# School Test Performance

Christopher Pearson

May 19, 2020

```
library(readr)
library(lme4)
library(dplyr)
library(tidyr)
library(tidyselect)
library(stringr)
library(ggplot2)
library(compositions)
```

## BUSINESS UNDERSTANDING

It is sometimes said that the three rules of business are location, location, location; but the same might be said for schooling, and employees with children. One of the greatest challenges for parents is to know where to live to get their children into a good school. For businesses, this may present itself as a challenge for retaining employees. Being in a county with poor-performing schools could make it more difficult for businesses to retain employees. On the other hand, having access to quality and affordable schools could make it easier to recruit top talent.

As we are focused on school quality, our metric is performance on statewide tests, and finding the variables that leads to schools performing better. To avoid the additional difficulty of finding State effects, we will keep this study focused on Washington State and the Washington Assessment of Student Learning (WASL), a statewide test used between 1997 and 2009.

As the effects of location are key, we will be aiming to use a hierarchic model with county as a grouping variable.

## DATA UNDERSTANDING

The original data-set has 17 variables. Three of these identify the State, School, and Grade taking the exam (10th or 5th). Another eight identify how many students met state requirements, and how many students took the exam at each school. Then there is the school's total enrollment. The last six variables are metrics that may help explain why the school performs well: the percent of students whom are white, the percent of students on food assistance, the percent of teachers with an MA degree, mean teacher experience, and teachers per student.

There are a number of NAs throughout the dataset. Most of these are with regard to test performance. As this is our target variable, it would be best to drop rather than estimate the missing data. On the other hand, some of these are related to the school metrics we use to measure classroom performance. It is only 66 columns, only one of which has NAs in four of the potential independent variables. Rather than risk good data, we will simply drop the one column with four NAs, and replace the rest with mean-values. This will not only offer a fair approximation, but minimize their leverage should they be too far off.

```r
# Preliminary Steps
#Get and clean the data
wasl<-read_csv("wasl.csv", skip = 1,
               col_types = 'ccnnnnnnnnnnnnnnn',
               col_names = c('COUNTY','SCHOOL',"GRADE","MATH_MET","READING_MET","WRITING_MET","SCIENCE_M
                             "MATH_TOOK","READING_TOOK","WRITING_TOOK","SCIENCE_TOOK","TOTAL_ENROLL",
                             "P_WHITE","STU_PER_TEACHER","FOOD_ASSIST","MEAN_TEACHER_EXP","MEAN_MA_DEGRE
  summary(wasl)
```

```
##     COUNTY             SCHOOL              GRADE          MATH_MET
##  Length:1481        Length:1481        Min.   : 5.000   Min.   :  0.00
##  Class :character   Class :character   1st Qu.: 5.000   1st Qu.: 19.00
##  Mode  :character   Mode  :character   Median : 5.000   Median : 35.00
##                                        Mean   : 6.404   Mean   : 52.69
##                                        3rd Qu.:10.000   3rd Qu.: 57.00
##                                        Max.   :10.000   Max.   :496.00
##                                                         NA's   :8
##   READING_MET      WRITING_MET      SCIENCE_MET       MATH_TOOK
##  Min.   :  3.00   Min.   :  2.00   Min.   :  0.00   Min.   :  6.00
##  1st Qu.: 32.00   1st Qu.: 20.75   1st Qu.: 10.00   1st Qu.: 43.00
##  Median : 50.50   Median : 79.00   Median : 22.50   Median : 67.00
##  Mean   : 77.92   Mean   :140.52   Mean   : 35.21   Mean   : 94.67
##  3rd Qu.: 76.00   3rd Qu.:250.25   3rd Qu.: 39.25   3rd Qu.: 95.00
##  Max.   :698.00   Max.   :670.00   Max.   :355.00   Max.   :785.00
##  NA's   :9        NA's   :1069     NA's   :21
##   READING_TOOK     WRITING_TOOK     SCIENCE_TOOK     TOTAL_ENROLL
##  Min.   :  7.00   Min.   :  0.00   Min.   :  5.00   Min.   :   0.0
##  1st Qu.: 43.00   1st Qu.:  0.00   1st Qu.: 43.00   1st Qu.: 307.0
##  Median : 67.00   Median :  0.00   Median : 68.00   Median : 439.0
##  Mean   : 95.28   Mean   : 46.66   Mean   : 93.47   Mean   : 522.3
##  3rd Qu.: 95.00   3rd Qu.: 16.00   3rd Qu.: 95.00   3rd Qu.: 568.0
##  Max.   :803.00   Max.   :792.00   Max.   :751.00   Max.   :2997.0
##
##     P_WHITE       STU_PER_TEACHER   FOOD_ASSIST     MEAN_TEACHER_EXP
##  Min.   :  0.00   Min.   :  4.00   Min.   :  0.00   Min.   : 2.20
##  1st Qu.: 55.93   1st Qu.: 15.00   1st Qu.: 22.06   1st Qu.:11.80
##  Median : 74.19   Median : 17.00   Median : 35.96   Median :13.70
##  Mean   : 66.58   Mean   : 17.21   Mean   : 39.46   Mean   :13.72
##  3rd Qu.: 85.76   3rd Qu.: 19.00   3rd Qu.: 54.23   3rd Qu.:15.60
##  Max.   :100.00   Max.   :166.00   Max.   :100.00   Max.   :31.30
##                   NA's   :53       NA's   :9        NA's   :53
##  MEAN_MA_DEGREE
##  Min.   : 16.70
##  1st Qu.: 51.45
##  Median : 60.90
##  Mean   : 59.92
##  3rd Qu.: 68.80
##  Max.   :100.00
##  NA's   :58
```

```r
wasl %>%
  filter( (is.na(STU_PER_TEACHER) | is.na(FOOD_ASSIST) | is.na(MEAN_TEACHER_EXP) | is.na(MEAN_MA_DEGREE)
  summarize(`Possible Independent Variables with NA Values` = length(FOOD_ASSIST))
```

```
## # A tibble: 1 x 1
##   `Possible Independent Variables with NA Values`
##                                             <int>
## 1                                              66
```

```
replace_na_mean <- function(vec){
  if(is.character(vec) ) return(vec)
  return(if_else(is.na(vec), mean(vec, na.rm = TRUE), vec) )
}

load_and_clean <- function() {
    read_csv("wasl.csv", skip = 1,
                col_types = 'ccnnnnnnnnnnnnnnn',
                col_names = c('COUNTY','SCHOOL',"GRADE","MATH_MET","READING_MET","WRITING_MET","SCIENCE_
                              "MATH_TOOK","READING_TOOK","WRITING_TOOK","SCIENCE_TOOK","TOTAL_ENROLL",
                              "P_WHITE","STU_PER_TEACHER","FOOD_ASSIST","MEAN_TEACHER_EXP","MEAN_MA_DEGR
    mutate(P_WHITE = P_WHITE / 100, FOOD_ASSIST = FOOD_ASSIST/100, FOOD_ASSIST = FOOD_ASSIST / 100,
           LOG_T_EXP = log(MEAN_TEACHER_EXP), LOG_ENROLL = log(TOTAL_ENROLL) ) %>%
    group_by(COUNTY, SCHOOL) %>%  # essentially group by row
    filter( TOTAL_ENROLL >= max(MATH_MET,READING_MET,WRITING_MET,SCIENCE_MET,MATH_TOOK,READING_TOOK,WRI
                                na.rm = TRUE) ) %>%
    ungroup() %>%
    mutate(ONES = 1, ID = cumsum(ONES) ) %>%
    select( - ONES ) %>%
    select(ID, everything() ) %>%
    mutate(TEACHERS_PER_STU = STU_PER_TEACHER^-1) %>%
    mutate(LOG_STU_PER_TEACHER = log(STU_PER_TEACHER) ) %>%
    mutate( TEMP_MATH = MATH_MET + MATH_TOOK / 10000, TEMP_READING = READING_MET + READING_TOOK / 10000
                TEMP_WRITING = WRITING_MET + WRITING_TOOK / 10000, TEMP_SCIENCE = SCIENCE_MET + SCIEN
    select( - MATH_MET, - MATH_TOOK,
            - READING_MET, - READING_TOOK,
            - WRITING_MET, - WRITING_TOOK,
            - SCIENCE_MET, - SCIENCE_TOOK ) %>%
    pivot_longer( cols =starts_with('TEMP'), names_to = 'TEST', values_to = 'TEMP' ) %>%  # perhaps cre
    mutate(TEST = stringr::str_to_lower(substring(TEST, 6) ),
           PASSED = trunc(TEMP),
           TOOK = (TEMP %% 1) * 10000,
           SHARE_PASSED = PASSED / TOOK,
           SHARE_TOOK  = TOOK / TOTAL_ENROLL) %>%
    select( - TEMP ) %>%
    filter(!is.na(TOOK) & !(is.na(PASSED)) ) %>%
    filter( !(is.na(STU_PER_TEACHER) & is.na(FOOD_ASSIST) & is.na(MEAN_TEACHER_EXP) & is.na(MEAN_MA_DEG
    mutate_all(replace_na_mean) %>%
    return()
}
```

```
wasl <- load_and_clean()
```

## Quick Answers

For those of you interested in what schools performed best, it it is interesting to note that half of these are in King County. So if a company hopes to locate to take advantage of the best schools, most of them are in the Seattle Area. Moreover, the King County schools also have higher enrollments. However, the quality of some of the schools in King County is counterbalanced by much higher demand for placement in those schools andmuch higher property values. Clallam County to the West of Seattle and Klickitat County in

South Central Seattle both offer excellent schools for those looking ofr an out-of-the-way place.

A list of the worst schools by exam is somewhat more difficult though, as stacking at 0 means there are 43 of them rather than 7 (5th grade does not take the writing exam). While King County is well represented on this list, the overall list is much more geographically diverse than the list of best schools.

```
wasl %>%
  group_by(TEST, GRADE) %>%
#  mutate(SHARE_PASSED = PASSED / TOOK ) %>%
  filter(SHARE_PASSED == max(SHARE_PASSED) ) %>%
  ungroup() %>%
  mutate(SHARE_PASSED = round(SHARE_PASSED, 4) * 100,
         GRADE_TEST = paste('grade',GRADE,TEST,'test') ) %>%
  arrange(GRADE_TEST) %>%
  select(COUNTY, SCHOOL, GRADE_TEST, SHARE_PASSED, TOTAL_ENROLL)
```

```
## # A tibble: 12 x 5
##    COUNTY    SCHOOL                  GRADE_TEST        SHARE_PASSED TOTAL_ENROLL
##    <chr>     <chr>                   <chr>                    <dbl>        <dbl>
##  1 King      International School    grade 10 math te~         98.3          492
##  2 Clallam   Clallam Bay High & Ele~ grade 10 reading~        100            189
##  3 Grant     Wilson Creek High       grade 10 reading~        100             69
##  4 King      Federal Way Public Aca~ grade 10 reading~        100            308
##  5 Klickitat Trout Lake School       grade 10 reading~        100            115
##  6 King      International Communit~ grade 10 science~         96.1          380
##  7 Clallam   Clallam Bay High & Ele~ grade 10 writing~        100            189
##  8 King      Federal Way Public Aca~ grade 10 writing~        100            308
##  9 Klickitat Trout Lake School       grade 10 writing~        100            115
## 10 Spokane   Libby Center            grade 5 math test        100            150
## 11 King      Family Learning Center  grade 5 reading ~        100            267
## 12 King      Lowell Elementary Scho~ grade 5 science ~         97.2          267
```

```
wasl %>%
  filter(TOOK > 0) %>%
  group_by(TEST, GRADE) %>%
  filter(SHARE_PASSED == min(SHARE_PASSED) ) %>%
  ungroup() %>%
  mutate(SHARE_PASSED = round(SHARE_PASSED, 4) * 100,
       GRADE_TEST = paste('grade',GRADE,TEST,'test') ) %>%
  arrange(GRADE_TEST) %>%
  select(COUNTY, SCHOOL, GRADE_TEST, SHARE_PASSED, TOTAL_ENROLL)
```

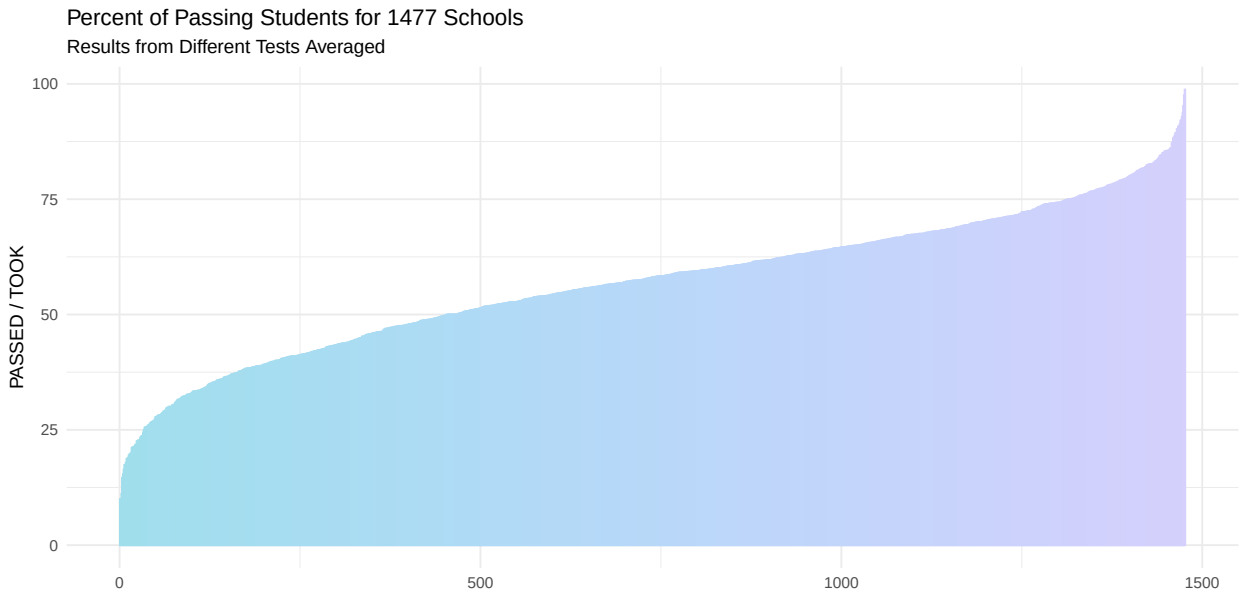
```
## # A tibble: 43 x 5
##    COUNTY    SCHOOL                  GRADE_TEST         SHARE_PASSED TOTAL_ENROLL
##    <chr>     <chr>                   <chr>                     <dbl>        <dbl>
##  1 King      Interagency Programs    grade 10 math test            0          514
##  2 Kitsap    Spectrum Community Sc~ grade 10 math test            0          105
##  3 Snohomish Heritage School         grade 10 math test            0           40
##  4 Spokane   Barker Center           grade 10 math test            0           70
##  5 Yakima    Stanton Alternative S~ grade 10 math test            0          668
##  6 Yakima    Compass High School     grade 10 math test            0           92
##  7 Yakima    Pride High School       grade 10 math test            0          121
##  8 Yakima    Eagle High School       grade 10 math test            0          199
##  9 King      New Start               grade 10 math test            0           67
## 10 Yakima    Compass High School     grade 10 reading ~           18.8          92
## # ... with 33 more rows
```

# DATA PREPARATION

## Data Cleaning and Wrangling

In its present form, the data is rather difficult to work with–so one of our primary goals is to create a pivot-table that has the test-type in one column, and data on how many students passed and took it in two others. Then we can handle NAs as we dicussed in the Data Understanding section.

```
wasl %>%
  group_by(ID) %>%
  summarize( MEAN_PASSED = sum(PASSED) / sum(TOOK) * 100 ) %>%
  ungroup() %>%
  arrange(MEAN_PASSED) %>%
  mutate(ONES = 1, ID = cumsum(ONES) ) %>%
  select( - ONES) %>%
  ggplot( aes(x = ID, y = MEAN_PASSED ) ) +
  geom_bar(stat = 'identity', col = hcl(210 + 60/1476 * 1:1476 ), fill = hcl(210 + 60/1476 * 1:1476 ) )
  xlab('') + ylab('PASSED / TOOK') + theme_minimal() +
  ggtitle('Percent of Passing Students for 1477 Schools', subtitle = 'Results from Different Tests Avera
```



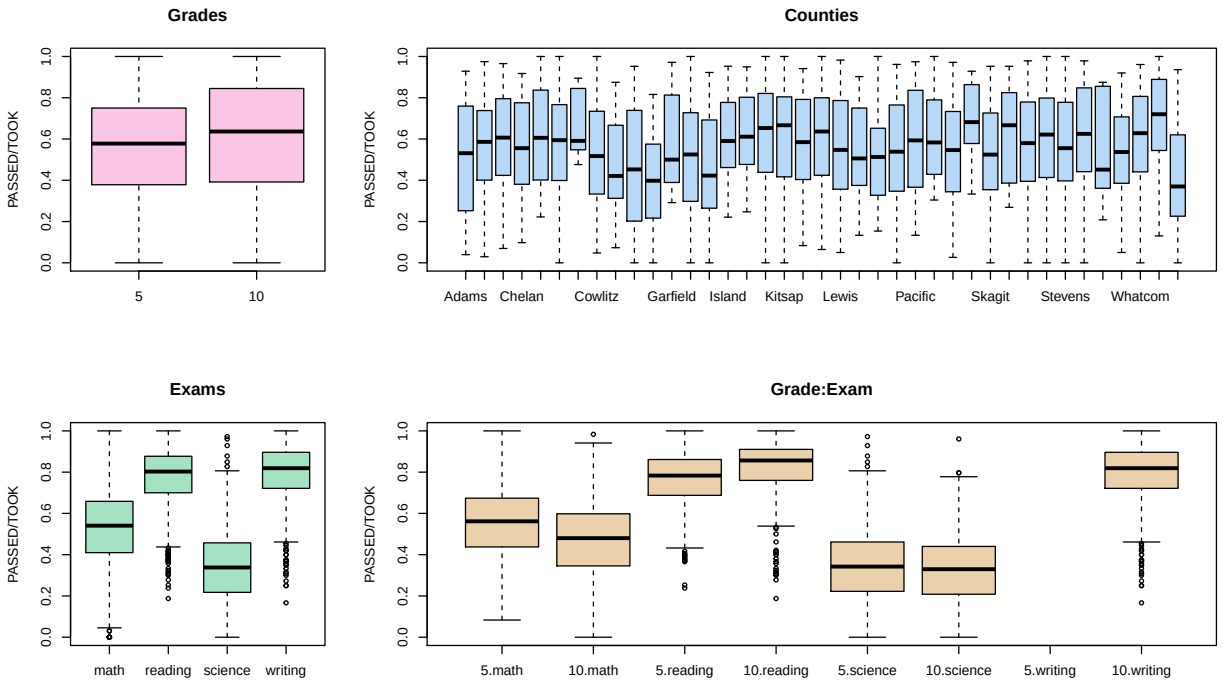

### Variable Selection

Most of these data need to be adjusted by the total enrollment of the school. For the target, we either want to work with passed / took, or logratio transformations of passed / enrolled, failed / enrolled, and untried / enrolled (more on this latter option later). This would put the target variables in a much more similar range to most of our features, all but one of which is in the 0 to 1 range. Mean teaching experience on the other hand ranges from 2.2 to 31.3, so a log transformation would be appropriate to account for diminishing returns of additional years of experience (2 vs 4 years of experience should mean more than 28 vs 30 years of experience), and bring its range more in line with the other variables.

We could also include variables for grade and the type of test. On the one hand, there are clear differences in performance between exam types this could greatly improve the accuracy of our predictions. On the other hand, including these variables does little to nothing to help us determine what a good school is. One cannot choose between Math and Reading the same way one chooses between schools; and whether a child is in 5th or 10th grade is a matter of age rather than choice of schools. However, this data is clearly important to

predict school performance, so we will try constructing both a "stoic model" that only takes into account variables we have more control of, and an "intuitive model" that includes the data on tests and grade.

```
layout(matrix(c(1,1,2,2,2,2,2,3,3,4,4,4,4,4), 2, 7, byrow = TRUE))

boxplot(wasl$PASSED/wasl$TOOK~wasl$GRADE,main="Grades", xlab = NULL, ylab = 'PASSED/TOOK', col = hcl(h =
boxplot(wasl$PASSED/wasl$TOOK~wasl$COUNTY,main="Counties", xlab = NULL, ylab = 'PASSED/TOOK', col = hcl
boxplot(wasl$PASSED/wasl$TOOK~wasl$TEST,main="Exams", xlab = NULL, ylab = 'PASSED/TOOK', col = hcl(h = 
boxplot(wasl$PASSED/wasl$TOOK~wasl$GRADE + wasl$TEST,main="Grade:Exam", xlab = NULL, ylab = 'PASSED/TOO
```



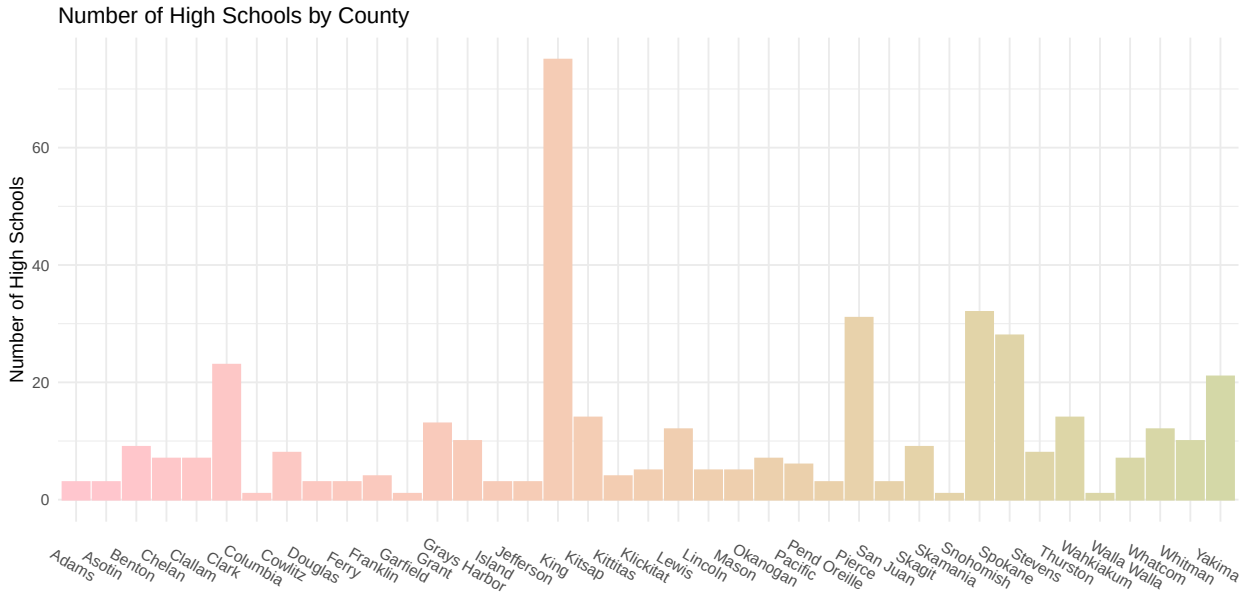

### Grouping Variable Selection

For the grouping variables, we could try using schools, grades or counties. With hierarchic models though, the number of variables quickly rises: including a random slow and intercept for one with respect to schools for one variable is the equivalent of adding close to 3,000 additional variables. Schools thus make a bad grouping variable here. Grade would not be bad, but as predictive as it could be, we are not interested in how 5th graders vs 10th graders perform. So instead, it makes the most sense to look at county-level data.

Looking at the number of high schools per county, we can see that most counties have fewer than 10, a substantial number have 11-35, and only King County, with over 70 high schools, has anymore. This is a good grouping variable, though it may make some sense to come back later and subdivide King County into 7 or so sub-counties.

```
wasl %>%
  filter(GRADE == 10) %>%
  select(COUNTY, SCHOOL) %>%
  distinct() %>%
  group_by(COUNTY) %>%
  summarize( NUMBER_OF_SCHOOLS = length(COUNTY) ) %>%
  ungroup() %>%
  ggplot( aes(x = COUNTY, y = NUMBER_OF_SCHOOLS ) ) + geom_bar(stat = 'identity', col = hcl(90 / 39 * 1
```

```
  xlab('') + ylab('Number of High Schools') + ggtitle('Number of High Schools by County') + theme_minima
```

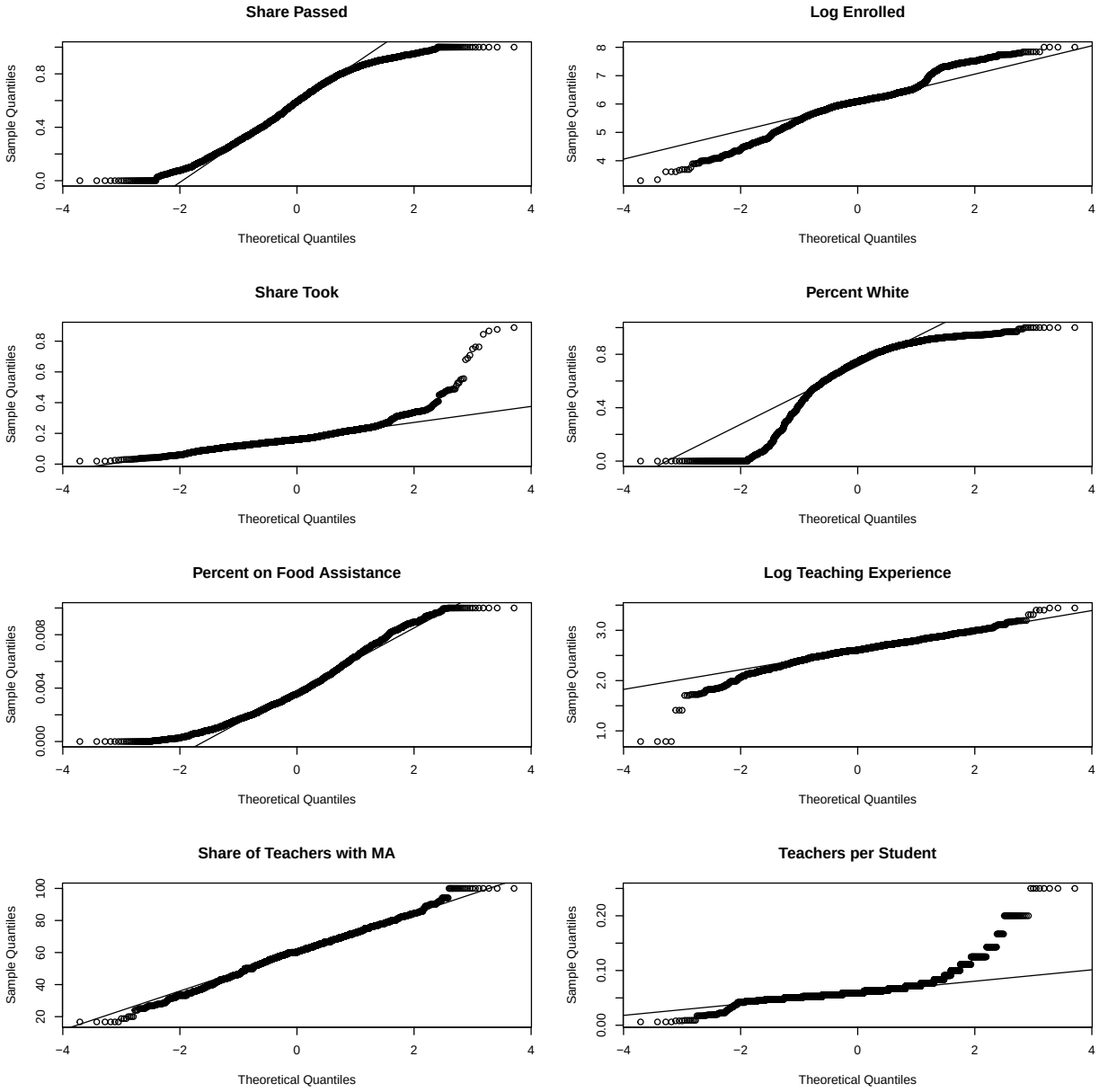



```
#  scale_x_discrete(labels = NULL)
```

## Explore Data: QQ-Norms

Turning to qqnorm plots of non-categorical variables, about half of these seem to be normally distributed with some issues near the tails. The measure of students who met the standard is tail-heavy, suggesting the variable is pressing up against the insurmountable barriers of 100% and 0% at some schools. Log enrolled is a little thin in the tails; and the share of students taking the exam is a little skewed toward few students taking the test. The percent of students whom are white also is skewed toward whiteness–not surprising when one takes more rural areas into account. Food assistance looks tail-heavy, suggesting some class segregation. The log of teaching experience and share of teachers with MAs look about right though. Finally, teachers per student is a little thin in the tails.

```
# don't want the shared y-axis form the facet_wrap
layout(matrix(1:8, 4, 2, byrow = TRUE))

qqnorm(wasl$SHARE_PASSED,main="Share Passed");qqline(wasl$SHARE_PASSED);
qqnorm(wasl$LOG_ENROLL,main="Log Enrolled");qqline(wasl$LOG_ENROLL);
qqnorm(wasl$SHARE_TOOK,main="Share Took");qqline(wasl$SHARE_TOOK);
qqnorm(wasl$P_WHITE,main="Percent White");qqline(wasl$P_WHITE);
qqnorm(wasl$FOOD_ASSIST,main="Percent on Food Assistance");qqline(wasl$FOOD_ASSIST);
qqnorm(wasl$LOG_T_EXP,main="Log Teaching Experience");qqline(wasl$LOG_T_EXP);
qqnorm(wasl$MEAN_MA_DEGREE,main="Share of Teachers with MA");qqline(wasl$MEAN_MA_DEGREE);
qqnorm(wasl$TEACHERS_PER_STU,main="Teachers per Student");qqline(wasl$TEACHERS_PER_STU)
```

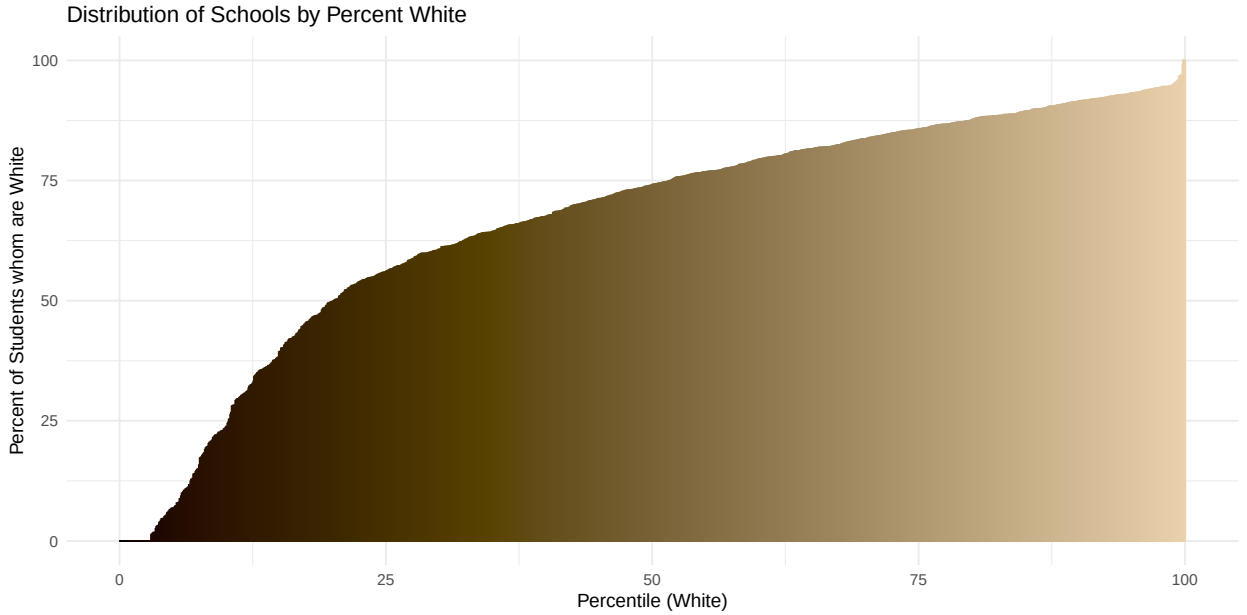



### Focus in on Percent White

Taking a closer look at the race distribution of the schools, the shape of the QQ-norm plot makes more sense. Only about 20% of schools are majority non-white, while the remaining 80% is majority white, with roughly 20% reaching 90% white. However, contrasting with this, there seem to be more schools that are 100% non-white than 100% white.

```
wasl %>%
  filter(TEST == 'math') %>%
  mutate(P_WHITE = P_WHITE * 100) %>%
  arrange(P_WHITE) %>%
  mutate(ONES = 1, PERCENTILE = cumsum(ONES) ) %>%
  mutate(PERCENTILE = PERCENTILE / max(PERCENTILE) * 100 ) %>%
  select( - ONES) %>%
```

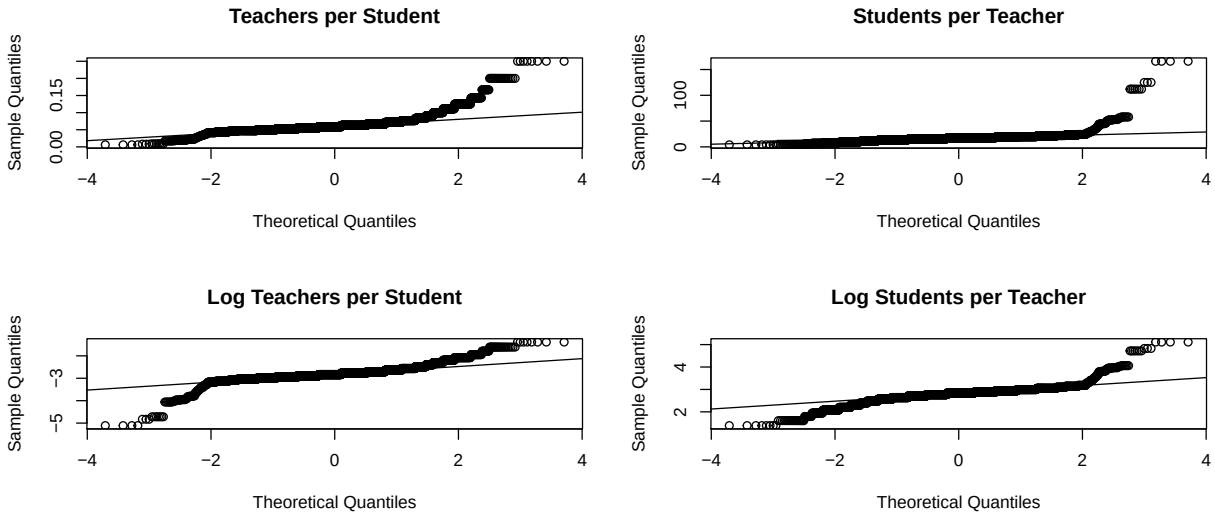
```
ggplot( aes(x = PERCENTILE, y = P_WHITE ) ) +
geom_bar(stat = 'identity', col = hcl(h = 60, l = 85/sum(wasl$TEST=='math') * 1:sum(wasl$TEST=='math'))
xlab('Percentile (White)') + ylab('Percent of Students whom are White') + ggtitle('Distribution of Sch
```




### Focus in on Teachers per Student

Turning to the teachers per student variable, there are a number of variations we could try: its inverse, its log, and its log inverse. Looking at these, we might get better results using its log inverse–log students_per_teacher, though any gains look like they would be slight to insignificant–so for now we will stick to using teachers per student.

```
layout(matrix(1:4, 2, 2, byrow = TRUE))
qqnorm(wasl$TEACHERS_PER_STU,main="Teachers per Student");qqline(wasl$TEACHERS_PER_STU)
qqnorm(wasl$STU_PER_TEACHER,main="Students per Teacher");qqline(wasl$STU_PER_TEACHER)
qqnorm(log(wasl$TEACHERS_PER_STU),main="Log Teachers per Student");qqline(log(wasl$TEACHERS_PER_STU))
qqnorm(log(wasl$STU_PER_TEACHER),main="Log Students per Teacher");qqline(log(wasl$STU_PER_TEACHER) )
```




# MODELING

To model the data, we are ultimately going to rely on using a hierarchic model using the variables we have available, but before we do that we will try a series of one-variable simple regressions using all of the independent variables available to determine what would be the best for this project. Then we can go on to the hierarchic model focusing on those variables that meet a P-Value threshold of 0.05.

```
white.fit.simple<-lm(SHARE_PASSED ~ P_WHITE, wasl)
food.fit.simple<- lm(SHARE_PASSED ~ FOOD_ASSIST, wasl)
exp.fit.simple<- lm(SHARE_PASSED ~ MEAN_TEACHER_EXP, wasl)
L.EXP.simple <- lm(SHARE_PASSED ~ LOG_T_EXP, wasl)
ma.fit.simple<- lm(SHARE_PASSED ~ MEAN_MA_DEGREE, wasl)
spt.fit.simple<- lm(SHARE_PASSED ~ LOG_STU_PER_TEACHER, wasl)
tps.fit.simple<- lm(SHARE_PASSED ~ TEACHERS_PER_STU, wasl)
sTook.fit.simple<- lm(SHARE_PASSED ~ SHARE_TOOK, wasl)
ssize.fit.simple<- lm(SHARE_PASSED ~ LOG_ENROLL, wasl)

CID.fit.simple<-lm(SHARE_PASSED~as.factor(COUNTY),data=wasl)
Grade.fit.simple<-lm(SHARE_PASSED~as.factor(GRADE),data=wasl)
SID.fit.simple<-lm(SHARE_PASSED~as.factor(SCHOOL),data=wasl)
Exam.fit.simple<-lm(SHARE_PASSED~as.factor(TEST),data=wasl)

comparison<-rbind(data.frame("Var"="P.White",
                             "Coefficient"=white.fit.simple$coefficients[2],
                             "P-Val"=anova(white.fit.simple)$`Pr(>F)`[1]),
                  data.frame("Var"="FoodAssist",
                             "Coefficient"=food.fit.simple$coefficients[2],
                             "P-Val"=anova(food.fit.simple)$`Pr(>F)`[1]),
                  data.frame("Var"="Experience",
                             "Coefficient"=exp.fit.simple$coefficients[2],
                             "P-Val"=anova(exp.fit.simple)$`Pr(>F)`[1]),
                  data.frame("Var"="Log Experience",
                             "Coefficient"=L.EXP.simple$coefficients[2],
                             "P-Val"=anova(L.EXP.simple)$`Pr(>F)`[1]),
                  data.frame("Var"="Hold MA",
```

```
                            "Coefficient"=ma.fit.simple$coefficients[2],
                            "P-Val"=anova(ma.fit.simple)$`Pr(>F)`[1]),
                  data.frame("Var"="Log Students/Teacher",
                            "Coefficient"=spt.fit.simple$coefficients[2],
                            "P-Val"=anova(spt.fit.simple)$`Pr(>F)`[1]),
                  data.frame("Var"="Teachers/Students",
                            "Coefficient"=tps.fit.simple$coefficients[2],
                            "P-Val"=anova(tps.fit.simple)$`Pr(>F)`[1]),
                  data.frame("Var"="Took Exam/Enroll",
                            "Coefficient"=sTook.fit.simple$coefficients[2],
                            "P-Val"=anova(sTook.fit.simple)$`Pr(>F)`[1]),
                  data.frame("Var"="Log Enrolled",
                            "Coefficient"=ssize.fit.simple$coefficients[2],
                            "P-Val"=anova(ssize.fit.simple)$`Pr(>F)`[1]))
comparison2<-rbind(data.frame("Var"="County ID",
                              "P-Value"=anova(CID.fit.simple)$`Pr(>F)`[1]),
                    data.frame("Var"="School ID",
                              "P-Value"=anova(SID.fit.simple)$`Pr(>F)`[1]),
                    data.frame("Var"="Grade",
                              "P-Value"=anova(Grade.fit.simple)$`Pr(>F)`[1]),
                    data.frame("Var"="Test Type",
                              "P-Value"=anova(Exam.fit.simple)$`Pr(>F)`[1]))

cat("Simple models w/ one independent variable, coefficients and p-values:\n")
```

```
## Simple models w/ one independent variable, coefficients and p-values:
```

```
print.data.frame(comparison,row.names=FALSE)
```

```
##                   Var  Coefficient         P.Val
##               P.White   0.27906396  7.288335e-91
##            FoodAssist -42.23201958 4.814015e-170
##            Experience   0.00213677  7.206373e-02
##        Log Experience   0.03497559  2.252749e-02
##               Hold MA   0.00135714  8.277510e-07
##  Log Students/Teacher   0.01758684  1.591254e-01
##     Teachers/Students  -0.43739279  1.066286e-02
##      Took Exam/Enroll   0.42923141  8.197086e-17
##          Log Enrolled   0.05244221  3.210370e-26
```

```
cat("\n\nSimple models w/ one categorical variable, p-values:\n")
```

```
##
##
## Simple models w/ one categorical variable, p-values:
```

```
print.data.frame(comparison2,row.names=FALSE)
```

```
##        Var      P.Value
##  County ID 3.992614e-34
##  School ID 1.613694e-11
##      Grade 1.424889e-08
##  Test Type 0.000000e+00
```

### Simple Models: Results

Looking at the results from the simple models, most of the variables look significant but there are a few that stand out. Raw experience seems insigificant, but log experience just meets our threshold. The log of students per teacher seems insignificant, but regular teachers per student meets does seem significant. Finally, food assistance could hardly be more significant than it already is. However, students per teacher, teacher experience, and (to a degree) teacher experience are orders of magnitude less important than whiteness, food assistance, and the log of the enrollment figures.

Turning to the categorical varaibles, all seem quite significant–particularly the test type. Curiously, COUNTY is more significant than the much finer school ID. As we are less interested in differences by test than county, it makes sense to use county as our grouping variable.

## From Simple to Hierarchical Models

As mentioned previously, we are going to create two different sets of models. The first, the "stoic models" will exclude data on testing and grade as these are outside of our interest. The second, the "intuitive models", will include all the variables deemed best for trying to predic the result. With both of these, we will experment with including or excluding teacher's experience and education.

For these variables, there is some question as to which to allow the slopes to vary for. In other words, which variables are likely to have different slopes depending on the county. The log of the enrollment seems like it will make little difference by county: overcrowded in Spokane is little different than overcrowded in Seattle (King County). One could say the same about the teachers' experience and education, as well as teachers per student.

This leaves two variables: food assistance and race (percent white). As the per capita income and living expenses vary by county, it would make sense for some rural counties to have higher levels of food assistance without as much of the social issues that tend to accompany poverty. For race... including county could be a good way to account for more than white vs nonwhite. Perhaps in Yakima non-white tends to mean Asian, but in Spokane it means Indian. However, as the price of including a varying slope for a variable is high, it makes the most sense to only include a random slope for food assistance.

To evaluate between these models, we are using Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). Looking at the early results, both AIC and BIC agree with respect to the best models. Among both the Stoic and Intuitive models, the criteria find the models that have fewer parameters to be preferable (the heaviest model even had an issue with a degenerate Hessian), but both also find that the more parameter-heavy intuitive fit works much better than the stoic fit. While initially I was not planning to use the more accurate stoic-fit model, the performance gap is sufficient to justify using the intuitive fit model instead.

## Comparing Complex Models

```
stoic_fit <- lmer(SHARE_PASSED ~ P_WHITE + LOG_ENROLL + STU_PER_TEACHER + (FOOD_ASSIST | COUNTY), data =

stoic_fit2 <- lmer(SHARE_PASSED ~ P_WHITE + LOG_ENROLL + STU_PER_TEACHER + MEAN_MA_DEGREE + (FOOD_ASSIS
                   data = wasl )
stoic_fit3 <- lmer(SHARE_PASSED ~ P_WHITE + LOG_ENROLL + STU_PER_TEACHER + LOG_T_EXP + (FOOD_ASSIST | C
                   data = wasl )
stoic_fit4 <- lmer(SHARE_PASSED ~ P_WHITE + LOG_ENROLL + STU_PER_TEACHER + LOG_T_EXP + MEAN_MA_DEGREE +
                   data = wasl )
intuitive_fit <- lmer(SHARE_PASSED ~ as.factor(TEST) + as.factor(GRADE) + LOG_ENROLL + STU_PER_TEACHER
                  (FOOD_ASSIST | COUNTY),
                  data = wasl )
intuitive_fit2 <- lmer(SHARE_PASSED ~ as.factor(TEST) + as.factor(GRADE) + LOG_ENROLL + STU_PER_TEACHER
                  (FOOD_ASSIST | COUNTY),
```

```
                  data = wasl )
intuitive_fit3 <- lmer(SHARE_PASSED ~ as.factor(TEST) + as.factor(GRADE) + LOG_ENROLL + STU_PER_TEACHER
                  (FOOD_ASSIST | COUNTY),
                  data = wasl )
intuitive_fit4 <- lmer(SHARE_PASSED ~ as.factor(TEST) + as.factor(GRADE) + LOG_ENROLL + STU_PER_TEACHER
                  (FOOD_ASSIST | COUNTY),
                  data = wasl )

# have tried adding P_WHITE to random effects with stoic_fit, only works with additional data on test a
cat('\n\n')
```

```
t(BIC(stoic_fit, stoic_fit2, stoic_fit3, stoic_fit4, intuitive_fit, intuitive_fit2, intuitive_fit3, intu
```

```
##     stoic_fit stoic_fit2 stoic_fit3 stoic_fit4 intuitive_fit intuitive_fit2
## df     8.0000     9.0000     9.0000    10.0000        12.000         13.000
## BIC -627.1414  -604.9764  -612.6074  -590.3087     -6007.308      -5991.414
##     intuitive_fit3 intuitive_fit4
## df          13.000         14.000
## BIC      -5983.715      -5967.745
```

```
t(AIC(stoic_fit, stoic_fit2, stoic_fit3, stoic_fit4, intuitive_fit, intuitive_fit2, intuitive_fit3, intu
```

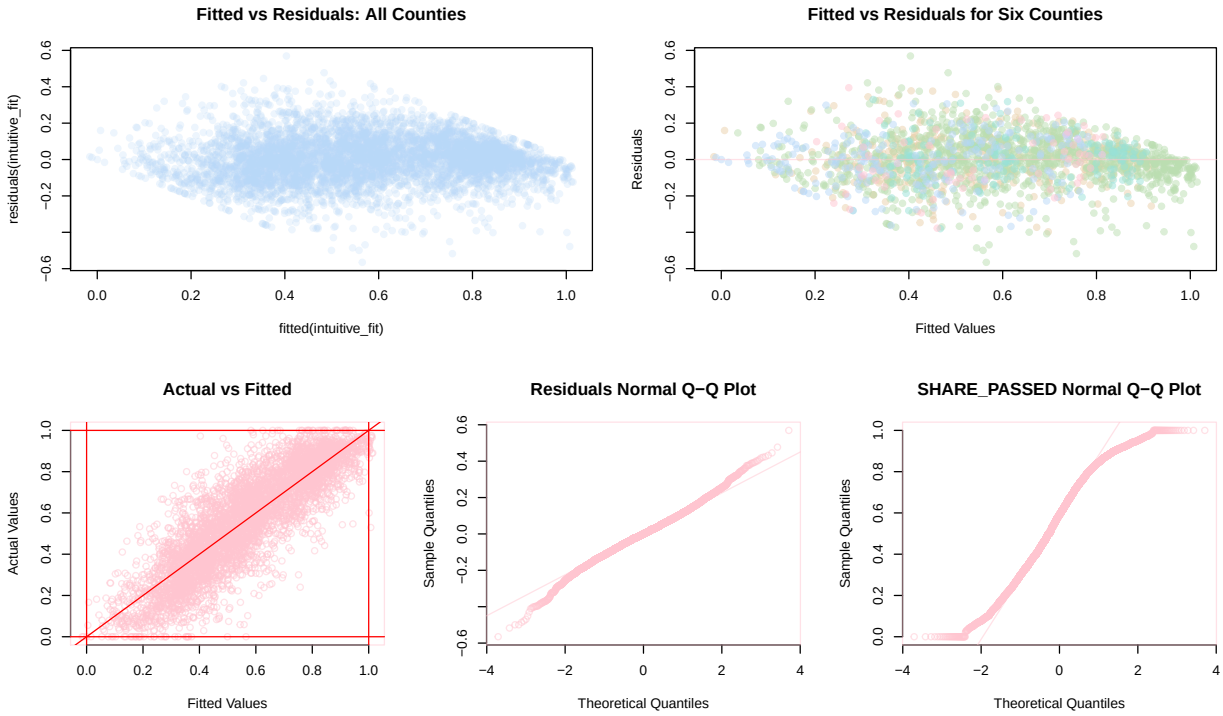
```
##     stoic_fit stoic_fit2 stoic_fit3 stoic_fit4 intuitive_fit intuitive_fit2
## df     8.0000     9.0000     9.0000    10.0000         12.00         13.000
## AIC -678.9557  -663.2675  -670.8985  -655.0766      -6085.03      -6075.613
##     intuitive_fit3 intuitive_fit4
## df          13.000          14.00
## AIC      -6067.913       -6058.42
```

# MODEL DIAGNOSTICS

Fitted vs residuals plots have very few points in the upper right and lower left corners. This makes it appear that there is some auto-correlation. However, this auto-correlation largely disappears when one shifts to lookin at just a sample of the counties involved. Looking instead at actual vs fitted, the reasons for this becomes clear: both actual and fitted values are approaching floors or cielings as to what they can perform. If the tests were to be made more difficult, the problem could eventually disappear for the upper bound, but get worse at the lower bound. Turning to a Normal Q-Q plot of the residuals for the model, they seem to be normally distributed, but a little light at the tails.

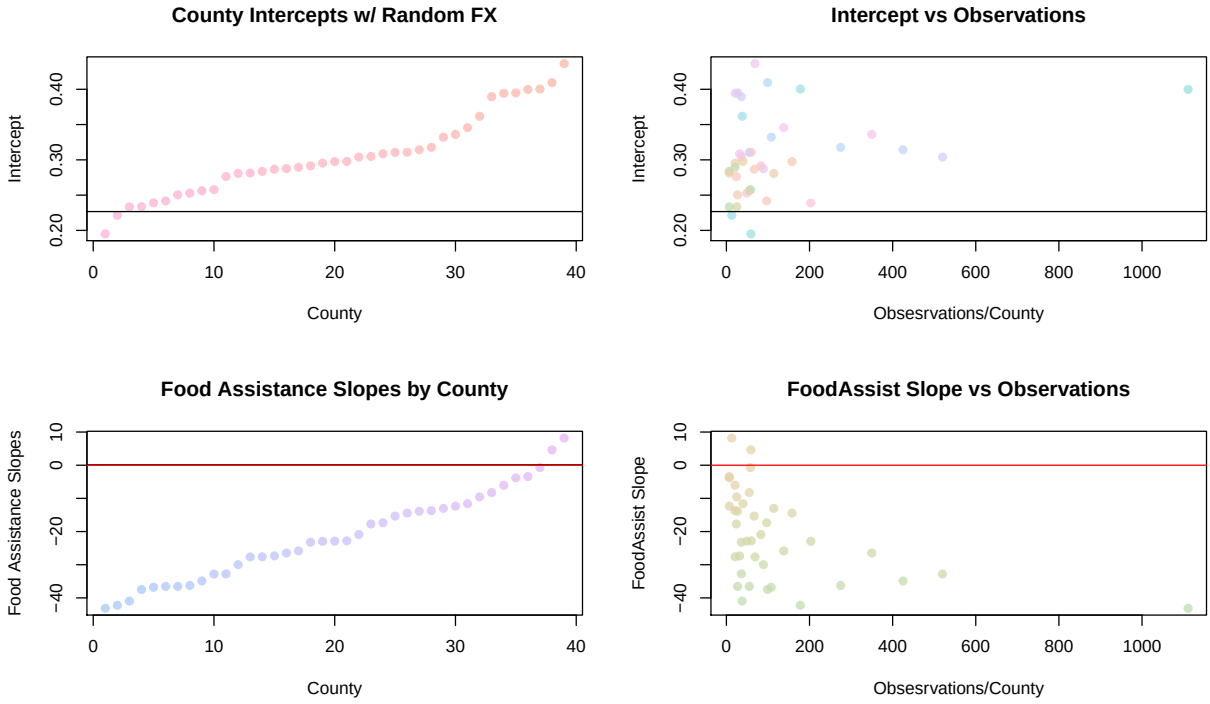
```
layout(matrix(c(1,1,1,2,2,2,3,3,5,5,4,4), 2, 6, byrow = TRUE))
valid<-!(is.na(wasl$SHARE_PASSED)|is.na(wasl$P_WHITE)|is.na(wasl$FOOD_ASSIST)|
           is.na(wasl$SHARE_TOOK)|is.na(wasl$LOG_ENROLL)|is.na(wasl$TEACHERS_PER_STU) )

plot(fitted(intuitive_fit),residuals(intuitive_fit),main="Fitted vs Residuals: All Counties", pch = 19,

plot(range(fitted(intuitive_fit),na.rm=TRUE),range(residuals(intuitive_fit),na.rm=TRUE),type="n",xlab="
G<-sample(unique(wasl$COUNTY),6)
iro<- hcl(360 / 6 * 1:6, alpha = 0.5)
these<-rep(FALSE,NROW(wasl))
for (i in 1:6)
{
  par(col=iro[i])
  points(fitted(intuitive_fit)[wasl$COUNTY==G[i]],residuals(intuitive_fit)[wasl$COUNTY==G[i]], pch = 19
}
abline(h=0)
```

```r
plot(fitted(intuitive_fit),wasl$SHARE_PASSED[valid],xlab="Fitted Values",ylab="Actual Values",main="Actu
abline(0,1,col="red");abline(v=0,col="red");abline(v=1,col="red")
abline(h=0,col="red");abline(h=1,col="red")

qqnorm(wasl$SHARE_PASSED,main="SHARE_PASSED Normal Q-Q Plot");qqline(wasl$SHARE_PASSED)
qqnorm(residuals(intuitive_fit),main="Residuals Normal Q-Q Plot");qqline(residuals(intuitive_fit))
```




```r
layout(matrix(c(1,3,2,4), 2, 2, byrow = TRUE))
rfx<-ranef(intuitive_fit)
ffx<-fixef(intuitive_fit)
fco<-summary(intuitive_fit)$coefficients

#ord<-order(rfx$SCHOOL[,1])
#plot(1:NROW(rfx$SCHOOL),(ffx[1]+rfx$SCHOOL[ord,1]),xlab="School",ylab="Intercept",main="School Interce
#abline(h=ffx[1])

ord<-order(rfx$COUNTY[,1])
plot(1:NROW(rfx$COUNTY),(ffx[1]+rfx$COUNTY[ord,1]),
     xlab="County",ylab="Intercept",main="County Intercepts w/ Random FX", pch = 19,
     col = hcl(330 + 60 / NROW(rfx$COUNTY) * 1:NROW(rfx$COUNTY ) ) )
abline(h=ffx[1])

ord<-order(rfx$COUNTY[,2])
plot(1:NROW(rfx$COUNTY),(rfx$COUNTY[ord,2]),
     xlab="County",ylab="Food Assistance Slopes",main="Food Assistance Slopes by County", pch = 19,
     col = hcl(240 + 60 / NROW(rfx$COUNTY) * 1:NROW(rfx$COUNTY) ) )
abline(h=ffx[8])
abline(h=0,col="red")
```

```r
sizes<-c(table(wasl$COUNTY))
#sizes2<-c(table(wasl$SCHOOL[valid]))

#plot(sizes2,(ffx[1]+rfx$SID[,1]),xlab="Observations/School",ylab="Intercept",main="Intercepts vs Obser
#abline(h=ffx[1])

plot(sizes,(ffx[1]+rfx$COUNTY[,1]),
     xlab="Obsesrvations/County",ylab="Intercept",main="Intercept vs Observations", pch = 19,
     col = hcl(150 + 60 / max(rfx$COUNTY) * rfx$COUNTY[,2], alpha = 0.75 ) )
abline(h=ffx[1])

plot(sizes,(rfx$COUNTY[,2]),
     xlab="Obsesrvations/County",ylab="FoodAssist Slope",main="FoodAssist Slope vs Observations", pch =
     col = hcl(60 + 60 / max(abs(rfx$COUNTY[,2])) * abs(rfx$COUNTY[,2]), alpha = 0.75 ) )
#abline(h=ffx[8])
abline(h=0,col="red")
```



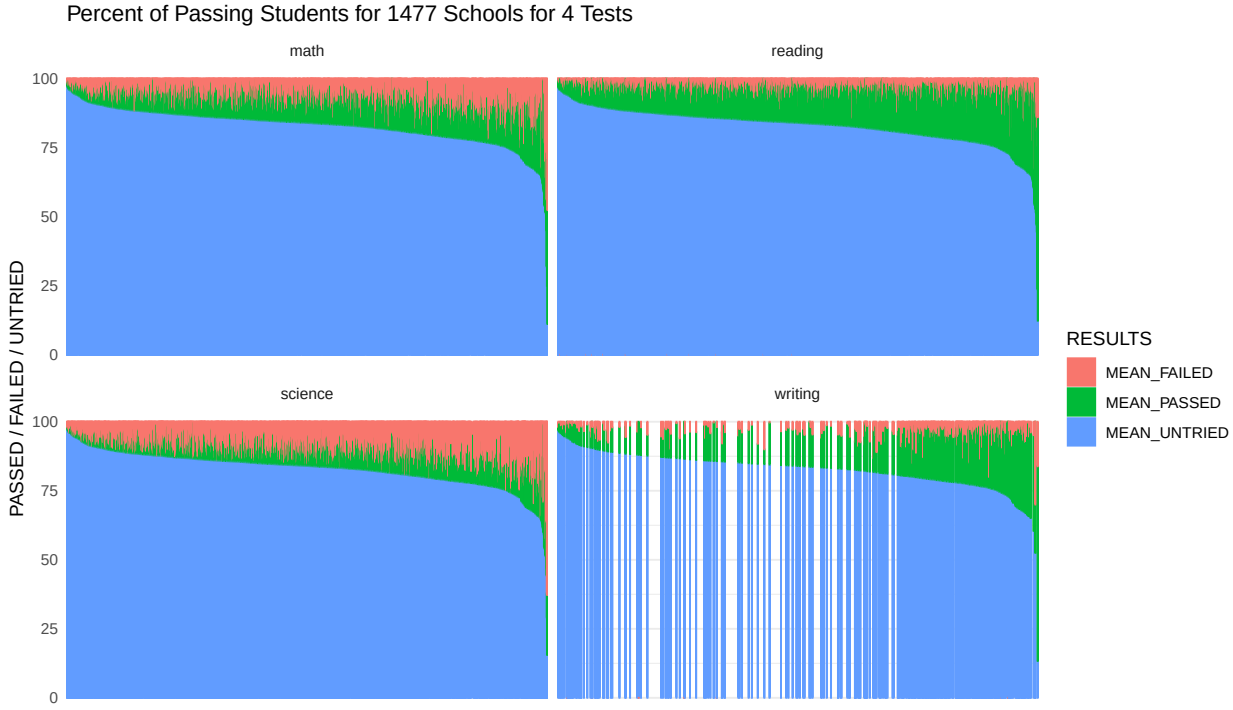

Looking instead at the number of observations for each county and comparing it to the random intercept and slopes for Food assist, we see that as the number of observations increase, the intercepts and slopes vary less and less. Moreover, the two counties wherin schools with more students on food assistance outperformed those with fewer students on food assistance also have very few observations. King County, which has the most observations, also had the most negative slope for the effects of food assistance. Howeer, looking King County itself does seem somewhat of an exception–as its intercept is also unusually high. The results at least indicate that King County itself should be grouped seperately from the other counties.

## Models With LogRatio Transformation

There is one major deficiency in the models we have been working with so far, and that is that our outcomes are compositional data. One cannot have a normal distribution when the results begin stacking up at 0% and

100%. Here, we have three different variables at work: the share who pass, the share who fail, and the share who were untried. The main trouble with logratios is that one has to create multiple models–and to make the results easily understood one needs to combine the results from all of the models and perform an inverse transformation. So for two variables, we will have to create two models.

```
wasl %>%
  arrange(TOOK / TOTAL_ENROLL) %>%
  mutate(ONES = 1,
  ID = cumsum(ONES) ) %>%
  group_by(ID, TEST) %>%
  summarize( MEAN_PASSED  = sum(PASSED) / sum(TOTAL_ENROLL) * 100,
             MEAN_FAILED  = sum(TOOK - PASSED) / sum(TOTAL_ENROLL) * 100,
             MEAN_UNTRIED = sum(TOTAL_ENROLL - TOOK) / sum(TOTAL_ENROLL ) * 100 ) %>%
  ungroup() %>%
  pivot_longer( cols = c('MEAN_PASSED', 'MEAN_FAILED', 'MEAN_UNTRIED'), names_to = 'RESULTS', values_to
  ggplot( aes(x = ID, y = MEAN, group = RESULTS, col = RESULTS, fill = RESULTS ) ) +
  geom_bar(stat = 'identity') +
  scale_x_discrete(labels = NULL) +
  xlab('') + ylab('PASSED / FAILED / UNTRIED') + ggtitle('Percent of Passing Students for 1477 Schools
  theme_minimal() + facet_wrap( ~ TEST)
```




```
wasl <- load_and_clean()
Y <- wasl %>%
  transmute(SHARE_PASSED = PASSED / TOTAL_ENROLL,
         SHARE_FAILED = (TOOK - PASSED) / TOTAL_ENROLL,
         SHARE_UNTRIED = (TOTAL_ENROLL - TOOK) / TOTAL_ENROLL )

Y <- ilr(Y)
names(Y) <- c('V1','V2')
```

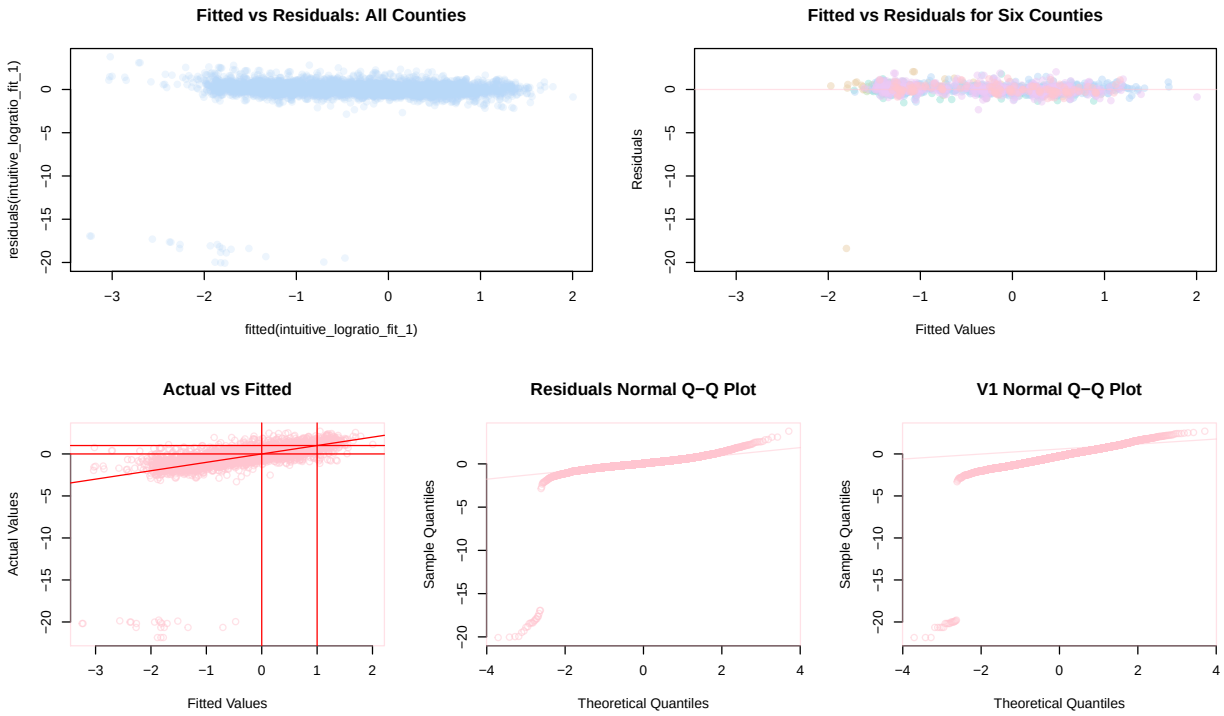
```
wasl <- cbind(wasl, Y)
```

```
#stoic_logratio_fit_1 <- lmer(V1 ~ P_WHITE + LOG_ENROLL + STU_PER_TEACHER + (FOOD_ASSIST | COUNTY), dat
#stoic_logratio_fit_2 <- lmer(V2 ~ P_WHITE + LOG_ENROLL + STU_PER_TEACHER + (FOOD_ASSIST | COUNTY), dat

intuitive_logratio_fit_1 <- lmer(V1 ~ as.factor(TEST) + as.factor(GRADE) + P_WHITE + LOG_ENROLL + STU_PE
intuitive_logratio_fit_2 <- lmer(V2 ~ as.factor(TEST) + as.factor(GRADE) + P_WHITE + LOG_ENROLL + STU_PE
```

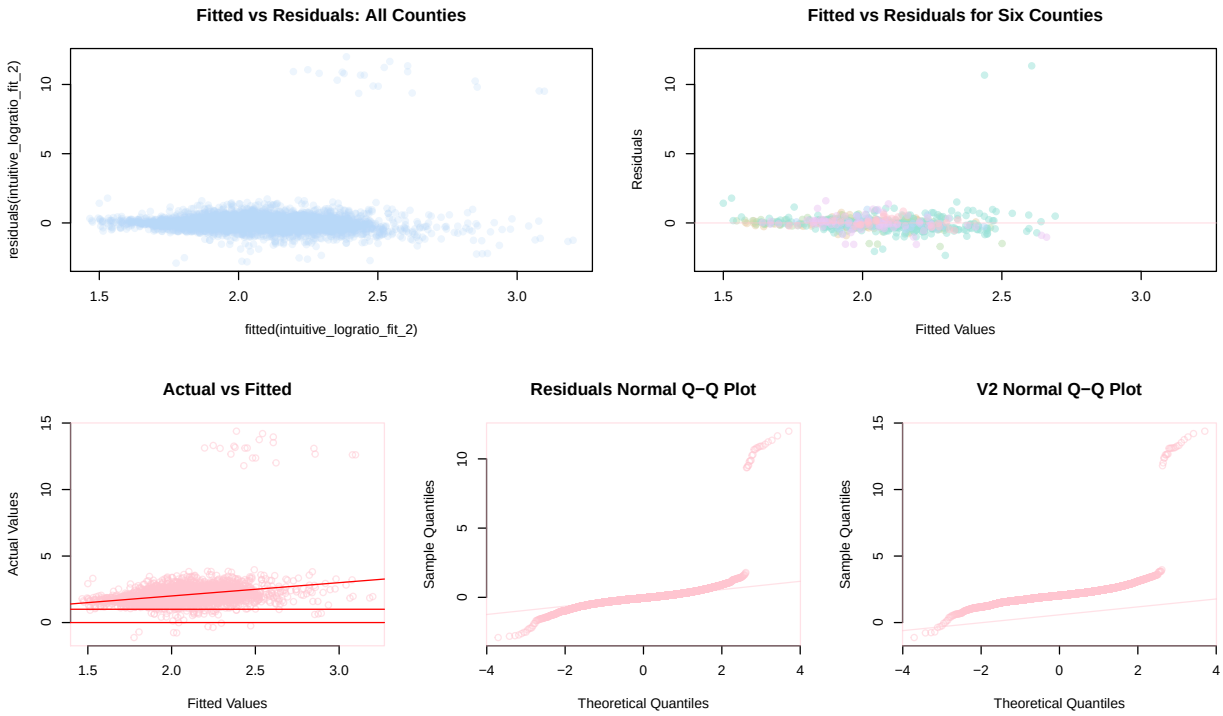
### Performance for Logratio Models: Not Combined

Looking at the two intuitive models that instead use logratios, the metrics for our two target variables both look fairly good aside from some clear outliers, which can be seen in all of the plots to evaluate metrics. Comparing the plots for V1 vs V2, the two are almost mirror images.

```
layout(matrix(c(1,1,1,2,2,2,3,3,5,5,4,4), 2, 6, byrow = TRUE))
valid<-!(is.na(wasl$V1)|is.na(wasl$P_WHITE)|is.na(wasl$FOOD_ASSIST)|
           is.na(wasl$SHARE_TOOK)|is.na(wasl$LOG_ENROLL)|is.na(wasl$TEACHERS_PER_STU) )

plot(fitted(intuitive_logratio_fit_1),residuals(intuitive_logratio_fit_1),main="Fitted vs Residuals: Al

plot(range(fitted(intuitive_logratio_fit_1),na.rm=TRUE),range(residuals(intuitive_logratio_fit_1),na.rm=
G<-sample(unique(wasl$COUNTY),6)
iro<- hcl(360 / 6 * 1:6, alpha = 0.5)
these<-rep(FALSE,NROW(wasl))
for (i in 1:6)
{
  par(col=iro[i])
  points(fitted(intuitive_logratio_fit_1)[wasl$COUNTY==G[i]],residuals(intuitive_logratio_fit_1)[wasl$C
}
abline(h=0)

plot(fitted(intuitive_logratio_fit_1),wasl$V1[valid],xlab="Fitted Values",ylab="Actual Values",main="Ac
abline(0,1,col="red");abline(v=0,col="red");abline(v=1,col="red")
abline(h=0,col="red");abline(h=1,col="red")

qqnorm(wasl$V1,main="V1 Normal Q-Q Plot");qqline(wasl$SHARE_PASSED)
qqnorm(residuals(intuitive_logratio_fit_1),main="Residuals Normal Q-Q Plot");qqline(residuals(intuitive
```




```
layout(matrix(c(1,1,1,2,2,2,3,3,5,5,4,4), 2, 6, byrow = TRUE))
valid<-!(is.na(wasl$V2)|is.na(wasl$P_WHITE)|is.na(wasl$FOOD_ASSIST)|
           is.na(wasl$SHARE_TOOK)|is.na(wasl$LOG_ENROLL)|is.na(wasl$TEACHERS_PER_STU) )

plot(fitted(intuitive_logratio_fit_2),residuals(intuitive_logratio_fit_2),main="Fitted vs Residuals: Al

plot(range(fitted(intuitive_logratio_fit_2),na.rm=TRUE),range(residuals(intuitive_logratio_fit_2),na.rm
G<-sample(unique(wasl$COUNTY),6)
iro<- hcl(360 / 6 * 1:6, alpha = 0.5)
these<-rep(FALSE,NROW(wasl))
for (i in 1:6)
{
  par(col=iro[i])
  points(fitted(intuitive_logratio_fit_2)[wasl$COUNTY==G[i]],residuals(intuitive_logratio_fit_2)[wasl$C
}
abline(h=0)

plot(fitted(intuitive_logratio_fit_2),wasl$V2[valid],xlab="Fitted Values",ylab="Actual Values",main="Ac
abline(0,1,col="red");abline(v=0,col="red");abline(v=1,col="red")
abline(h=0,col="red");abline(h=1,col="red")

qqnorm(wasl$V2,main="V2 Normal Q-Q Plot");qqline(wasl$SHARE_PASSED)
qqnorm(residuals(intuitive_logratio_fit_2),main="Residuals Normal Q-Q Plot");qqline(residuals(intuitive_
```



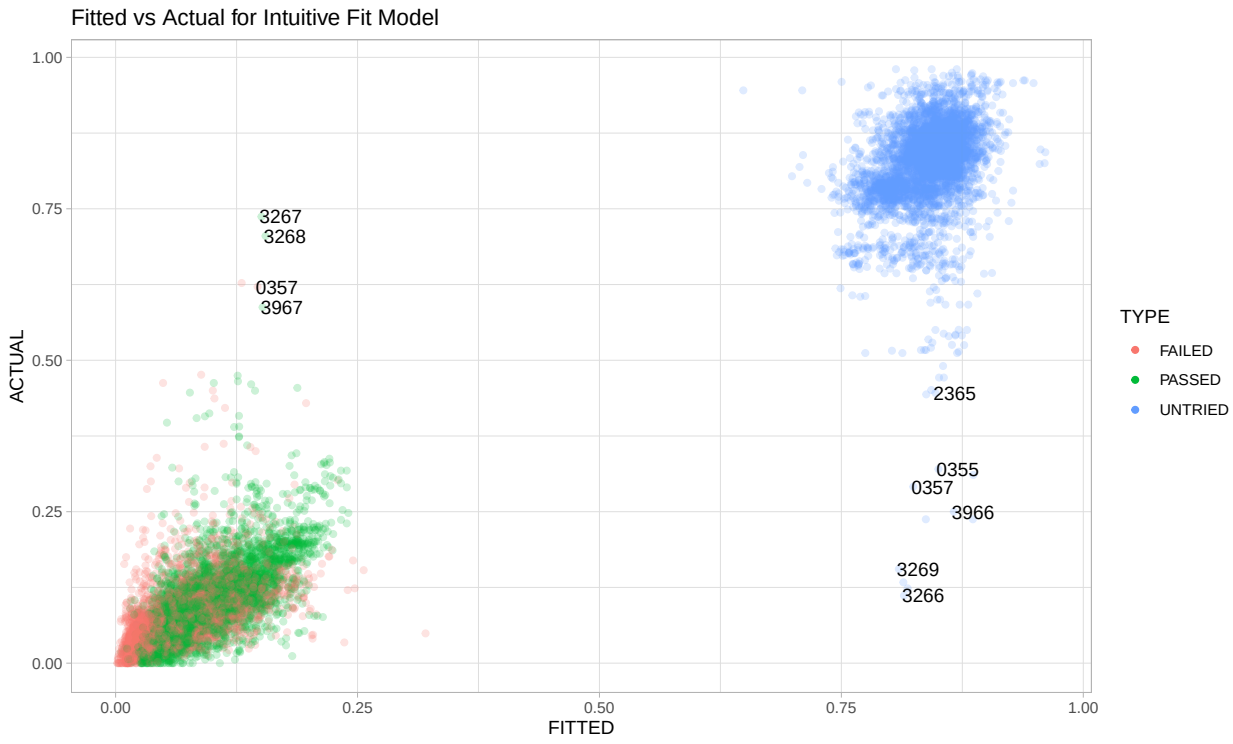

### Prediction vs Actual and Residuals

Combining the two models that use logratios and getting their predictions and residuals, we can see some of what was going on with the earlier outliers. A handful of schools participated in the exam to a far greater degree than their peers. These schools accordingly had much higher pass/fail rates than their peers.

```
# need to manually calculate residuals
Y <- wasl %>%
  transmute(SHARE_PASSED = PASSED / TOTAL_ENROLL,
         SHARE_FAILED = (TOOK - PASSED) / TOTAL_ENROLL,
         SHARE_UNTRIED = (TOTAL_ENROLL - TOOK) / TOTAL_ENROLL )
Y <- Y[valid,]
names(Y) <- c('PASSED','FAILED','UNTRIED')
Y_hat <- ilrInv(cbind(fitted(intuitive_logratio_fit_1), fitted(intuitive_logratio_fit_2)))
Y_hat <- as.data.frame(Y_hat)
names(Y_hat) <- c('PASSED','FAILED','UNTRIED')
res <- Y - Y_hat
names(res) <- c('PASSED','FAILED','UNTRIED')
Y <- Y %>%
  mutate( ONES = 1, ID_ = str_pad(cumsum(ONES), width = 4, pad = '0' ) ) %>%
  select( - ONES) %>%
  pivot_longer( cols = ends_with('D'), names_to = 'TYPE', values_to = 'ACTUAL') %>%
  rename(ID = ID_) %>%
  mutate(ID = paste(ID, TYPE) ) %>%
  select( - TYPE)
Y_hat <- Y_hat %>%
  mutate( ONES = 1, ID_ = str_pad(cumsum(ONES), width = 4, pad = '0' ) ) %>%
  select( - ONES) %>%
  pivot_longer( cols = ends_with('D'), names_to = 'TYPE', values_to = 'FITTED') %>%
  rename(ID = ID_) %>%
```

```
  mutate(ID = paste(ID, TYPE) ) %>%
  select( - TYPE )
res <- res %>%
  mutate( ONES = 1, ID_ = str_pad(cumsum(ONES), width = 4, pad = '0' ) ) %>%
  select( - ONES) %>%
  pivot_longer( cols = ends_with('D'), names_to = 'TYPE', values_to = 'RES') %>%
  rename(ID = ID_) %>%
  mutate(ID = paste(ID, TYPE) )
intuitive_metrics <- Y %>%
  full_join(Y_hat, by = 'ID') %>%
  full_join(res, by = 'ID') %>%
  mutate(ID = substring(ID, 1, 4) ) %>%
  select(ID, TYPE, everything() )
rm(Y, res, Y_hat)
```

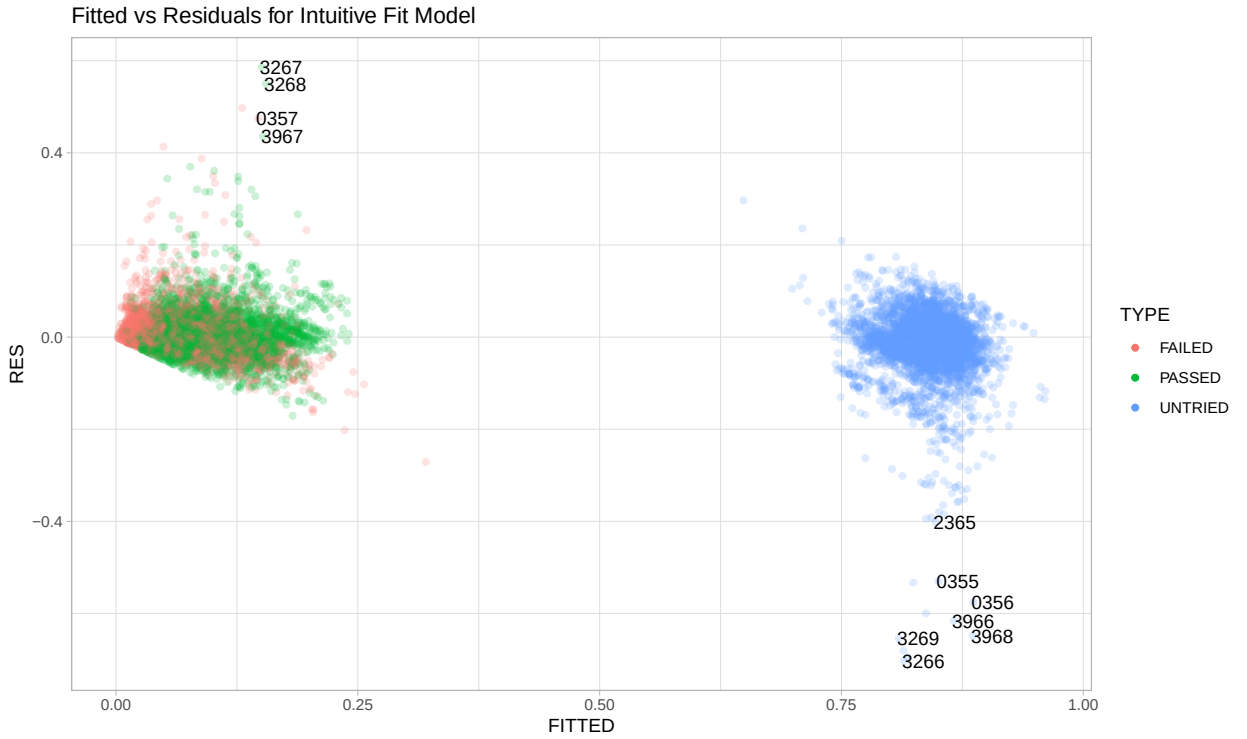
```
extreme_intuitive_metrics<- intuitive_metrics %>%
  filter( (TYPE %in% c('PASSED', 'FAILED' ) & ACTUAL > 0.5) | (TYPE == 'UNTRIED' & ACTUAL < 0.45) )

intuitive_metrics %>%
  ggplot( aes(x = FITTED, y = ACTUAL, group = TYPE, col = TYPE, fill = TYPE, alpha = I(0.2) ) ) + geom_
  geom_text(data = extreme_intuitive_metrics,
            aes(x = FITTED, y = ACTUAL, label = ID, group = TYPE, alpha = I(1) ),
            inherit.aes = FALSE, check_overlap = TRUE, nudge_x = 0.02 ) +
  ggtitle( 'Fitted vs Actual for Intuitive Fit Model ')
```




```
#  geom_label(pos = 'nudge')
```

```
extreme_intuitive_metrics<- intuitive_metrics %>%
  filter( (TYPE %in% c('PASSED', 'FAILED' ) & ACTUAL > 0.5) | (TYPE == 'UNTRIED' & ACTUAL < 0.45) )
```

```
intuitive_metrics %>%
  ggplot( aes(x = FITTED, y = RES, group = TYPE, col = TYPE, fill = TYPE, alpha = I(0.2) ) ) + geom_poi
  geom_text(data = extreme_intuitive_metrics,
            aes(x = FITTED, y = RES, label = ID, group = TYPE, alpha = I(1) ),
            inherit.aes = FALSE, check_overlap = TRUE, nudge_x = 0.02 ) +
  ggtitle('Fitted vs Residuals for Intuitive Fit Model')
```




```
#  geom_label(pos = 'nudge')
```

# EVALUATION

Given that we have found a fairly good model, what does this data mean for school select? One of the best ways to show this type of data is a ropeletter used to model what a small change in some of the features would do to performance.

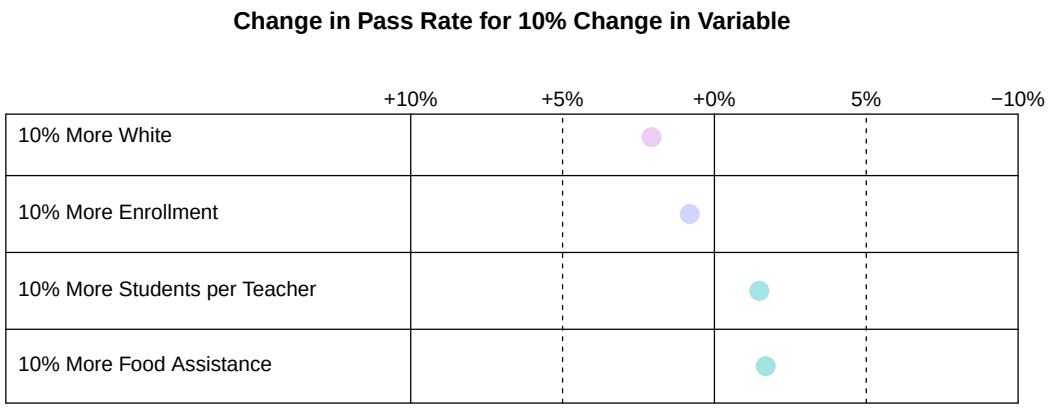
```
shift_predict <- function(model_1, model_2, data, col, shift = 0.1){
    y_hat <- cbind(predict(model_1), predict(model_2)) %>%
      ilrInv()
  if(grepl('log',stringr::str_to_lower(col) ) ){
    data[,col] <- data[,col] + shift
  }
  else{
    data[,col] <- data[,col] * (1 + shift)
  }
  y_hat2 <- cbind(predict(model_1, newdata = data), predict(model_2, newdata = data) ) %>%
    ilrInv()
  return( (y_hat2[,1] - y_hat[,1]) / y_hat[,1] )
}
```

```
predictions <- NULL
predictions <- c(predictions, mean(shift_predict(intuitive_logratio_fit_1, intuitive_logratio_fit_2, was
predictions <- c(predictions, mean(shift_predict(intuitive_logratio_fit_1, intuitive_logratio_fit_2, was
predictions <- c(predictions, mean(shift_predict(intuitive_logratio_fit_1, intuitive_logratio_fit_2, was
predictions <- c(predictions, mean(shift_predict(intuitive_logratio_fit_1, intuitive_logratio_fit_2, was
```

## Ropeladders

The first ropeladder looks at how changing the variables white, total enrollment, students per teacher, and food assistance would affect the performance of the school (logged results were handled so as to effectly increase the pre-logged value by 10%). Here, the only surprising result is that schools with more enrollment actually perform slightly better–but that is perhaps because we are also taking account of students per teacher, which has a stronger negative effect. If one thinks about the total number of enrolled students as a sign of popularity (parents will move to get into a good school), then it would make sense that total enrollment might be a good sign so long as the students per teacher are not also increasing.

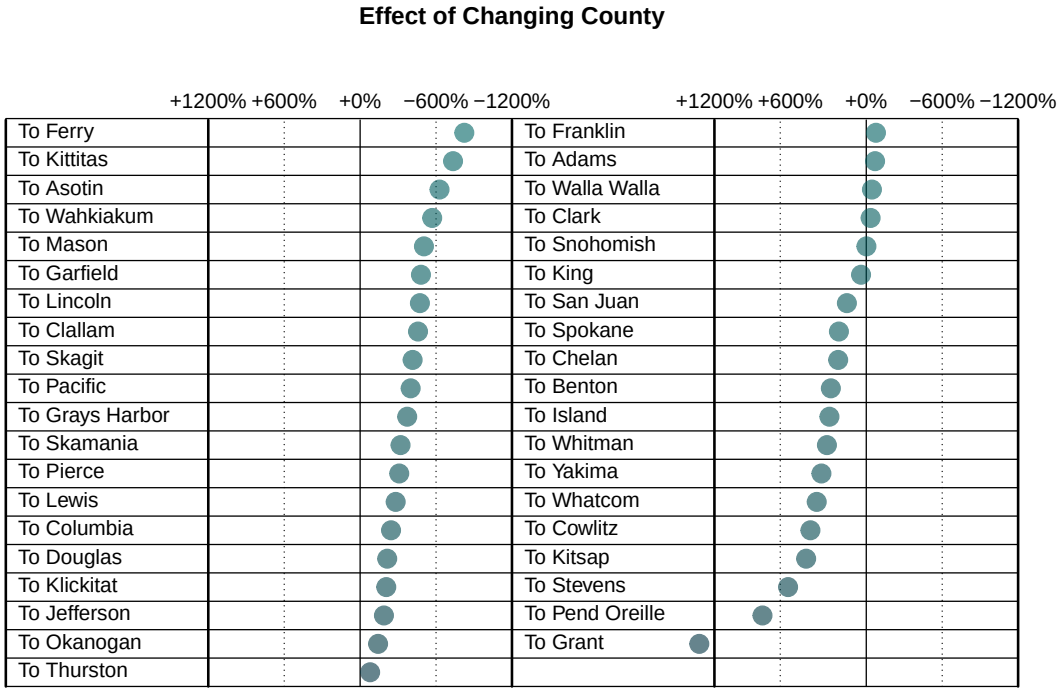
```
plot.new()
title('Change in Pass Rate for 10% Change in Variable')
segments(x0 = c(0, 0.4, 0.7, 1.0), x1 = c(0, 0.4, 0.7, 1.0), y0 = -0.04, y1 = 0.90 )
segments(x0 = c(0.55, 0.85), x1 = c(0.55, 0.85), y0 = -0.04, y1 = 0.90, lty = 2 )
segments(x0 = 0, x1 = 1, y0 = c(0.90,.70, .45, .20, -0.04) )
text(0.4, 0.95, '+10%'); text(0.55, 0.95, '+5%'); text(0.7, 0.95, '+0%'); text(0.85, 0.95, '5%'); text(

text(0, .825, '10% More White', pos = 4)
text(0, .575, '10% More Enrollment', pos = 4)
text(0, .325, '10% More Students per Teacher', pos = 4)
text(0, .080, '10% More Food Assistance', pos = 4)
#predictions <- c(0.25, 0.125, -0.125, -0.25)
points(predictions * (- 0.3 / 0.25 * 8 / 5 ) + 0.7, c(.825, .575, .325, .080),
       pch = 19,
       col = hcl(h = predictions / max(predictions) * 60 + 240 , alpha = 0.9 ), cex = 2 )
```




```
county_shift_predict <- function(model_1, model_2, data, shift = 'next', baseline = 'all'){
  y_hat <- cbind(predict(model_1), predict(model_2)) %>%
    ilrInv()
  if( shift == 'next' ){
```

```
    if (baseline == 'all') shift <- unique(data$COUNTY)[1]
    else shift <- grep(unique(baseline, data$COUNTY) ) + 1
  }
  data$COUNTY <- shift
  y_hat2 <- cbind(predict(model_1, newdata = data), predict(model_2, newdata = data) ) %>%
    ilrInv()
  return( sum( (y_hat2[,1] - y_hat[,1]) / y_hat[,1] ) )
}

county_shifts <- data.frame(COUNTY = unique(wasl$COUNTY), SHIFT = 0)

for (county in county_shifts$COUNTY ){
  county_shifts$SHIFT[county_shifts$COUNTY == county] <- county_shift_predict(intuitive_logratio_fit_1,
}
county_shifts <- county_shifts %>%
  mutate(COUNTY = paste('To', COUNTY ) ) %>%
  arrange(SHIFT)
```

```
# need to make a function to do this...or a geom
plot.new()
title('Effect of Changing County')
segments(x0 = c(0, 0.2, 0.5, 0.7, 1), y0 = 0.00, y1 = 0.95, lwd = 2)
segments(x0 = 0, x1 = 1, y0 = 0:20 / 21 )
segments(x0 = c(0.35, 0.85), y0 = 0, y1 = 0.95)
segments(x0 = c(0.275,0.425, 0.765, 0.925), y0 = 0, y1 = 0.95, lty = 3)
text(x = c(0.2, 0.275, 0.35, 0.425, 0.5, 0.7, 0.775, 0.85, 0.925, 1), y = 0.95, labels = c('+1200%','+6
text(x = 0,   y = 0:19 / 21 + 1/21/2, labels = county_shifts$COUNTY[20:1], pos = 4)
text(x = 0.5, y = 1:19 / 21 + 1/21/2, labels = county_shifts$COUNTY[39:21], pos = 4)
points(x = county_shifts$SHIFT[20:1] / - 1200 * 0.15 + 0.35, y = 0:19 / 21 + 1 / 21 / 2, pch = 19, cex =
       col = hcl(h = 190 + 30/max(abs(county_shifts$SHIFT)) * abs(county_shifts$SHIFT),
                 l = 40 + 40 /max(abs(county_shifts$SHIFT) ) * county_shifts$SHIFT,
                 alpha = 0.6) )
points(x = county_shifts$SHIFT[39:21] / - 1200 * 0.15 + 0.85, y = 1:19 / 21 + 1 / 21 / 2, pch = 19, cex
       col = hcl(h = 190 + 30/max(abs(county_shifts$SHIFT)) * abs(county_shifts$SHIFT),
                 l = 40 + 40 /max(abs(county_shifts$SHIFT) ) * county_shifts$SHIFT,
                 alpha = 0.6) )
```




The second ropeladder asks the ludicrous question of how Washington State's schools would perform if all of the counties had the same effect as a particular county. While the idea is a little absurd–there is a sharp limit to how much urban King County can become like rural Stevens County, it is a handy way to separate the effects of each county from the effects of a county being more affluent or having more teachers per student.

## Revisiting the Best and Worst Schools

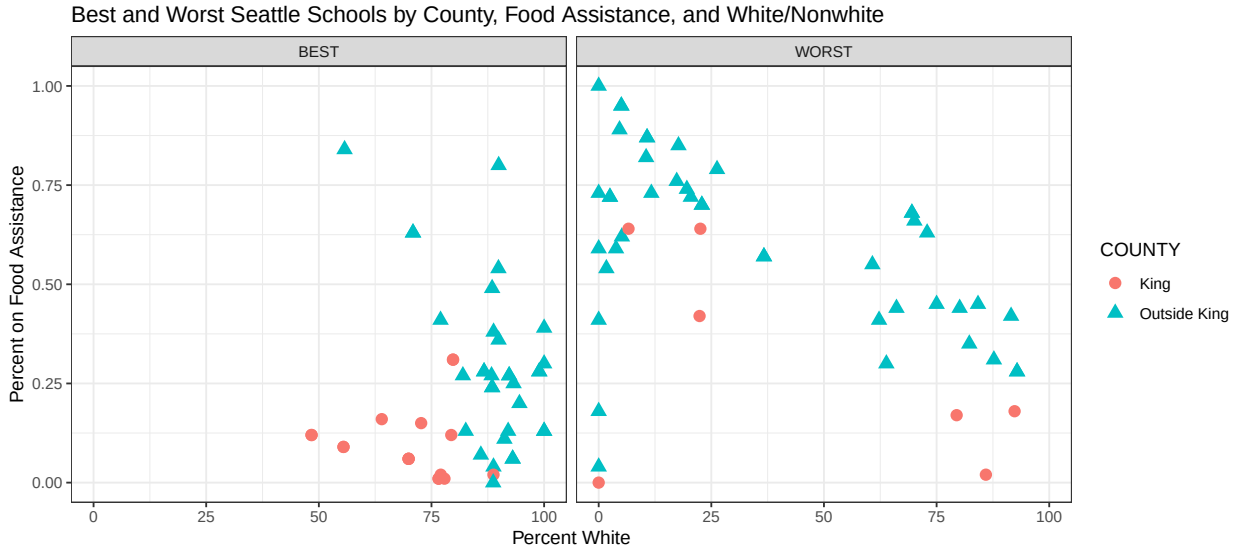
Looking at poverty and racial makeup for the 50 best and worst schools, we can see these variables in action. The best schools tend to be in counties with less poverty and whiter populations, while the worst tend to be more impoverished and less white.

The relationship seems weaker for the worst performing schools than the best performing schools. Two of the worst performing schools in King County, one entirely non-white and the other mostly white, have almost no poverty. In general though, the schools with the lowest poverty are mostly absent from the list of the worst schools.

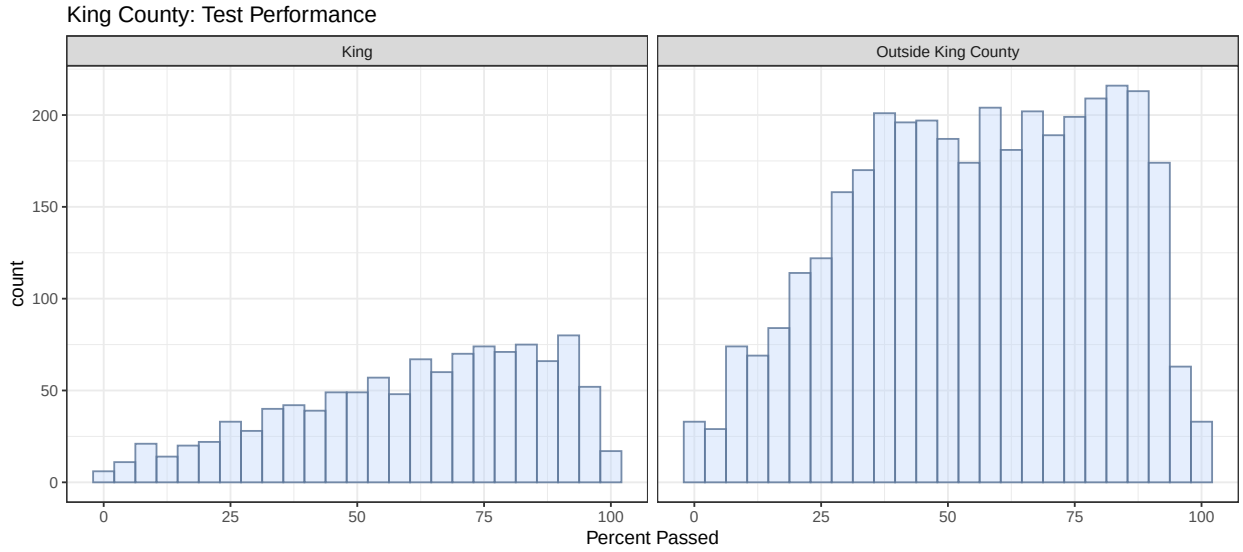
For the best schools, one can see the variables measuring poverty and whiteness at work. However, there is also a strong distinction between the schools inside and outside of King County. Those within King County tend to be more diverse, but that is coupled with lower levels of poverty. The additional racial diversity in the best-performing Seattle Schools seems to be coming from wealthy families. Meanwhile, those outside King Coutny tend to be more white, but their schools also tend to have more students on food assistance.

```
wasl %>%
  arrange(SHARE_PASSED) %>%
  slice(c(1:50, (length(ID) - 50:1) ) ) %>%
  mutate(CLASS = if_else(SHARE_PASSED > median(SHARE_PASSED), 'BEST','WORST' ) ) %>%
  mutate(P_WHITE = round(P_WHITE, 4) * 100, FOOD_ASSIST = round(FOOD_ASSIST, 4) * 100 ) %>%
```

```
  mutate(COUNTY = if_else(COUNTY == 'King', 'King', 'Outside King') ) %>%
  ggplot( aes(x = P_WHITE, y = FOOD_ASSIST, col = COUNTY, shape = COUNTY, fill = COUNTY) ) + geom_point
  ylab('Percent on Food Assistance') + xlab('Percent White') + facet_wrap(~CLASS) +
  ggtitle('Best and Worst Seattle Schools by County, Food Assistance, and White/Nonwhite') + theme_bw()
```




```
wasl %>%
  mutate(COUNTY = if_else(COUNTY == 'King', 'King', 'Outside King County' ) ) %>%
  mutate(SHARE_PASSED = SHARE_PASSED * 100) %>%
  ggplot( aes(x = SHARE_PASSED) ) + geom_histogram(bins = 25, col = hcl(247, l = 50, alpha = 0.85 ), fi
  ggtitle('King County: Test Performance') + facet_wrap(~COUNTY) + xlab('Percent Passed') + theme_bw()
```




This likely reflects how the housing market and school competition works differently in urban vs rural areas. Due to the smaller school zones and higher property values, it is more difficult for a poor family to locate themselves in a good urban school district. In contrast, the lower property values and larger school zones of less urban counties makes it much easier for a poor family deeply concerned wtih their children's education to move to a good school district.

Shifting to the perspective of the wealthy on the other hand, King County offers urbanites the possibility of

moving to a wealthy district without greating lengthening their commute, while in less urban moving to a better school district can come at the cost of much longer commuting times.

### Next Steps

It may be a good idea to try redoing this project with more up-to-date data. Also, considering how many of the best schools are outliers with unsually high numbers of students taking the exam, it could be effective to redo the exam focusing on just those schools. That said, doing so would greatly reduce the number of variables, making it harder to make meaningful predictions. Instead, at that point, it may be necessary to switch to more qualitative methods. However, that would be the work of someone more concerned with teaching methodologies than finding the best schools. Finally, it may be useful to get more detailed data on race, but perhaps that is a rabbit-hole we do not want to go any farther down.

## DEPLOYMENT

If a small to medium-sized company is seeking to take advantage of its location to attract good talent, there are plenty of opportunities within Washington State to do so. While competition for quality schools is fiercest within King County, but most of the State is outside of Seattle, and so too is the majority of the best schools. As enticing as a major city like Seattle is, the school and cost of living are high enough that many a familiy would accept a lower cost of living to reduce the stress caused by both. Whether one is looking more inland to Grant County, or a little West of Seattle to Clallam Bay, there are many excellent options.