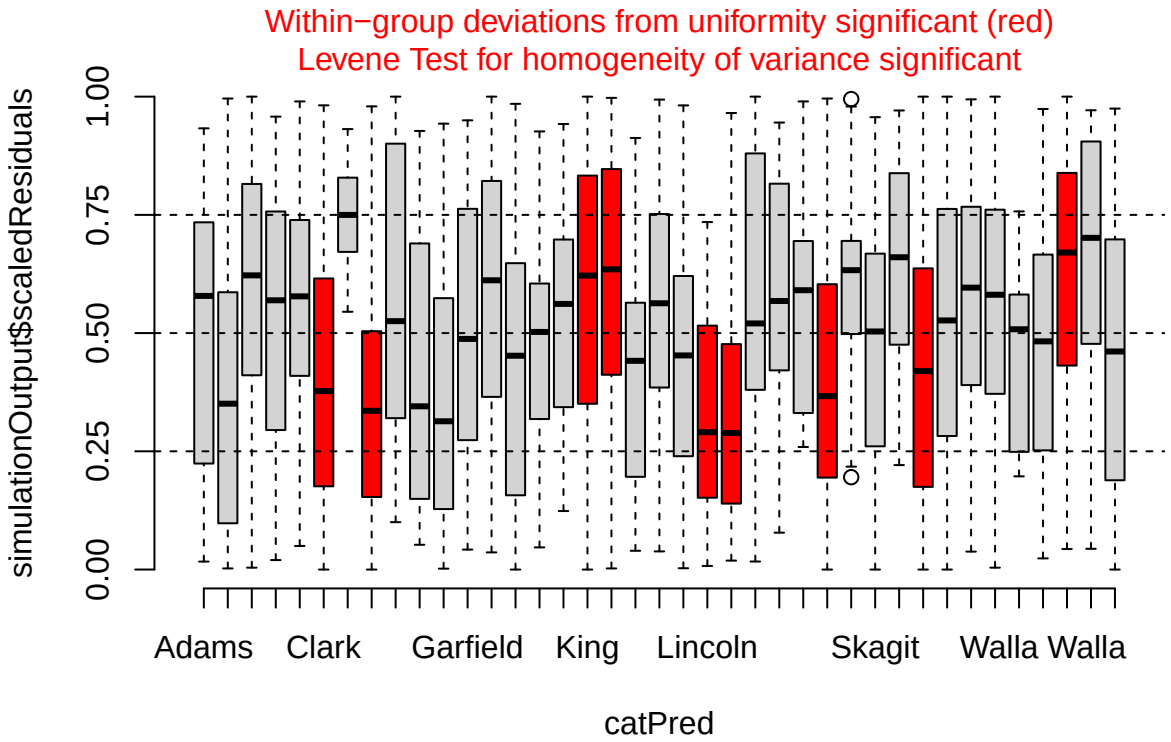


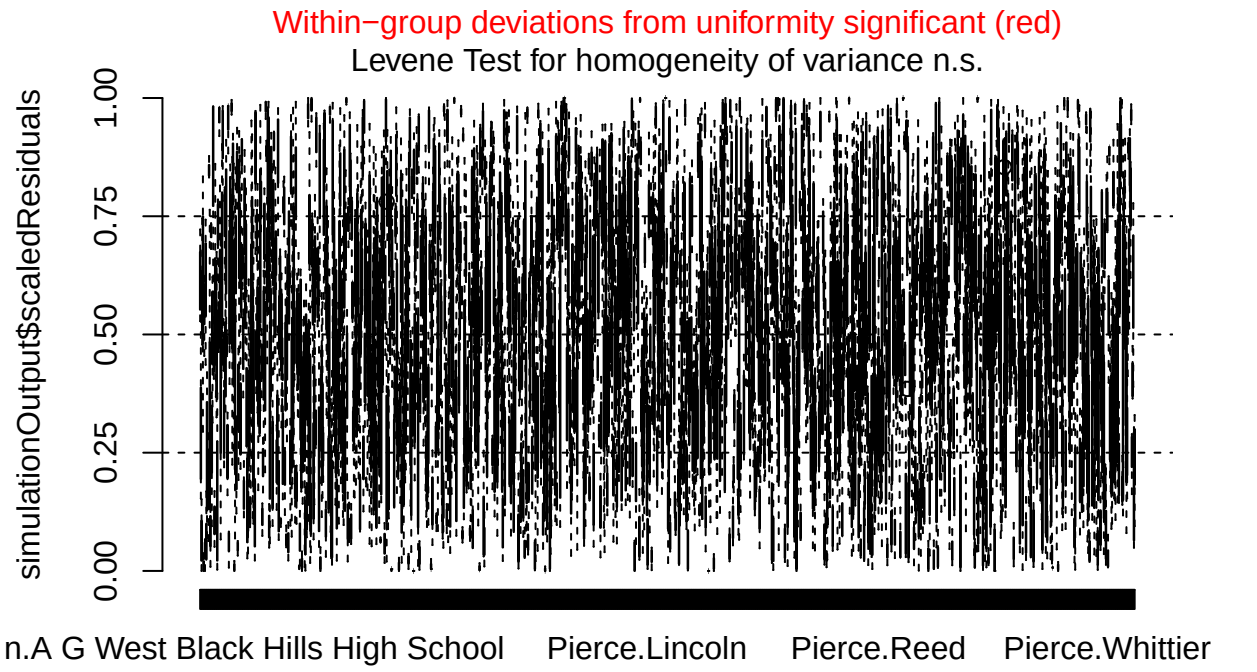


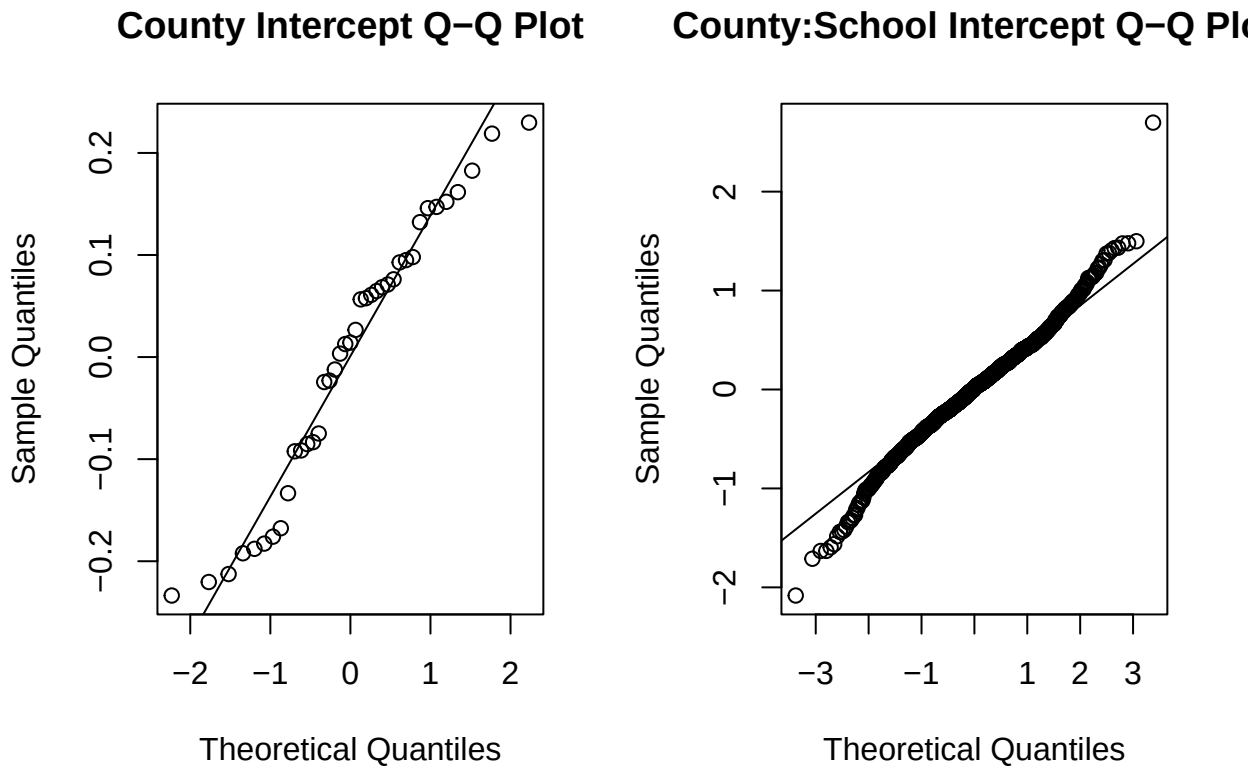

```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.055, p-value = 0.04
## alternative hypothesis: two.sided
```



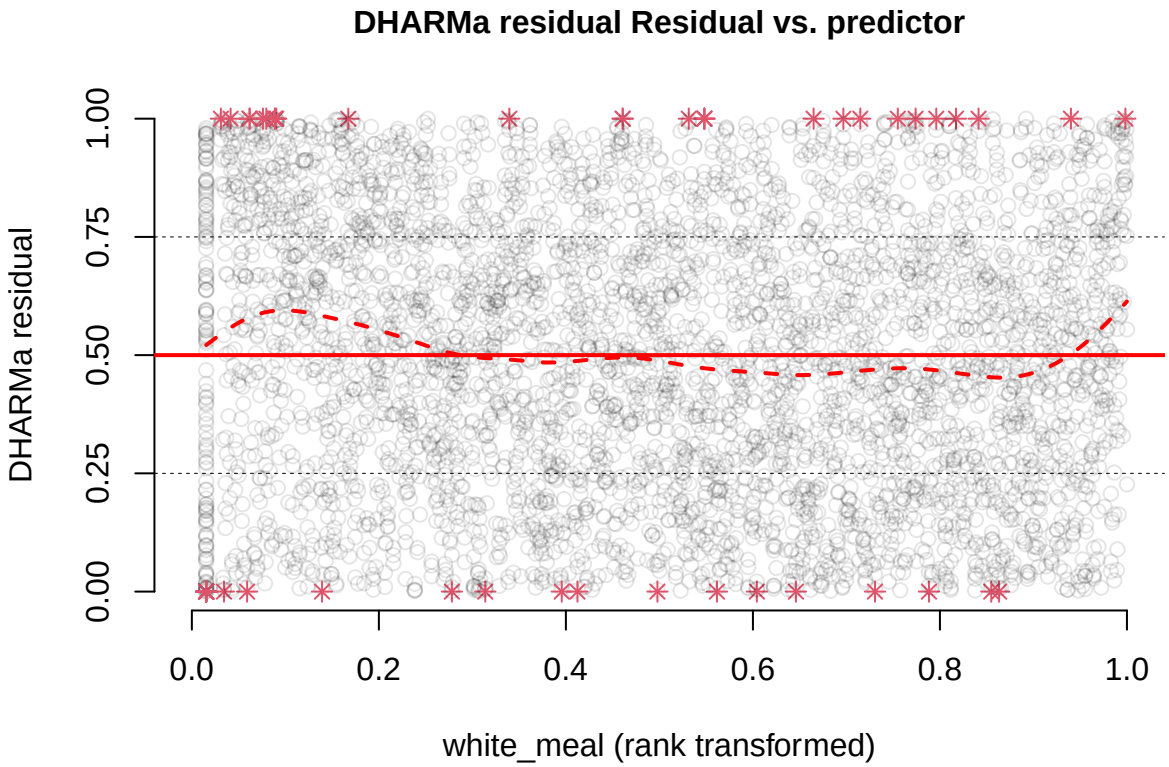

```
## Warning in ensurePredictor(simulationOutput, form): DHARMa:::ensurePredictor:
## character string was provided as predictor. DHARMa has converted to factor
## automatically. To remove this warning, please convert to factor before
## attempting to plot with DHARMa.
```







Including both %white and %reduced cost meals improved our BIC by a lot, and substantially decreased within-group deviation for Subject_Grade and County, as well as homogenized the County:School variance. Let's check to see if there may be an interaction effect betweeen these two variables



It doesn't look like much, but it's worth trying:

```
## Data: wasl_num
## Models:
## modwm: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals_logi
## modwm_int: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + white_meal_logit + Percen
##           npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## modwm       11 29514 29585 -14746    29492
## modwm_int   12 29516 29593 -14746    29492 0.4201  1     0.5169
```

It didn't help. Recall from before that there may be evidence for a random slope of %reduced price meals instead of (or perhaps along with) a fixed effect. I can't figure out how to plot it in a manner that will let me discern what the random intercept should be for the slope, so we will just have to check a few possible terms.

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient
```

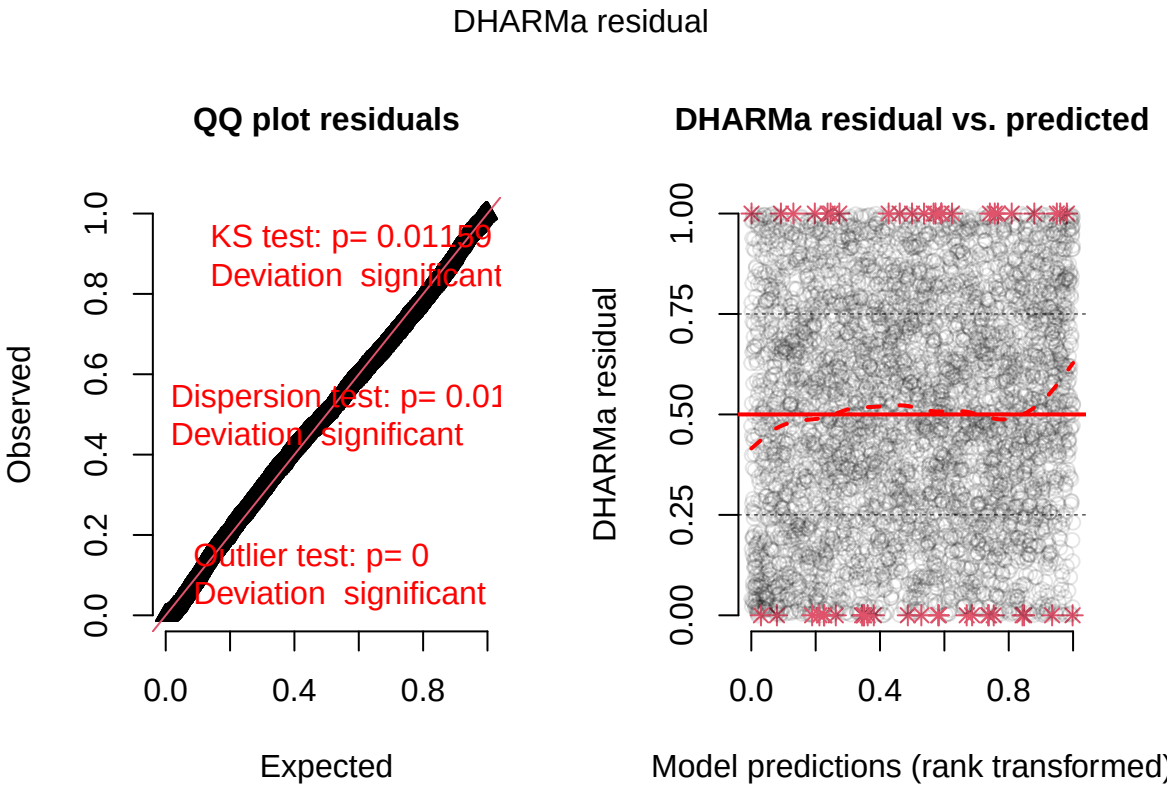
```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues
```

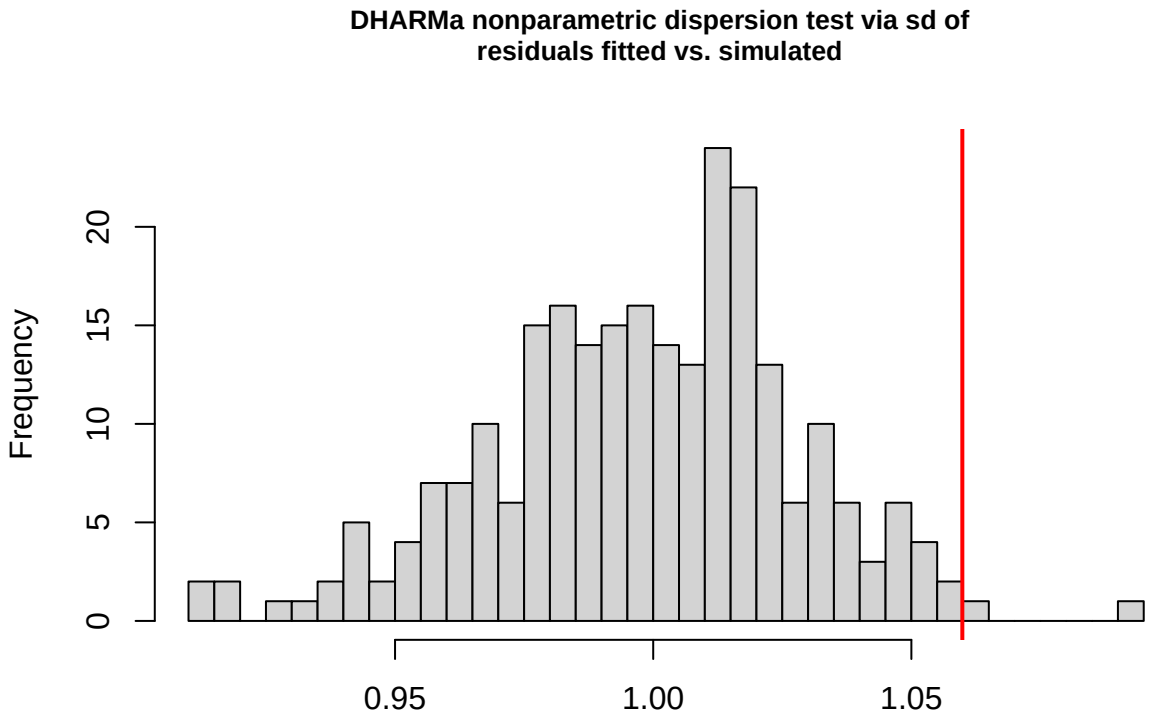
```
## boundary (singular) fit: see help('isSingular')
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues
```

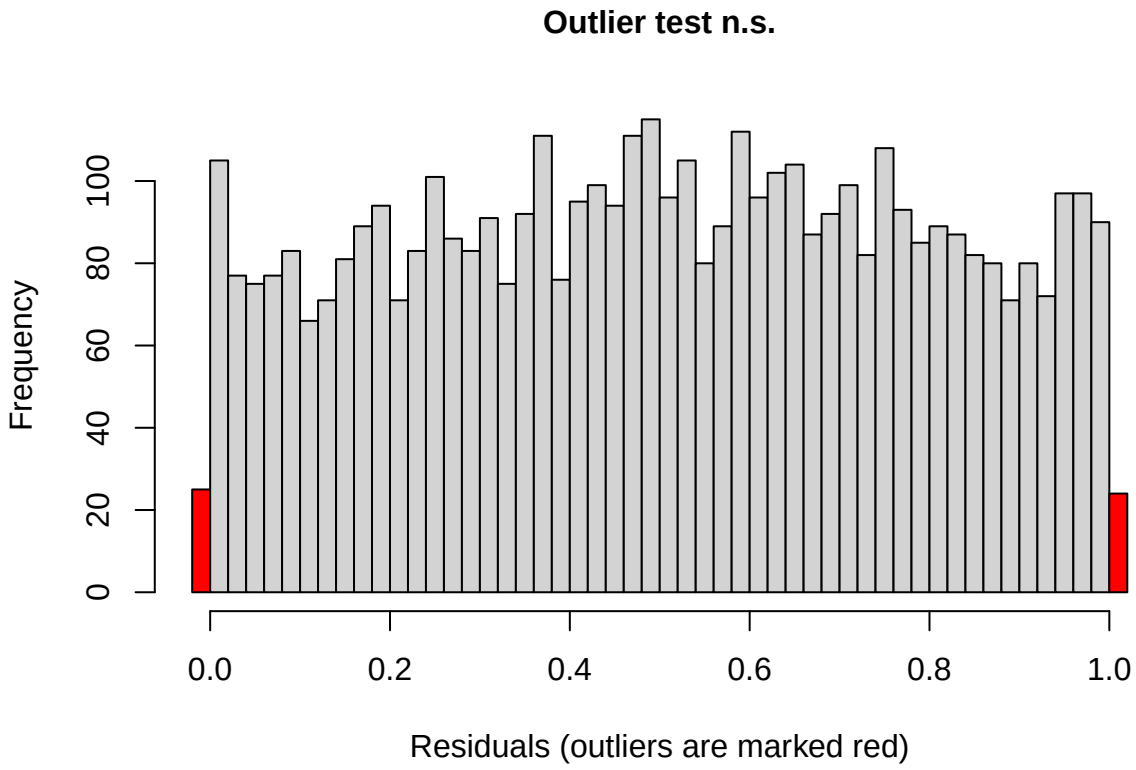
```
## Data: wasl_num
## Models:
## modwm: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals_logi
## modwm_slope_c: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + PercentWhite_logit +
## modwm_both_c: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMea
```

```
## modwm_slope_cs: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + PercentWhite_logit +
## modwm_both_cs: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMea
## modwm_both_c_corr: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPrice
## modwm_both_cs_corr: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPric
##                     npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## modwm                 11 29514 29585 -14746    29492
## modwm_slope_c         13 29530 29613 -14752    29504   0.00  2          1
## modwm_both_c          13 29474 29558 -14724    29448  55.38  0
## modwm_slope_cs        13 29540 29623 -14757    29514   0.00  0
## modwm_both_cs         13 29304 29388 -14639    29278 235.07  0
## modwm_both_c_corr     14 29476 29566 -14724    29448   0.00  1          1
## modwm_both_cs_corr    14 29300 29390 -14636    29272 176.54  0

## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dist
```

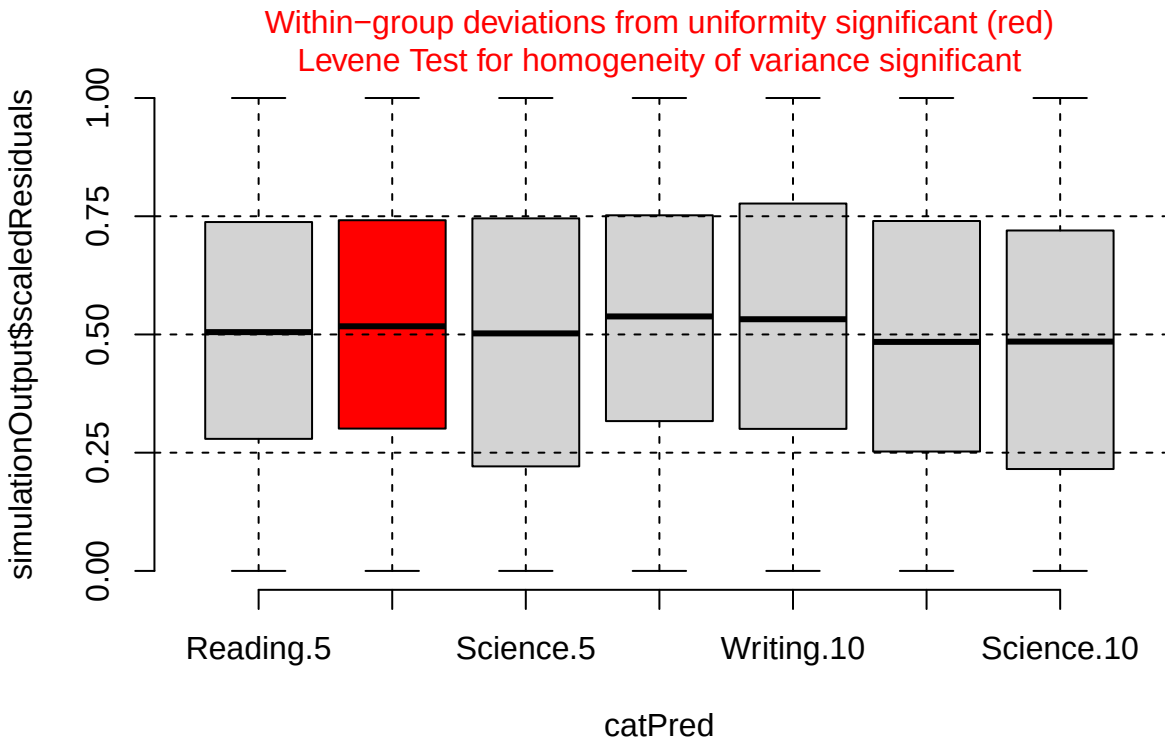




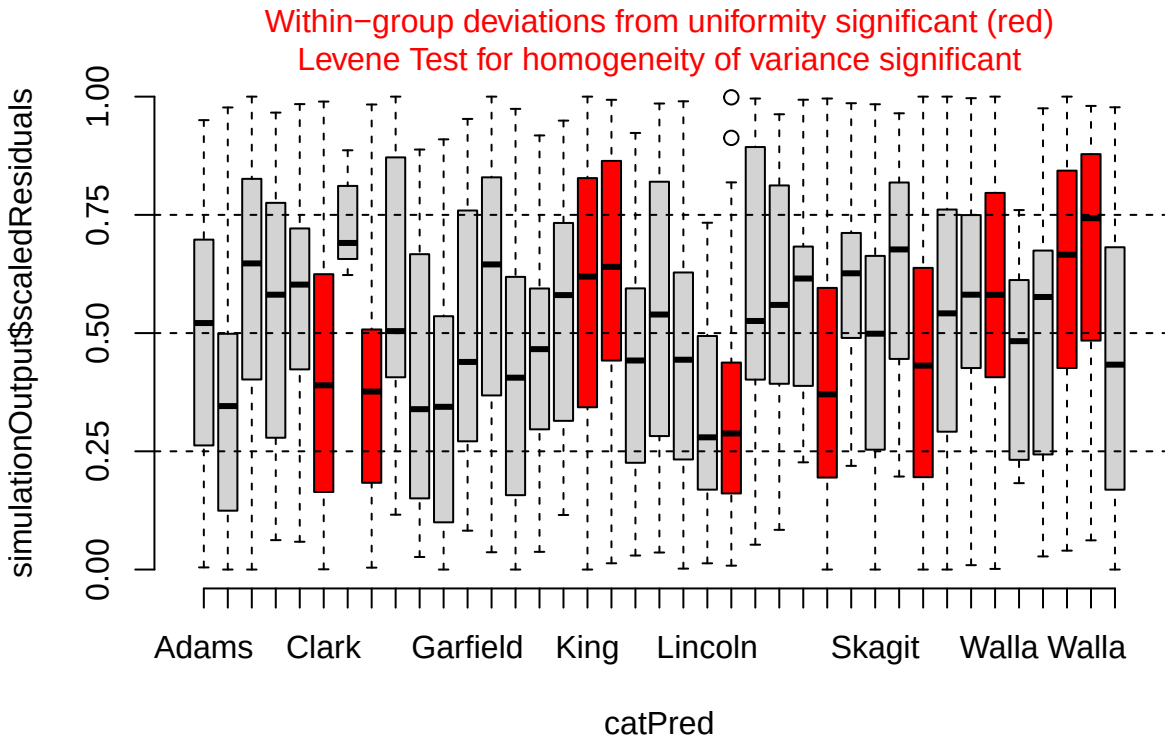

```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.062, p-value = 0.016
## alternative hypothesis: two.sided
```

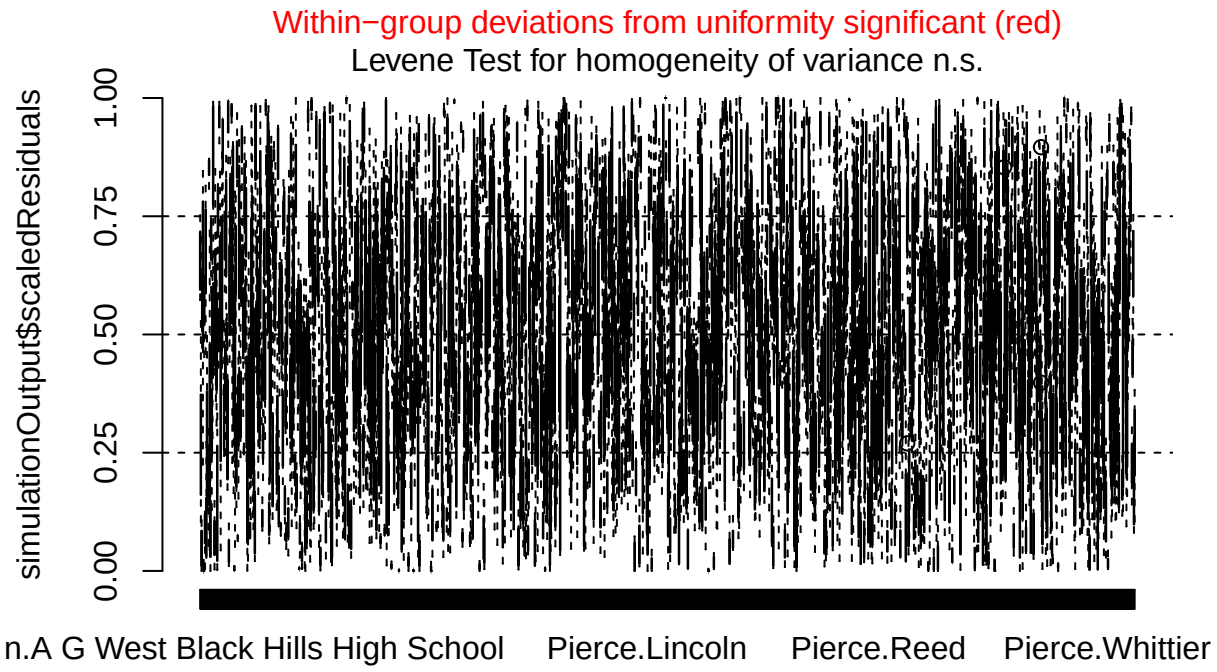


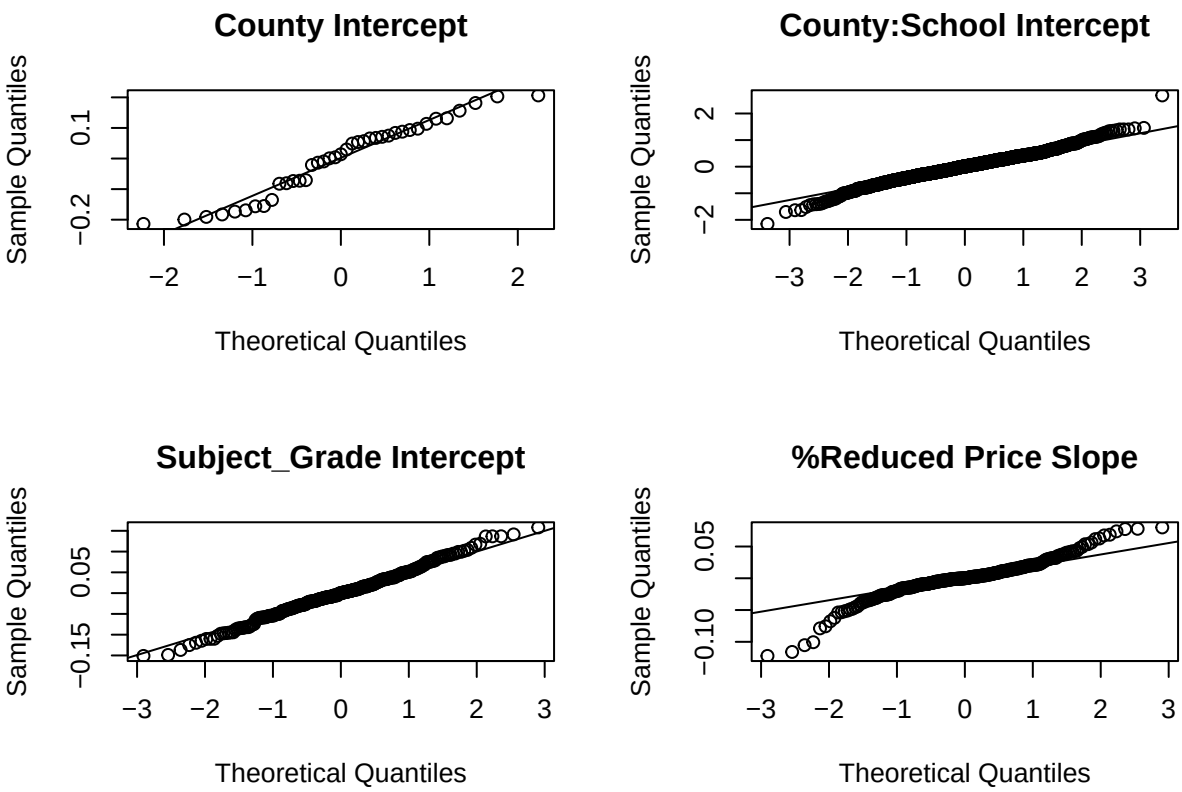

```
##
##  DHARMa bootstrapped outlier test
##
## data:  sim
## outliers at both margin(s) = 49, observations = 4525, p-value = 0.56
## alternative hypothesis: two.sided
##  percent confidence interval:
##  0.005944751 0.014745856
## sample estimates:
## outlier frequency (expected: 0.00959779005524862 )
##                                         0.01082873
```




```
## Warning in ensurePredictor(simulationOutput, form): DHARMa:::ensurePredictor:
## character string was provided as predictor. DHARMa has converted to factor
## automatically. To remove this warning, please convert to factor before
## attempting to plot with DHARMa.
```

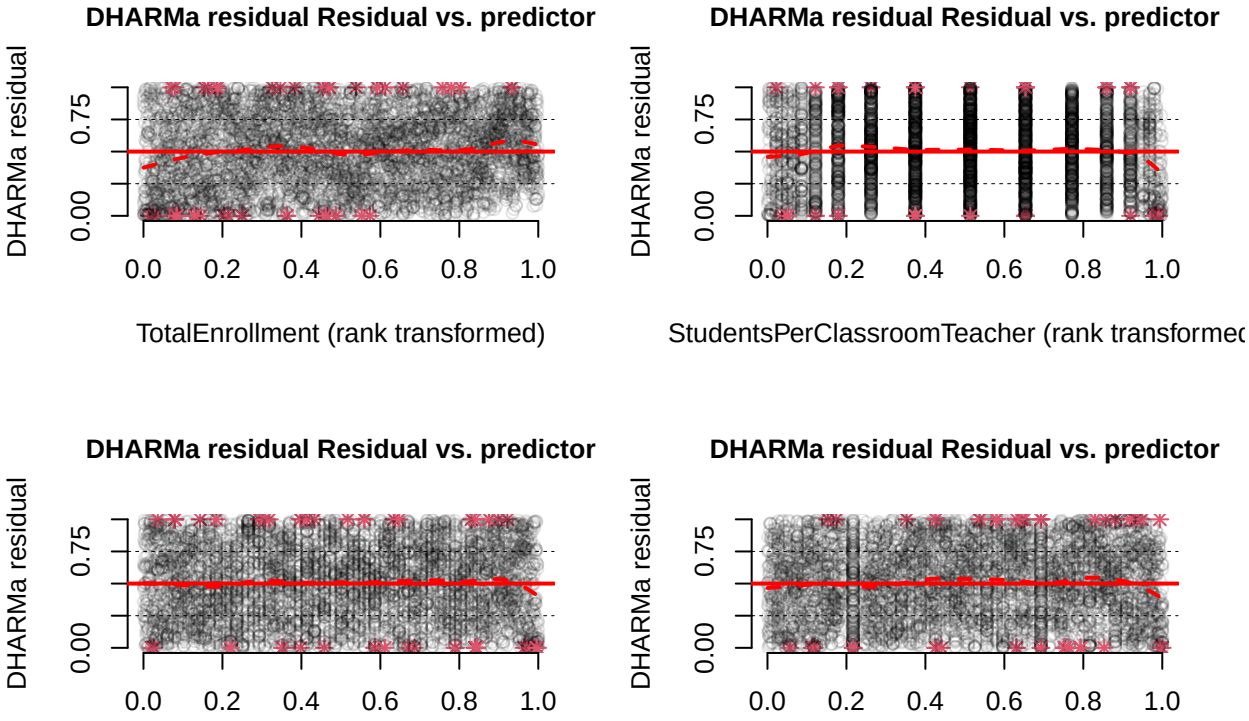






The uncorrelated random slope improves the model further, but also seems to have gotten us back in trouble in terms of residual variance metrics.. More importantly, does the term make sense? In our EDA, we saw that %reduced cost meals had pretty similar linear fits between grades, counties, and subjects. There was some erratic behavior with math and science for 10th graders, both in terms of direction and magnitude of the slopes, so it isn't shocking that the random slope would improve the model. But why uncorrelated? From

a practical standpoint, it makes the model less complicated. But we are concerned with why it made the model 'better'. Perhaps the counties in which we saw improvement in math and/or science with increases in %reduced cost meal plans (and therefore reduced expected scores and a lower intercept) were enough to move the needle. Regardless, with the demographic variables accounted for, we only have the resource variables to investigate.



It looks like there may be a slight effect of total enrollment, and maybe a negligible effect of student:teacher ratio at the high end, but the other two seem to have almost nothing. From our preliminary checks, it looked like it might be worth checking the effect of log(enrollment), so we will try that as well.

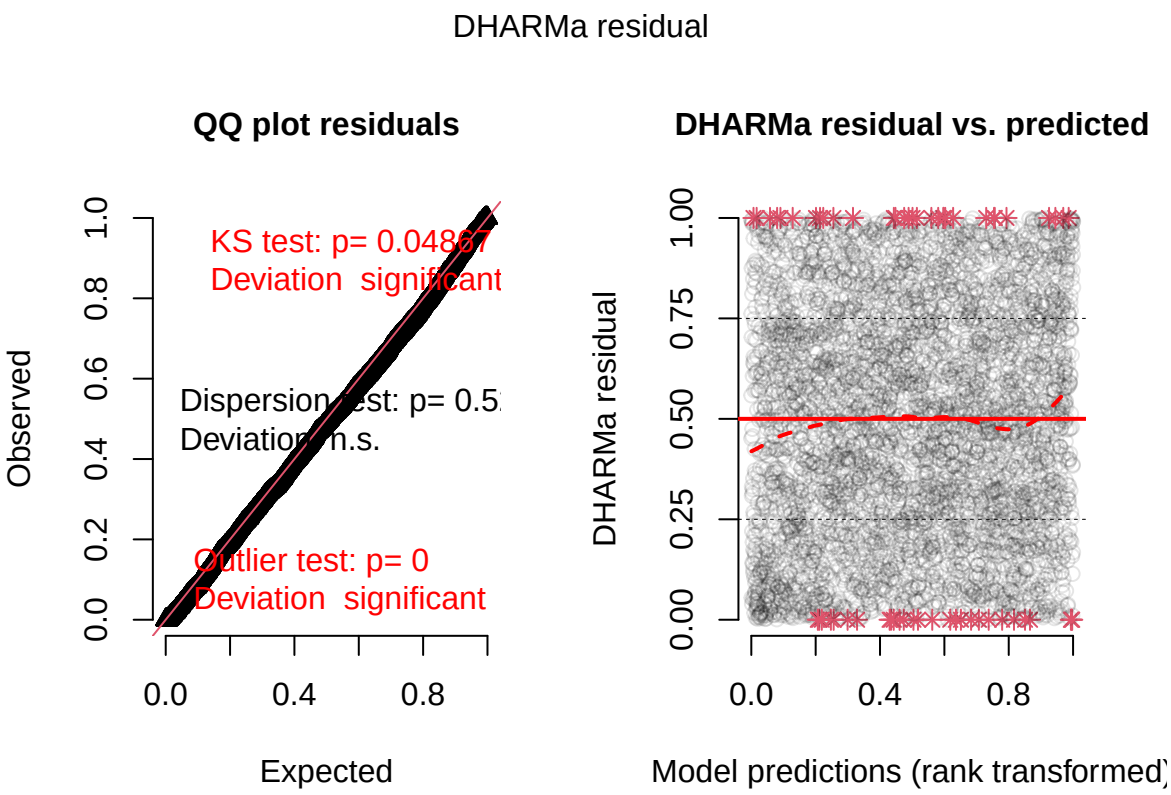
```
## Data: wasl_num
## Models:
## modwm_both_cs: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMea
## mod_enroll: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals
## mod_log: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals_log
## mod_str: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals_log
## mod_stroll: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals
## mod_stroll_slope: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPrice
##                  npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## modwm_both_cs      13 29304 29388 -14639    29278
## mod_enroll         14 29283 29372 -14627    29255 23.790  1  1.074e-06 ***
## mod_log            14 29269 29359 -14621    29241 13.397  0
## mod_str            14 29297 29387 -14634    29269  0.000  0
## mod_stroll         15 29269 29365 -14619    29239 30.141  1  4.017e-08 ***
## mod_stroll_slope   17 29180 29289 -14573    29146 93.079  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

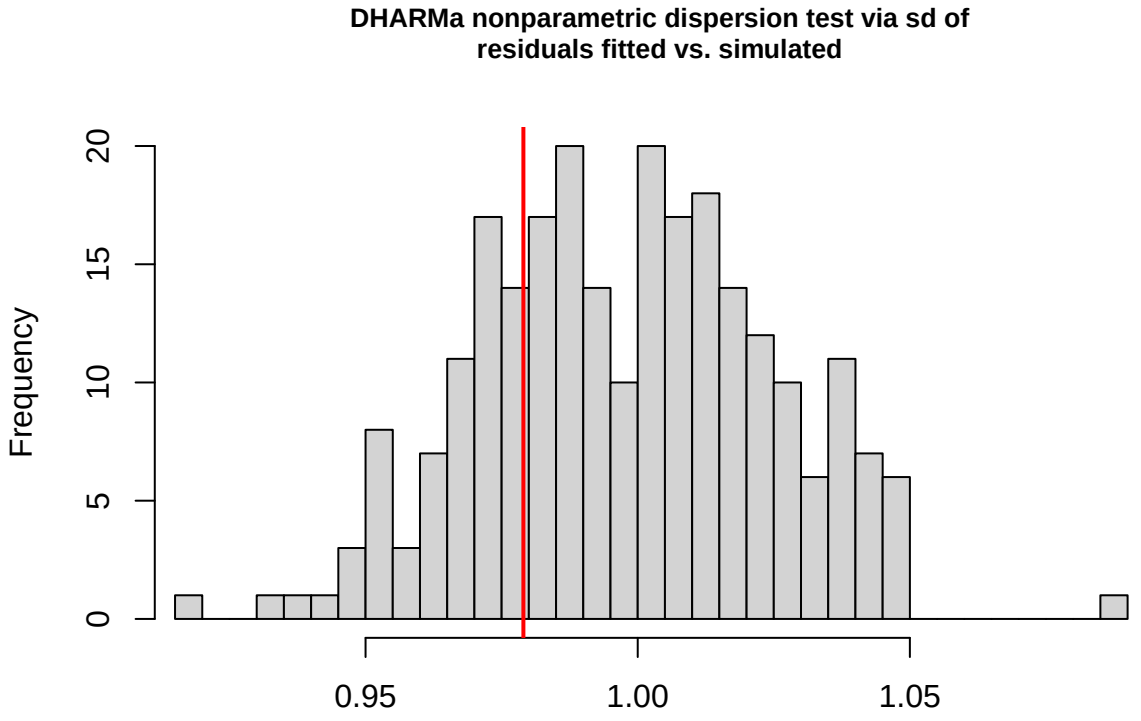
Adding enrollment and student:teacher ratio as fixed effects yielded some more good improvements, and adding a random slope for enrollment at the school within county level helped even more. It makes sense

that the effect of enrollment could vary between schools within counties. For instance, in a county with a lot of big schools (like where I went to school - Los Angeles County), having tons of students may not mean as much as demographic factors do, and maybe schools with fewer students (like small private schools) will outperform those with more students. However, in a smaller county, there may only be one or two schools with lots of students, and those schools will likely have disproportionate amounts of resources compared to the other small schools in the county, leading to generally better performance.

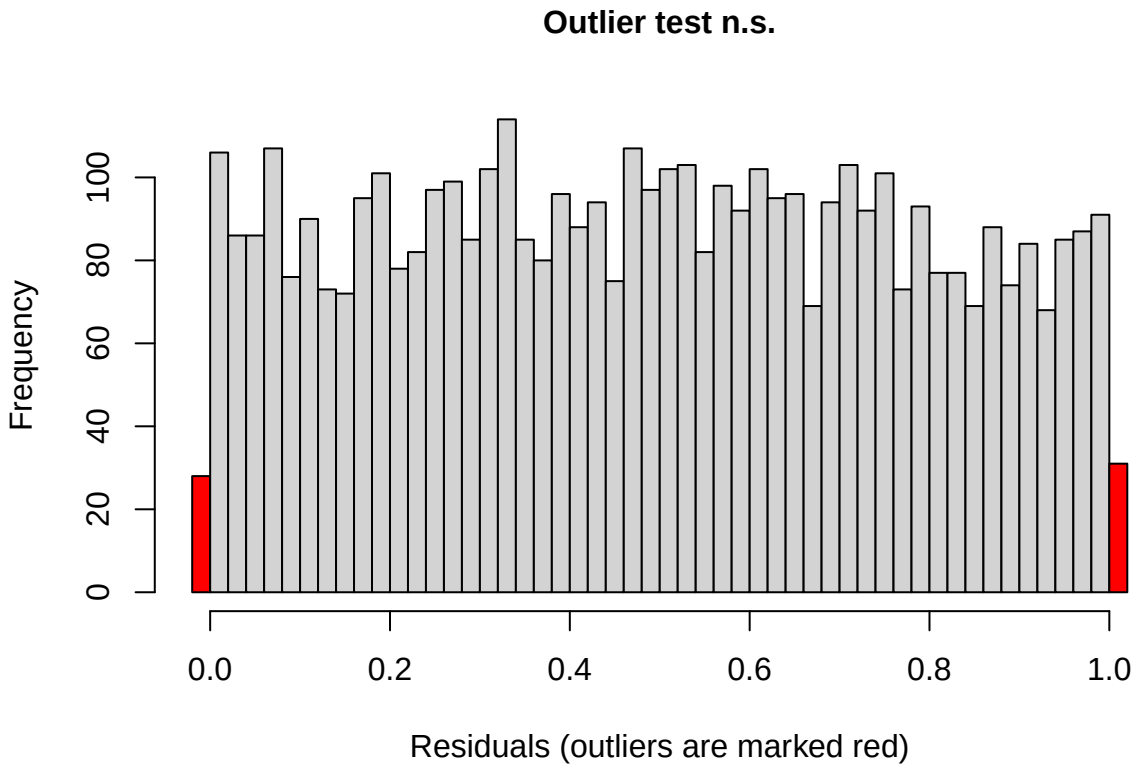
(note that using log(enroll) as a fixed effect and the scaled enroll for the random slope actually gave the best BIC, but I didn't think that was a justifiable addition, so I went with this model)

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```

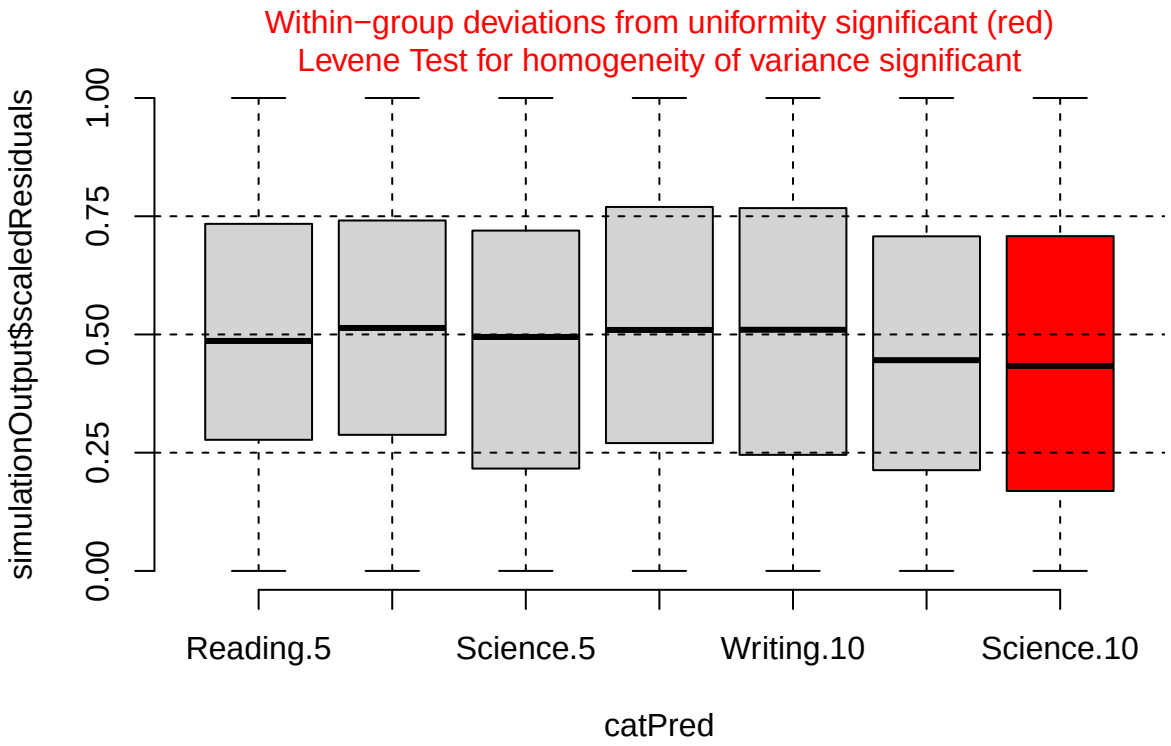




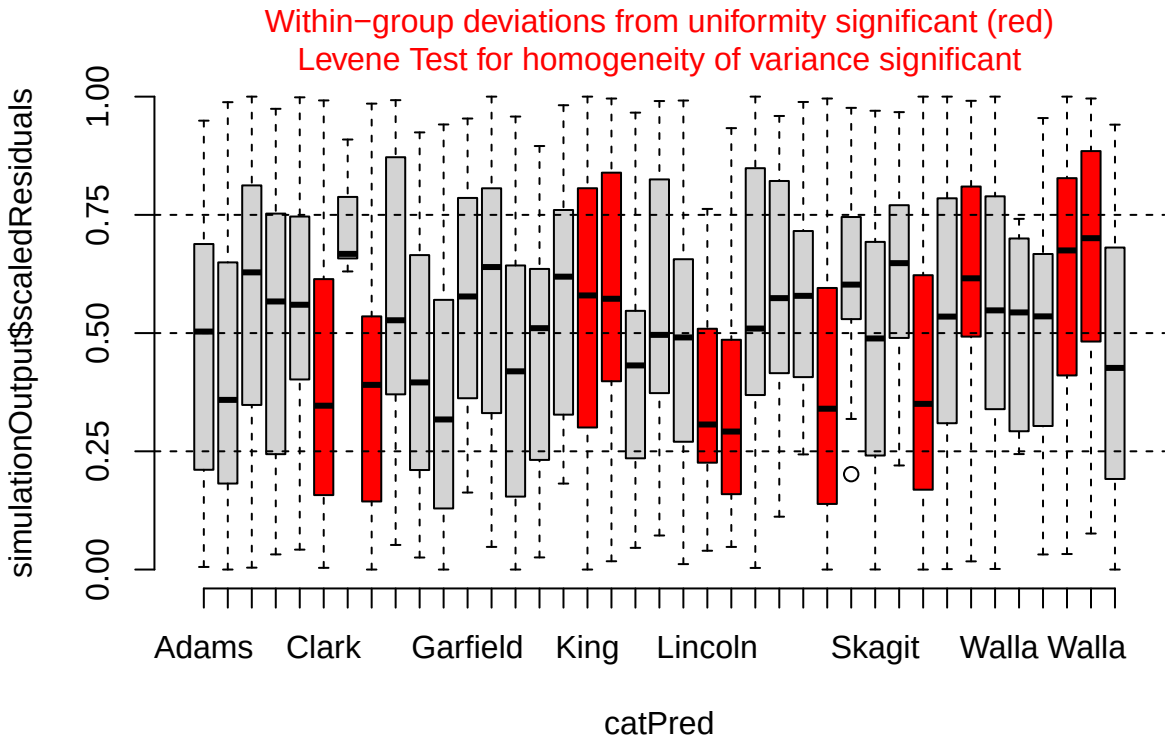

```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.98095, p-value = 0.52
## alternative hypothesis: two.sided
```

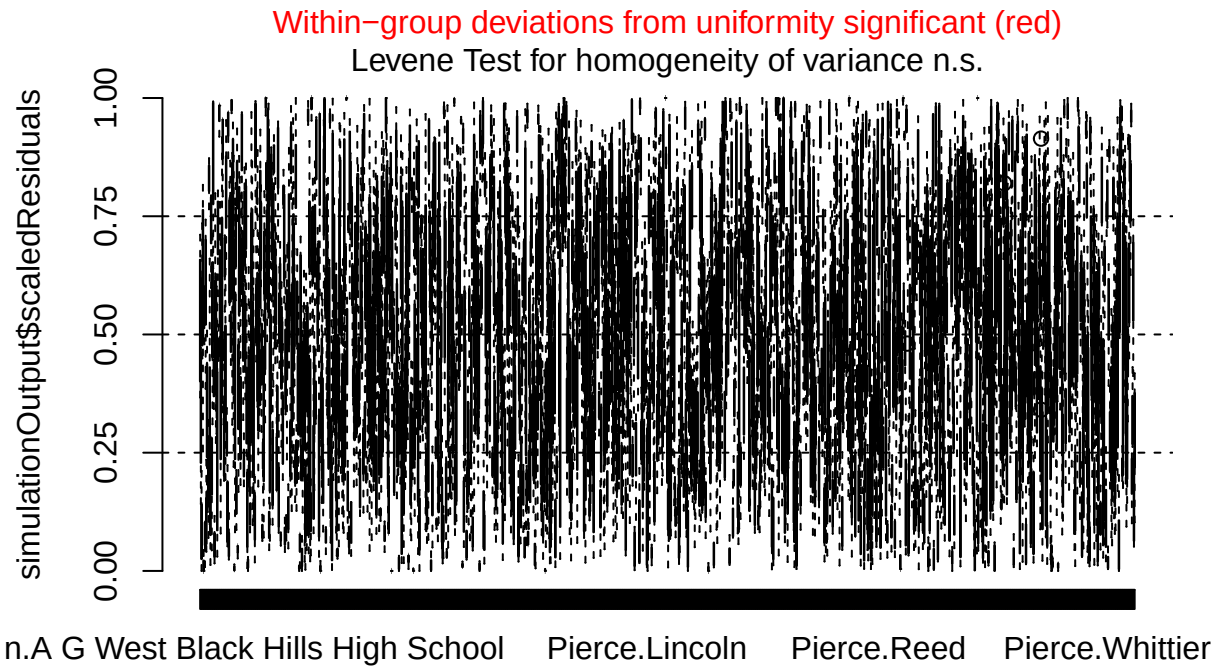


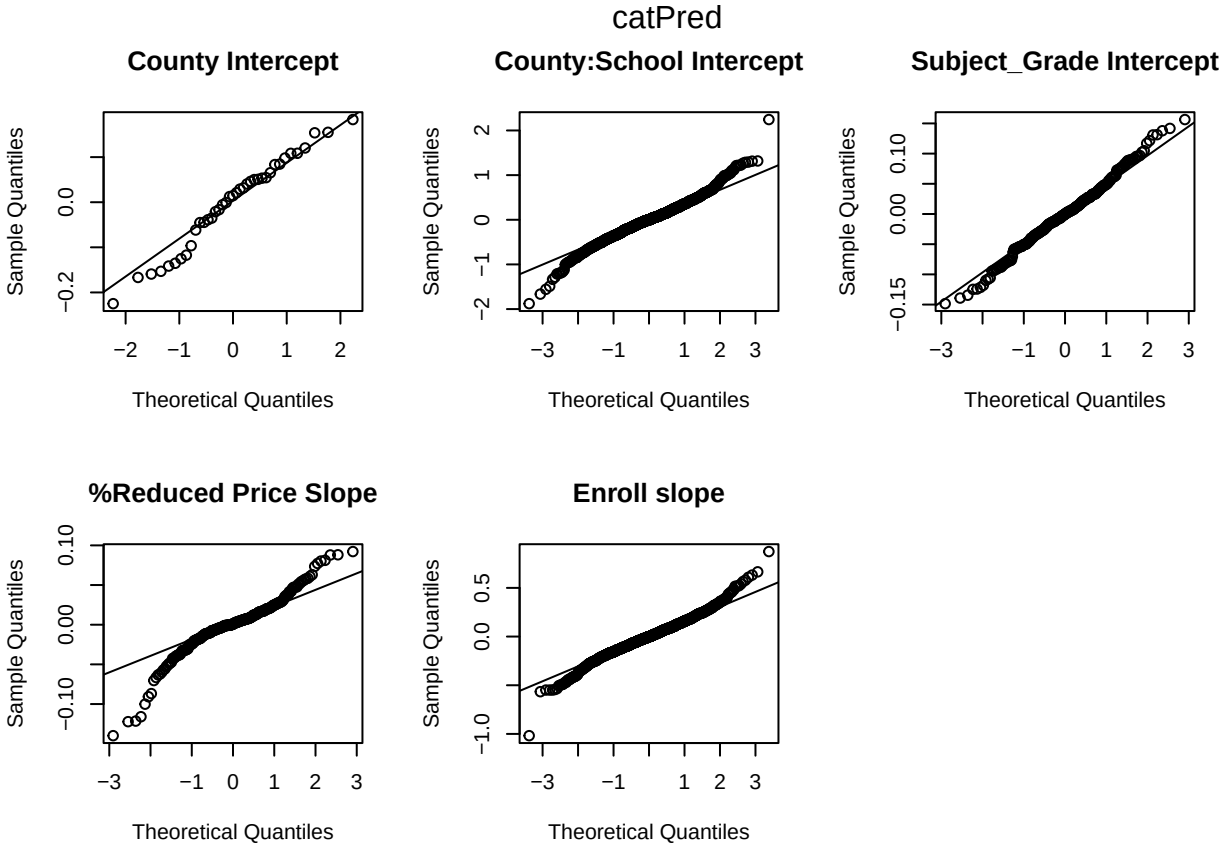

```
##
##  DHARMa bootstrapped outlier test
##
## data:  sim
## outliers at both margin(s) = 59, observations = 4525, p-value = 0.08
## alternative hypothesis: two.sided
##  percent confidence interval:
##  0.00640884 0.01371271
## sample estimates:
## outlier frequency (expected: 0.00947845303867403 )
##                                         0.01303867
```




```
## Warning in ensurePredictor(simulationOutput, form): DHARMa:::ensurePredictor:
## character string was provided as predictor. DHARMa has converted to factor
## automatically. To remove this warning, please convert to factor before
## attempting to plot with DHARMa.
```



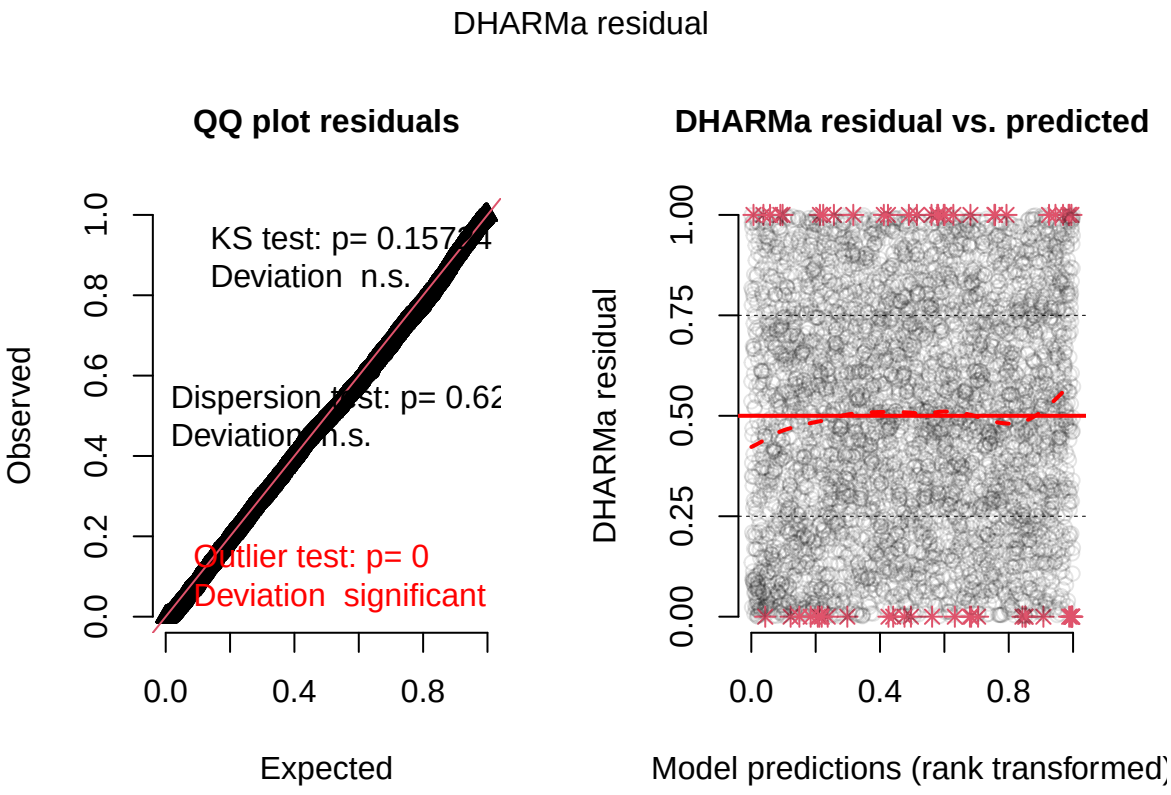


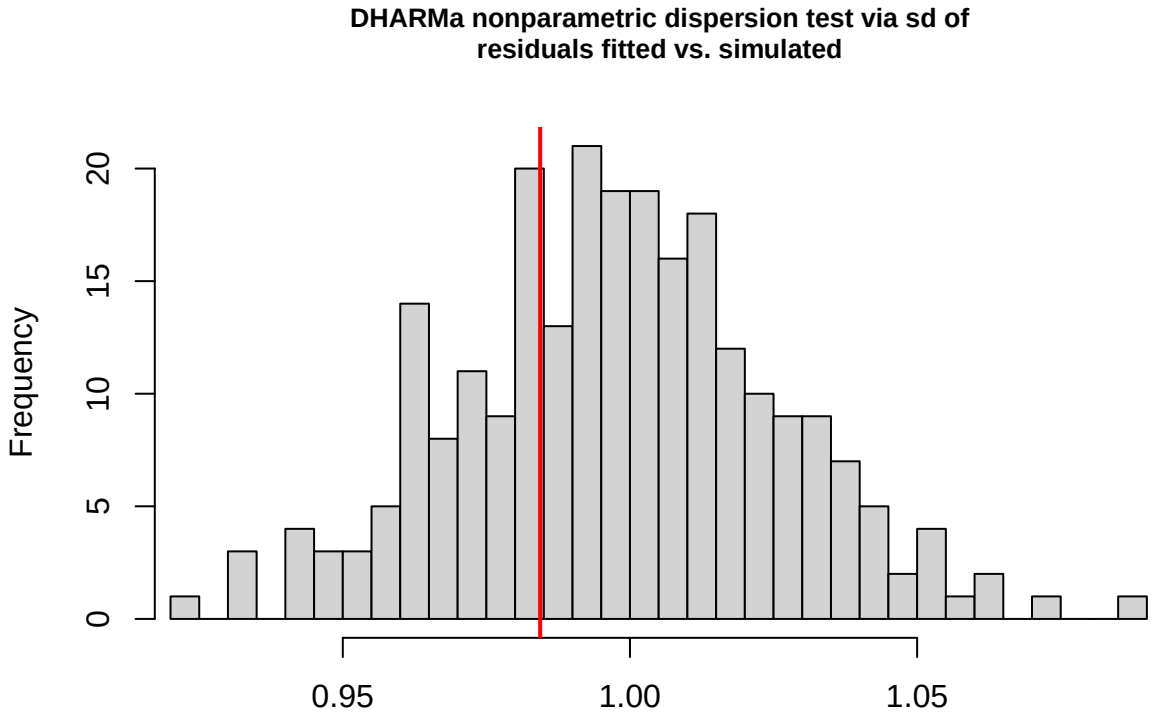


With the latest model, we have again cleared the dispersion and outlier hurdles, but our difficulties at the group level persist. One thing I am interested in testing out is whether %white could have a random slope. The plot earlier didn't give much indication of that, but we also couldn't see the grouping factors. Furthermore, our EDA showed that %white seemed to be all over the place between counties and subjects. Let's give it a try.

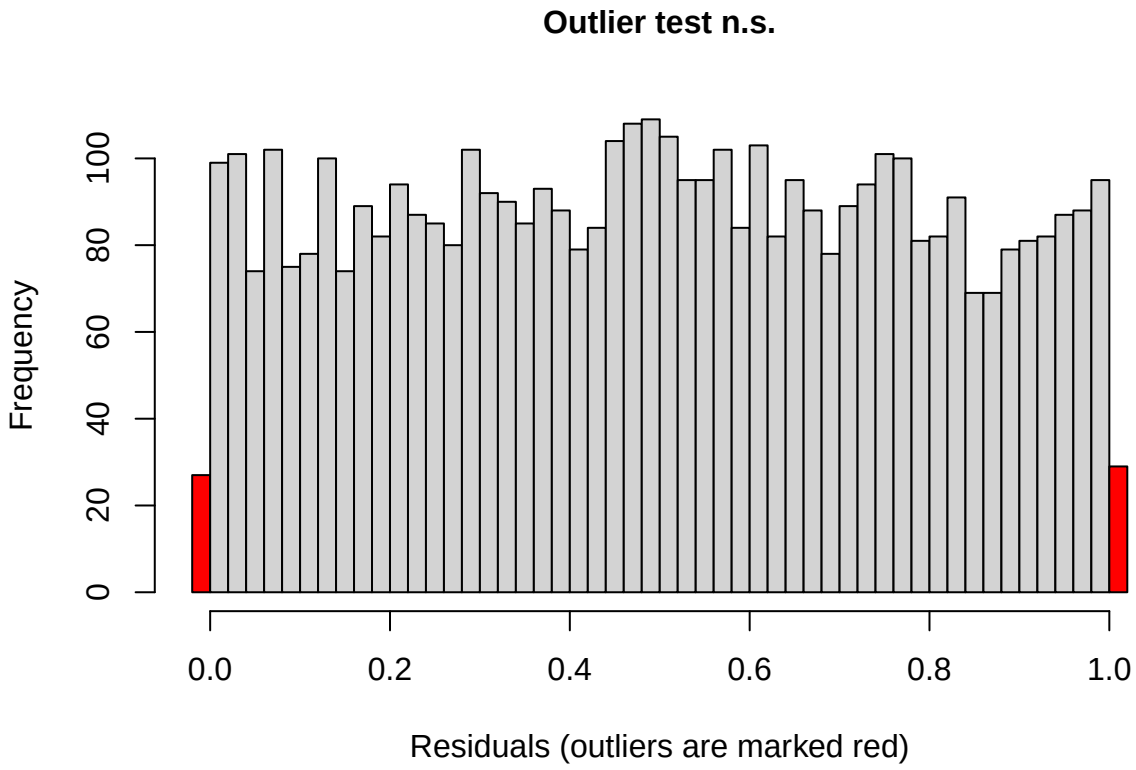
```
## Data: wasl_num
## Models:
## mod_stroll_slope: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPrice
## mod_stroll_wslope: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPric
##                   npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## mod_stroll_slope    17 29180 29289 -14573    29146
## mod_stroll_wslope   18 29172 29288 -14568    29136 9.3543  1   0.002225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```

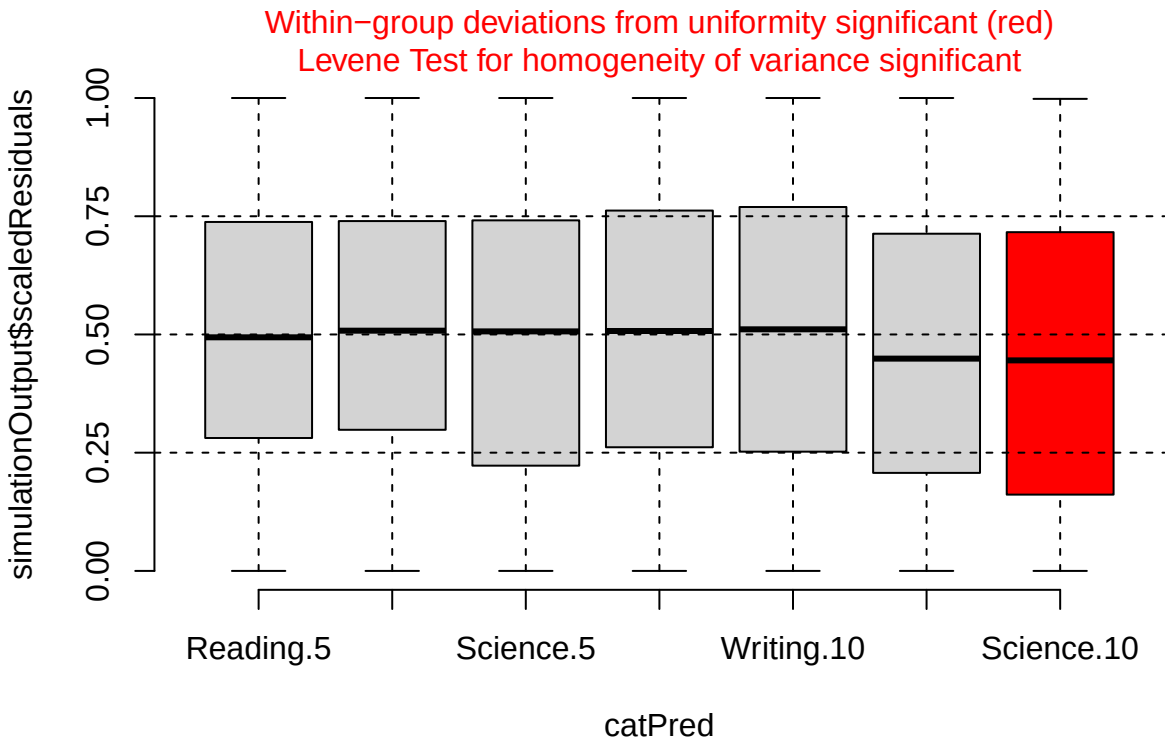




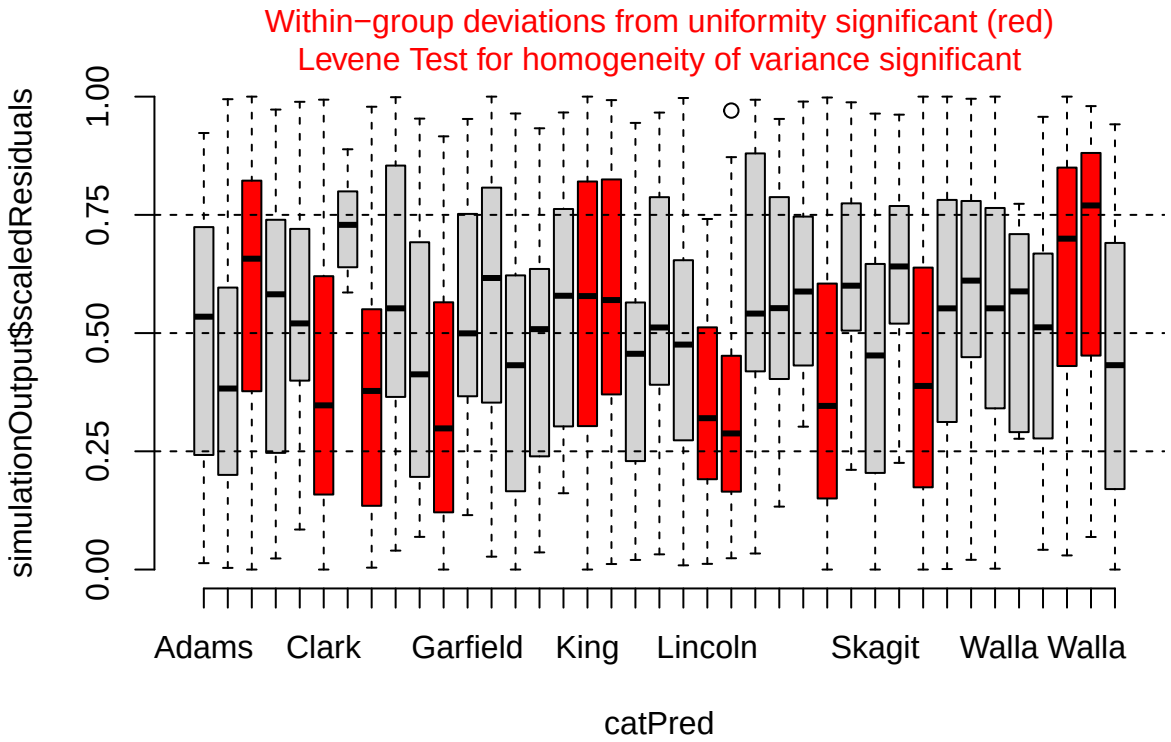

```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.98634, p-value = 0.624
## alternative hypothesis: two.sided
```

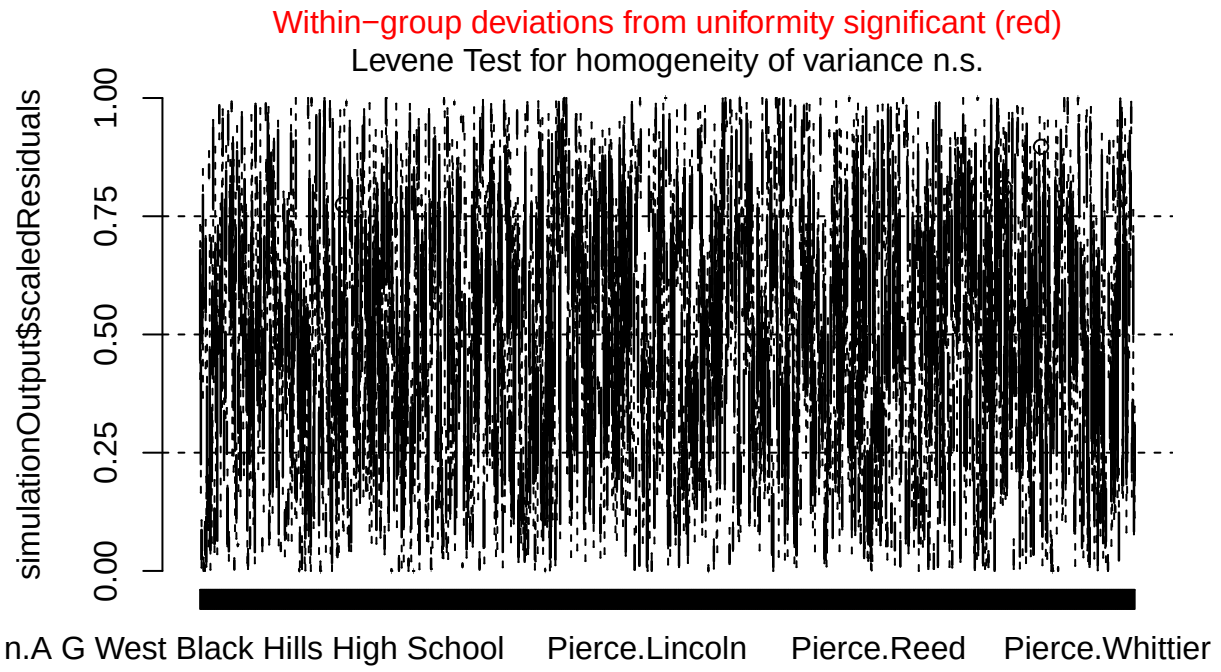


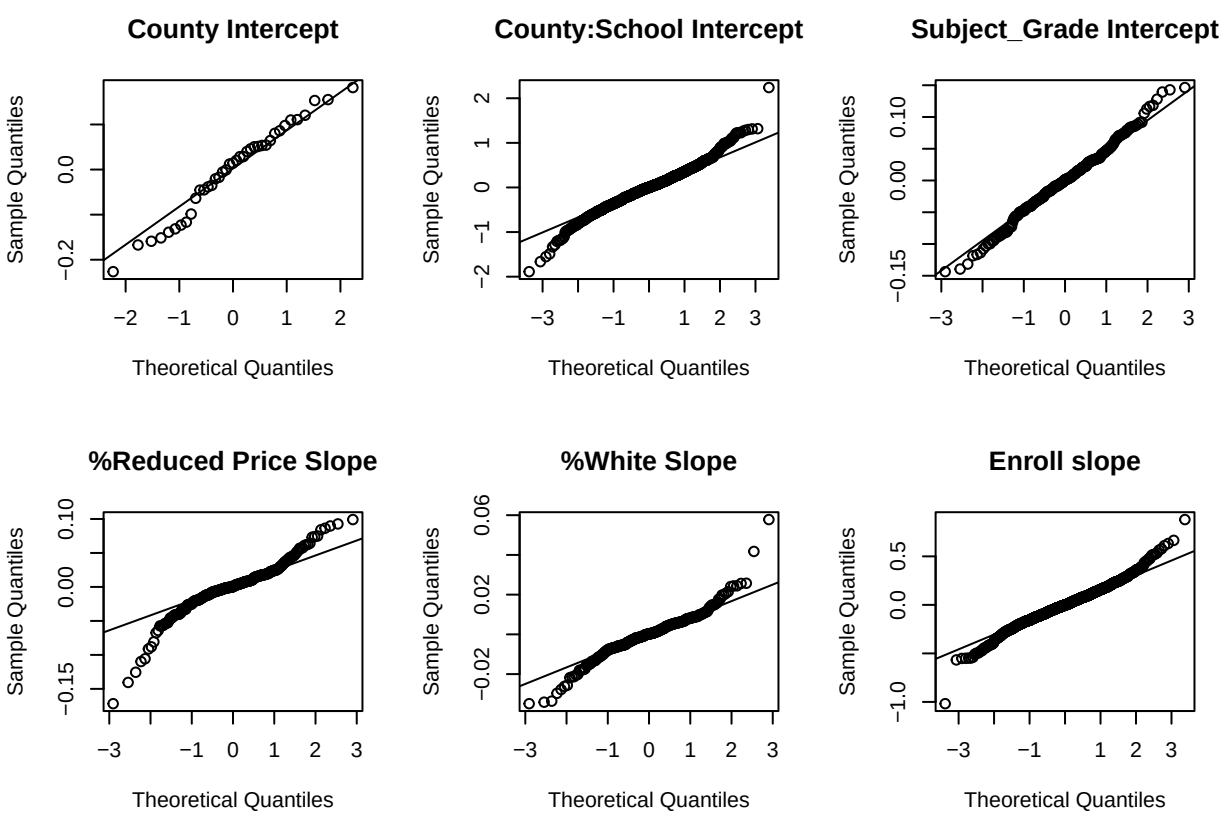

```
##
##  DHARMa bootstrapped outlier test
##
## data:  sim
## outliers at both margin(s) = 56, observations = 4525, p-value = 0.22
## alternative hypothesis: two.sided
##  percent confidence interval:
##  0.006618785 0.013491713
## sample estimates:
## outlier frequency (expected: 0.00959337016574586 )
##                                         0.01237569
```



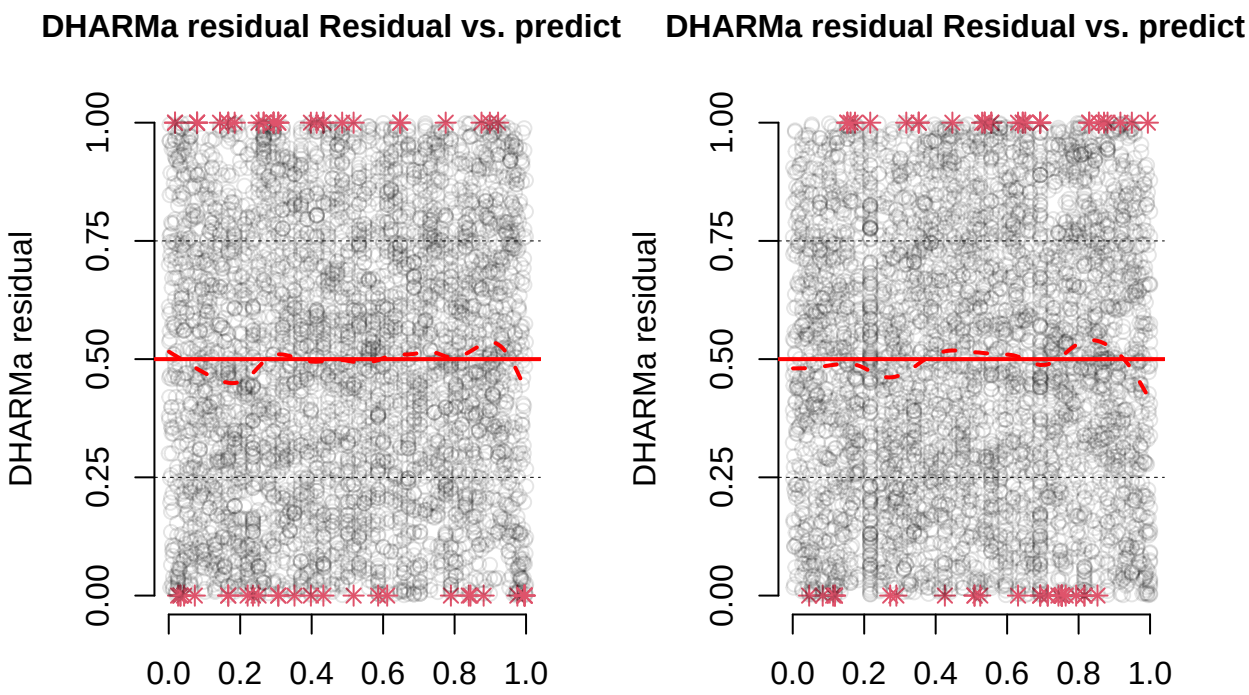

```
## Warning in ensurePredictor(simulationOutput, form): DHARMa:::ensurePredictor:
## character string was provided as predictor. DHARMa has converted to factor
## automatically. To remove this warning, please convert to factor before
## attempting to plot with DHARMa.
```







A whopping one point improvement in BIC! But at least we passed the big 3 DHARMa tests. Now we see if the last two predictors have anything to offer



vgYearsEducationalExperience (rank transfotTeachersWithAtLeastMasterDegree (rank t

It looks like they may be able to contribute a little bit. Worth a try.

```
## Data: wasl_num
## Models:
## mod_stroll_wslope: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPric
## mod_exp: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals_lo
## mod_mast: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade + FreeorReducedPricedMeals_l
##                   npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## mod_stroll_wslope   18 29172 29288 -14568    29136
## mod_exp             19 29174 29295 -14568    29136 0.8767  1     0.3491
## mod_mast            19 29173 29295 -14568    29135 0.3810  0
```

They were not able to contribute a little bit. Unfortunately, we are out of variables, and we are stuck with heteroscedasticity in our groups. Though of course that can happen, and shouldn't be shocking given the data we are working with (I think that if we could operate at the district level instead of county level, we would be able to achieve a much better fit).

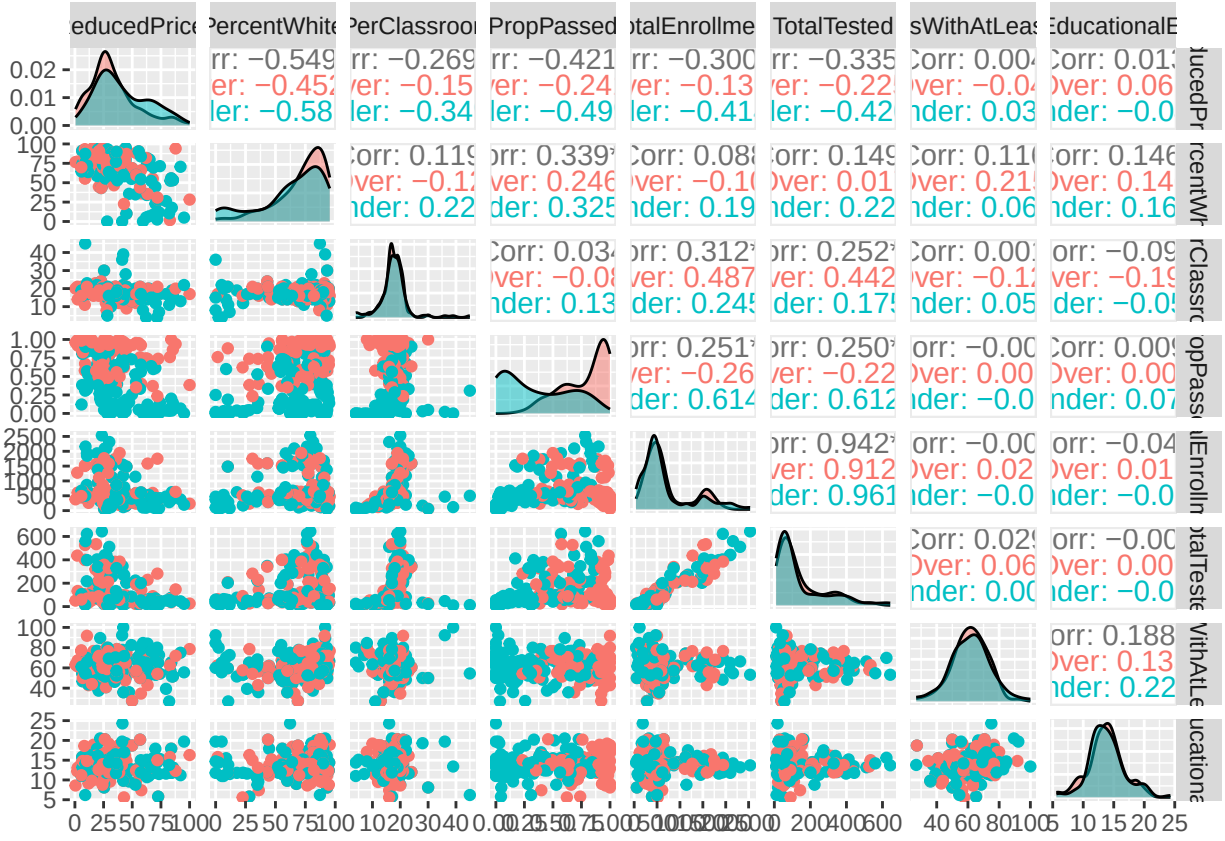
```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cbind(MetStandard, TotalTested - MetStandard) ~ Subject_Grade +
##     FreeorReducedPricedMeals_logit + PercentWhite_logit + stratio +
##     enroll + (FreeorReducedPricedMeals_logit + PercentWhite_logit ||
##     County:Subject_Grade) + (enroll | County:School) + (1 | County)
##    Data: wasl_num
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##  29172.3  29287.8 -14568.2  29136.3     4507
##
## Scaled residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4.0001 -0.6482  0.0150  0.6349  3.5184
##
## Random effects:
##  Groups                 Name                            Variance  Std.Dev. Corr
##  County.School          (Intercept)                     0.2008419 0.44815
##                         enroll                          0.0507576 0.22529  -0.82
##  County.Subject_Grade   PercentWhite_logit              0.0009638 0.03105
##  County.Subject_Grade.1 FreeorReducedPricedMeals_logit 0.0051545 0.07179
##  County.Subject_Grade.2 (Intercept)                     0.0084748 0.09206
##  County                 (Intercept)                     0.0198752 0.14098
## Number of obs: 4525, groups:
## County:School, 1369; County:Subject_Grade, 273; County, 39
##
## Fixed effects:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     1.09213    0.04203  25.986  < 2e-16 ***
## Subject_GradeMath.5            -1.06444    0.03133 -33.977  < 2e-16 ***
## Subject_GradeScience.5         -1.94799    0.03178 -61.306  < 2e-16 ***
## Subject_GradeReading.10         0.25646    0.04975   5.155 2.53e-07 ***
## Subject_GradeWriting.10         0.05458    0.04933   1.106    0.269
## Subject_GradeMath.10           -1.47354    0.04869 -30.263  < 2e-16 ***
## Subject_GradeScience.10        -2.17296    0.04896 -44.383  < 2e-16 ***
## FreeorReducedPricedMeals_logit -0.39934    0.02035 -19.620  < 2e-16 ***
## PercentWhite_logit              0.07999    0.01493   5.359 8.36e-08 ***
## stratio                        -0.09172    0.02089  -4.391 1.13e-05 ***
## enroll                          0.08131    0.01764   4.608 4.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) S_GM.5 S_GS.5 S_GR.1 S_GW.1 S_GM.1 S_GS.1 FrRPM_ PrcnW_
## Sbjct_GrM.5 -0.387
## Sbjct_GrS.5 -0.384  0.517
## Sbjct_GR.10 -0.473  0.324  0.324
## Sbjct_GW.10 -0.476  0.328  0.329  0.755
## Sbjct_GM.10 -0.484  0.332  0.331  0.769  0.772
## Sbjct_GS.10 -0.480  0.332  0.331  0.761  0.766  0.780
## FrrRdcdPrM_ -0.057  0.006  0.014  0.093  0.097  0.095  0.099
## PrcntWht_lg -0.234 -0.001 -0.005 -0.010 -0.013 -0.014 -0.014  0.455
## stratio     -0.005  0.001  0.004 -0.011 -0.009 -0.008 -0.006  0.203  0.228
## enroll       0.158 -0.002 -0.001 -0.382 -0.388 -0.391 -0.392  0.121  0.017
##             strati
## Sbjct_GrM.5
## Sbjct_GrS.5
## Sbjct_GR.10
## Sbjct_GW.10
## Sbjct_GM.10
## Sbjct_GS.10
## FrrRdcdPrM_
## PrcntWht_lg
## stratio
## enroll      -0.172
```

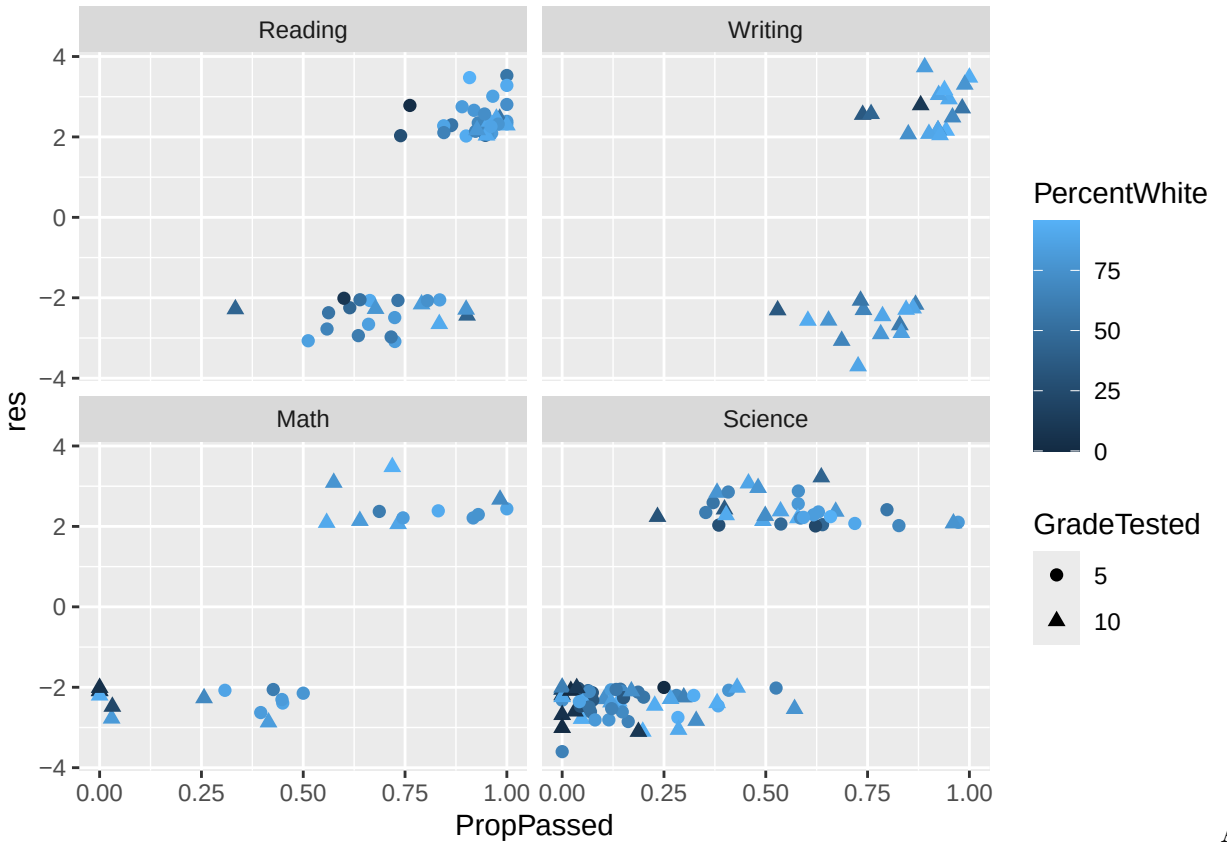
The County:School intercept does most of the heavy lifting for the random effects in our model, and is actually very correlated with its random slope `enroll`. This is rather concerning, but I think my rationale for including that term is sufficient to keep it as is. The County intercept accounts for the next-most variability. This is rather unsurprising; recall the massive jumps in BIC we saw in the initial models. We see a similar phenomenon in the fixed effects, with the various Subject_Grade combinations almost all having higher estimates than the remaining fixed effects. Back to the random effects though, the %white slope term is not doing much work in this model. Though it improved our BIC (by 1), this is probably the first term I would consider removing. However, the miniscule improvements it gave us also led to lower residual variance and fewer outliers.

Outlier Analysis

Now let's examine the outliers to find our over- and underperforming schools, and perhaps any patterns they may follow.



There is not much of a visibly discernible difference between any of ther variables for the outliers (except for `PropPassed`, which is to be expected). Let's see if there's anything at the subject level.



All that we have here is that math and writing yield the fewest outliers, reading has the most on the higher end, and science on the lower end. This is ostensibly the same as what we saw with `prop` above, given the trends in scores for each subject (and the sample size in the case of writing). Let's look at which schools had the highest over- and lowest underperformances.

School	County	Subject	Grade	Total Tested	Fitted	Observed	Residual
Kelso High School	Cowlitz	Writing	10	383	0.82	0.89	3.74
North Elementary	Grant	Reading	5	37	0.84	1.00	3.53
Kalama Jr Sr High	Cowlitz	Math	10	96	0.54	0.72	3.48
TEAM High School	Cowlitz	Writing	10	28	0.80	1.00	3.48
Hockinson Heights Intermediate	Clark	Reading	5	153	0.81	0.91	3.48
Skyline High School	King	Writing	10	436	0.96	0.99	3.30
Mullan Road Elementary	Spokane	Reading	5	68	0.92	1.00	3.28
Garfield High School	King	Science	10	341	0.56	0.64	3.22
Deer Park High School	Spokane	Writing	10	147	0.86	0.94	3.17
Gov John Rogers High School	Pierce	Math	10	534	0.51	0.57	3.09

School	County	Subject	Grade	Total Tested	Fitted	Observed	Residual
Woodland High School	Cowlitz	Writing	10	161	0.84	0.73	-3.70
Valley View Elementary	Yakima	Science	5	55	0.12	0.00	-3.60
Franklin High School	King	Science	10	203	0.30	0.19	-3.11
Kalama Jr Sr High	Cowlitz	Science	10	96	0.34	0.20	-3.10
Shadow Lake Elementary	King	Reading	5	80	0.86	0.72	-3.09
Salmon Creek Elementary	Clark	Reading	5	82	0.67	0.51	-3.07

School	County	Subject	Grade	Total Tested	Fitted	Observed	Residual
Wenatchee High School	Chelan	Writing	10	440	0.75	0.69	-3.07
West Valley High School	Yakima	Science	10	329	0.37	0.29	-3.05
Spokane Valley High School	Spokane	Science	10	31	0.14	0.00	-3.01
Selah Intermediate	Yakima	Reading	5	271	0.79	0.72	-2.97

We notice that the extreme outliers are usually either cases in which the model was unable to predict extremely low scores (close to 0 or 1), or schools with huge samples that the model missed by non-trivial amounts (or Kalama Jr Sr High, which for some reason our model just predicted way off on what were pretty moderate outomes)

```
## [1] 167
```

Interestingly, no schools were outliers for two different grade levels, and Medical Lake High School's 10th grade class was the only outlier in more than two subjects (Reading+, Writing+, Science-). Even more interestingly, although there were 33 schools who were outliers for two subjects, only two schools were outliers of the same type (i.e. overperformed in all subjects or underperformed in all subjects).

```
## # A tibble: 4 x 7
##   County   School                    Subject TotalTested   fit PropPassed   res
##   <chr>    <chr>                     <fct>         <dbl> <dbl>      <dbl> <dbl>
## 1 Franklin New Horizons High School  Math             32  0.17       0.03 -2.48
## 2 Franklin New Horizons High School  Science          22  0.09       0    -2.06
## 3 King     Newport Senior High School Reading        400  0.96       0.98  2.46
## 4 King     Newport Senior High School Writing        399  0.96       0.98  2.71
```

Though having multiple outlier scores is impressive, it isn't necessarily the best indication of overall unexpected performance. Let's find the schools who had the highest and lowest average residuals.

School	Grade	County	Mean Residual
Lowell Elementary School	5	King	1.760808
Clallam Bay High & Elementary	10	Clallam	1.497049
Libby Center	5	Spokane	1.290211
International School	10	King	1.280006
Trout Lake School	10	Klickitat	1.102496

School	Grade	County	Mean Residual
Echo Glen	10	King	-1.2768579
Trout Lake School	5	Klickitat	-1.0734804
Barker Center	10	Spokane	-1.0371864
Valley View Elementary	5	Yakima	-0.9607603
Curlew Elem & High School	5	Ferry	-0.9509335

We see from this that Lowell Elementary School is a cut above the rest, and Clallam Bay High School (funnily enough their 5th grade class slightly underperformed in all subjects) is a cut above the rest of the rest! Both noticeably outperform the rest of the schools at 'defeating' our model. However, it is reassuring to see that no schools were particularly close to being outliers in all subjects on average, and the 'extreme' averages on the negative side are especially nice to see. I think that it's just because even the most severely underperforming schools still have pass rates at or above 1/6 in reading and writing, which helps the model because it has difficulty predicting too close to 0% or 100%. I suppose it is also reassuring to me as an aspiring educator to see that the model has more trouble at the higher end because pass rates tend to be closer to 100% more

often than they are to 0%.

Conclusions/Further Work

I ran a model selection process based on working my way 'down' the hierarchy of the data, starting with county, subject, school, and proceeding from there. I used a combination of intuition and my previous visual exploration of the data to come up with models to test, and evaluated them based on BIC and intuitive plausibility. I then examined model assumptions using DHARMa to generate plots and conduct various tests of the model's fit/residual variance at different levels.

I found demographic factors (%white, %reduced cost meals) able to explain more variance in performance compared to school-resource-factors (enrollment, student:teacher ratio). Most interestingly, I found that schools with higher proportions of teacher's with master's degrees and more experienced teachers did not see any meaningful improvements in performance. The latter makes sense to some extent: there has always been the stereotype of the old teacher stuck in their ways who students accustomed to more modern methods of teaching have trouble learning from. It's a little saddening to see that borne out in the model, but I also shouldn't give my model too much credit here. These conclusions are certainly not... conclusive. From the very beginning, I made a decision to generally remove any rows with `NA` entries, except for those I felt certain I could discern the true value of. 70 or so rows may not be enough to change too much, but it would be interesting to try doing some kind of imputation to see if we could get decent estimates for some of those missing values. Or perhaps we could see if the data is online somewhere (I actually did this, but I didn't look very hard).

A second, and more consequential, shortcoming of my procedure came in my model selection. I think I was very lenient in terms of allowing myself to induce further variance within my grouping factors, writing it off as something that was unavoidable given the nature of the data. Not that I don't believe that to be true, but I also believe that there were ways it could have been mitigated. Perhaps not being so quick to add the School:County intercept, or the grouping of grade and subject. I could have chosen to completely disregard the writing test; having data on three subjects across two grade levels could have helped a lot (this was my initial approach but I convinced myself that it would not be received well for whatever reason).

A further issue with my model selection was my approach to variable selection. I think that the order I went in made sense and was totally reasonable, but I worry that I had too much of a one-track mind with exclusively using BIC to choose how to proceed. I don't think BIC is or was a bad metric to use, and my approach may turn out to have been the best one. But I think I should have done some further investigation into the merits of other approaches to model selection. In fact, I literally just finished a project on random jump Markov chain Monte Carlo, which is a tool that might have been great for this (probably too slow though). That's something I may look into over winter break.