R Notebook

```
data <- read_csv("SAT_School_Participation_and_Performance__2012-2013.csv")
```

```
## Parsed with column specification:
## cols(
##   `District Number` = col_double(),
##   District = col_character(),
##   School = col_character(),
##   `Test-takers: 2012` = col_double(),
##   `Test-takers: 2013` = col_double(),
##   `Test-takers: Change%` = col_double(),
##   `Participation Rate (estimate): 2012` = col_double(),
##   `Participation Rate (estimate): 2013` = col_double(),
##   `Participation Rate (estimate): Change%` = col_double(),
##   `Percent Meeting Benchmark: 2012` = col_double(),
##   `Percent Meeting Benchmark: 2013` = col_double(),
##   `Percent Meeting Benchmark: Change%` = col_double()
## )
```

```
#Alex's contribution: Tidying up the data
df <- data %>% select(-1, -6, -9, -12) %>% rename(district = "District", school = "School", t_takes2012
df <- df %>% dplyr::filter(!(is.na(t_takes2012) | is.na(t_takes2013) | is.na(part_rate2012) | is.na(par
df
```

```
## # A tibble: 187 x 8
##    district school t_takes2012 t_takes2013 part_rate2012 part_rate2013
##    <chr>    <chr>        <dbl>       <dbl>         <dbl>         <dbl>
##  1 Ansonia  Anson~         118         104            67            61
##  2 Avon     Avon ~         254         243            90            89
##  3 Berlin   Berli~         216         220            81            82
##  4 Bethel   Bethe~         200         190            86            82
##  5 Bloomfi~ Bloom~         116         130            79            89
##  6 Bloomfi~ Big P~          14          30           100           100
##  7 Bolton   Bolto~          62          70            85            96
##  8 Branford Branf~         196         213            77            84
##  9 Bridgep~ Bassi~         105         122            52            60
## 10 Bridgep~ Centr~         346         305            78            69
## # ... with 177 more rows, and 2 more variables: perc_mb2012 <dbl>,
## #   perc_mb2013 <dbl>
```

bmr = number of meeting Benchmark / number of total seniors = (t_takes*perc_mb) / (t_takes/part_rate) = pec_mb*part_rate bmr = perc_mb*part_rate*0.0001

We use bmr because it's a better measurement for comparing how well schools do. If 2 schools have the same percentage meeting benchmark, but one of them has a higher participation rate then the one with the higher participation rate is the better school.

```
#Alex's contribution: creating BMR formula
#df1 is for testtakers for each school+year
df1 <- df %>%
  select(1:4) %>%
  rename(`2012` = t_takes2012, `2013` = t_takes2013) %>%
  gather(3,4,key = "year", value = "t_takes") %>%
  arrange(school)

#df2 is participation rate for each school+year
df2 <- df %>% select(1,2,5,6) %>%
  rename(`2012` = part_rate2012, `2013` = part_rate2013) %>%
  gather(3,4,key = "year", value = "part_rate")

#df3 is percentage meeting benchmark for each school+year
df3 <- df %>%
  select(1,2,7,8) %>%
  rename(`2012` = perc_mb2012, `2013` = perc_mb2013) %>%
  gather(3,4,key = "year", value = "perc_mb")

#df4 combines them all
df4 <- df1 %>%
  full_join(df2,by = c("district","school","year")) %>%
  full_join(df3,by = c("district","school","year"))
df4 <- df4 %>%
  mutate(bmr = perc_mb*part_rate*1e-4)

df4
```

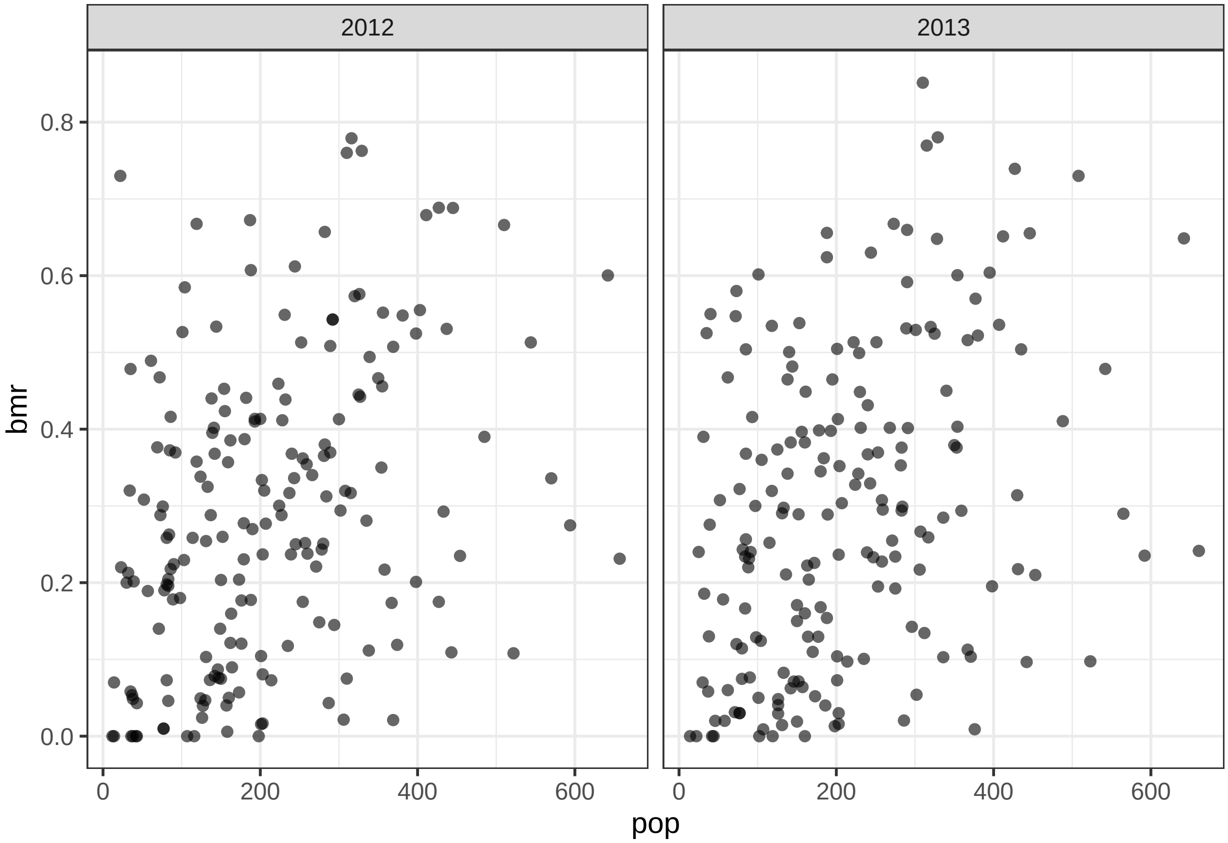
```
## # A tibble: 374 x 7
##    district            school          year  t_takes part_rate perc_mb    bmr
##    <chr>               <chr>           <chr>   <dbl>     <dbl>   <dbl>  <dbl>
##  1 Stamford            Academy of In~ 2012      133        82      47 0.385
##  2 Stamford            Academy of In~ 2013      142        88      51 0.449
##  3 Connecticut Techn~ Albert I Prin~ 2012       92        58       1 0.0058
##  4 Connecticut Techn~ Albert I Prin~ 2013       88        55       0 0
##  5 Amistad Academy D~ Amistad Acade~ 2012       34       100      32 0.32
##  6 Amistad Academy D~ Amistad Acade~ 2013       31       100      39 0.39
##  7 Regional 05         Amity Regiona~ 2012      381        87      61 0.531
##  8 Regional 05         Amity Regiona~ 2013      348        80      63 0.504
##  9 Ansonia             Ansonia High ~ 2012      118        67      18 0.121
## 10 Ansonia             Ansonia High ~ 2013      104        61      18 0.110
## # ... with 364 more rows
```

First we'll get the senior population for each school (denoted as pop)

```
data <- df4 %>% mutate(pop = floor(1e2*t_takes / part_rate))
```

Now lets plot it

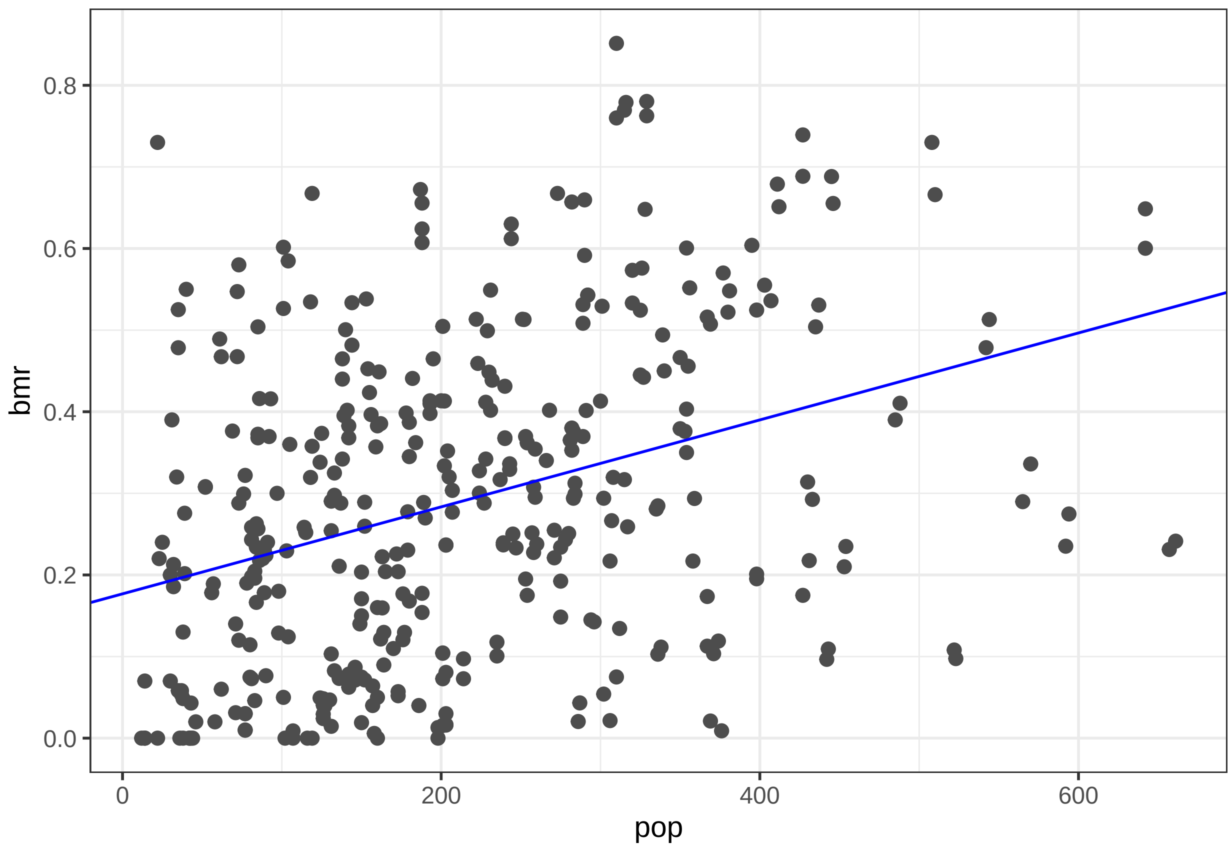
```
ggplot(data) +
  geom_point(aes(pop,bmr),alpha=3/5) +
  facet_wrap(~year) +
  theme_bw()
```



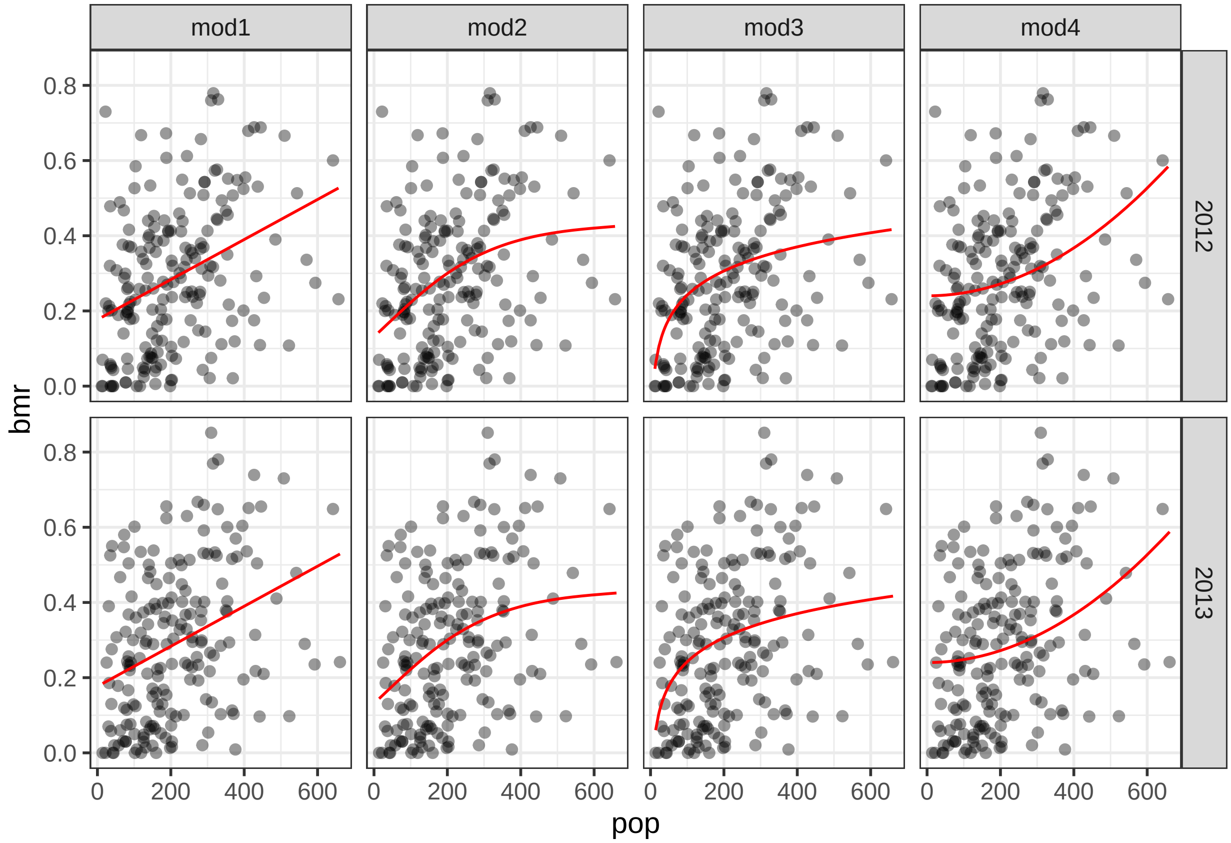
The data is relatively scattered, but we can see a weak positive linear trend.

Let's use mean-square residuals

```
#mean-square residuals
measure_distance <- function(mod,data){
  diff <- data$bmr - (mod[1] + data$pop*mod[2])
  sqrt(mean(diff^2))
}

best <- optim(c(0, 0), measure_distance, data = data)

ggplot(data, aes(pop, bmr)) +
  geom_point(size = 2, colour = "grey30") +
  geom_abline(color="blue",intercept = best$par[1], slope = best$par[2]) +
  theme_bw()
```

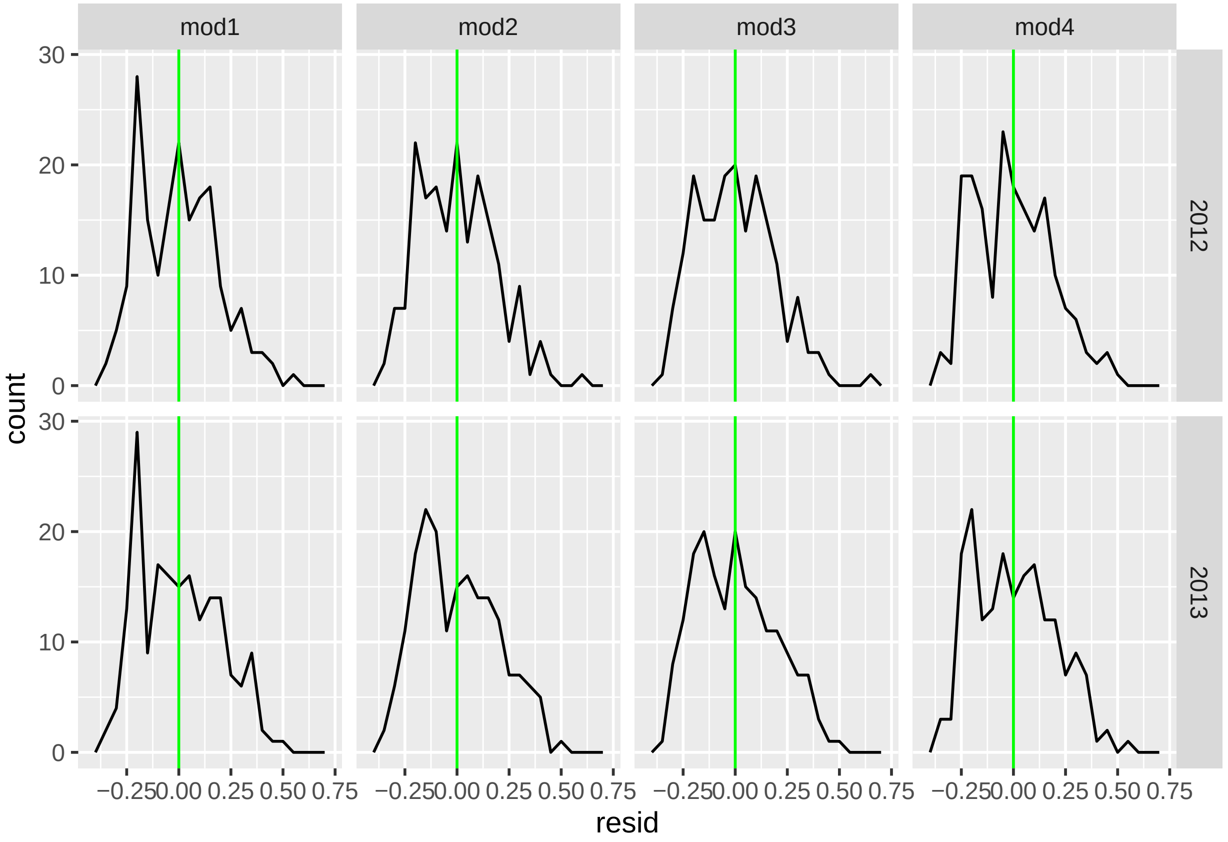


```
mod1 <- lm(bmr ~ ns(pop, 1), data = data)
mod2 <- lm(bmr ~ ns(pop, 2), data = data)
mod3 <- lm(bmr ~ log(pop, base = exp(1)), data = data)
mod4 <- lm(bmr ~ I(pop^2), data = data)

data %>%
  gather_predictions(mod1, mod2, mod3, mod4) %>%
  ggplot(aes(pop, bmr)) +
  geom_point(alpha=2/5) +
  geom_line(aes(pop,pred), colour = "red") +
  facet_grid(year~ model) +
  theme_bw()
```



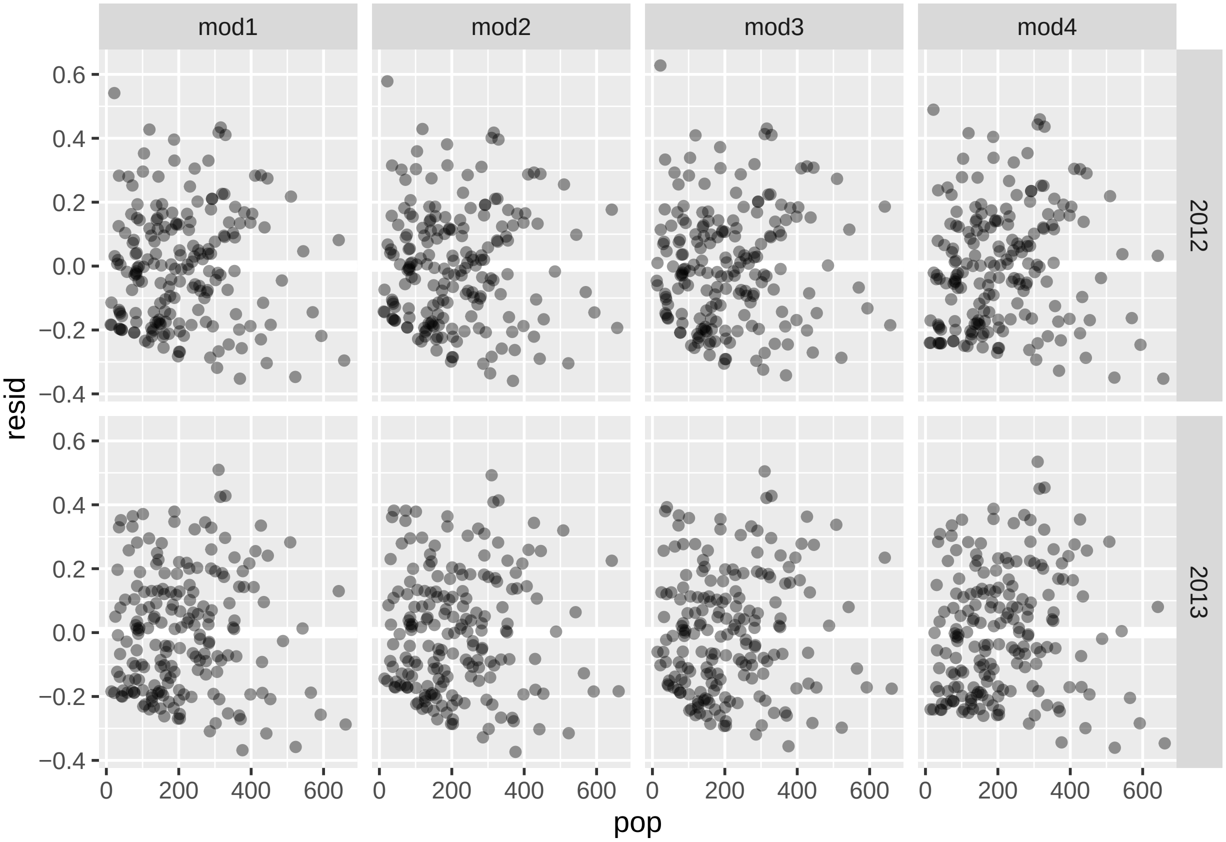
Let's check the residuals for any patterns

```
data %>%
  gather_residuals(mod1,mod2,mod3,mod4) %>%
  ggplot(aes(resid)) +
  geom_freqpoly(binwidth = 0.05) +
  geom_vline(xintercept = 0, colour = "Green", size=0.5) +
  facet_grid(year ~ model)
```



Looks approximately normal for all.

```
data %>%
  gather_residuals(mod1,mod2,mod3,mod4) %>%
  ggplot(aes(pop, resid)) +
  geom_hline(yintercept = 0, colour = "white", size = 2) +
  geom_point(alpha=2/5) +
  facet_grid(year ~ model)
```



Coefficient of Determination (r^2)

```
summary(mod1)$r.squared
```

```
## [1] 0.1260082
```

```
summary(mod2)$r.squared
```

```
## [1] 0.137638
```

```
summary(mod3)$r.squared
```

```
## [1] 0.1248974
```

```
summary(mod4)$r.squared
```

```
## [1] 0.09041305
```

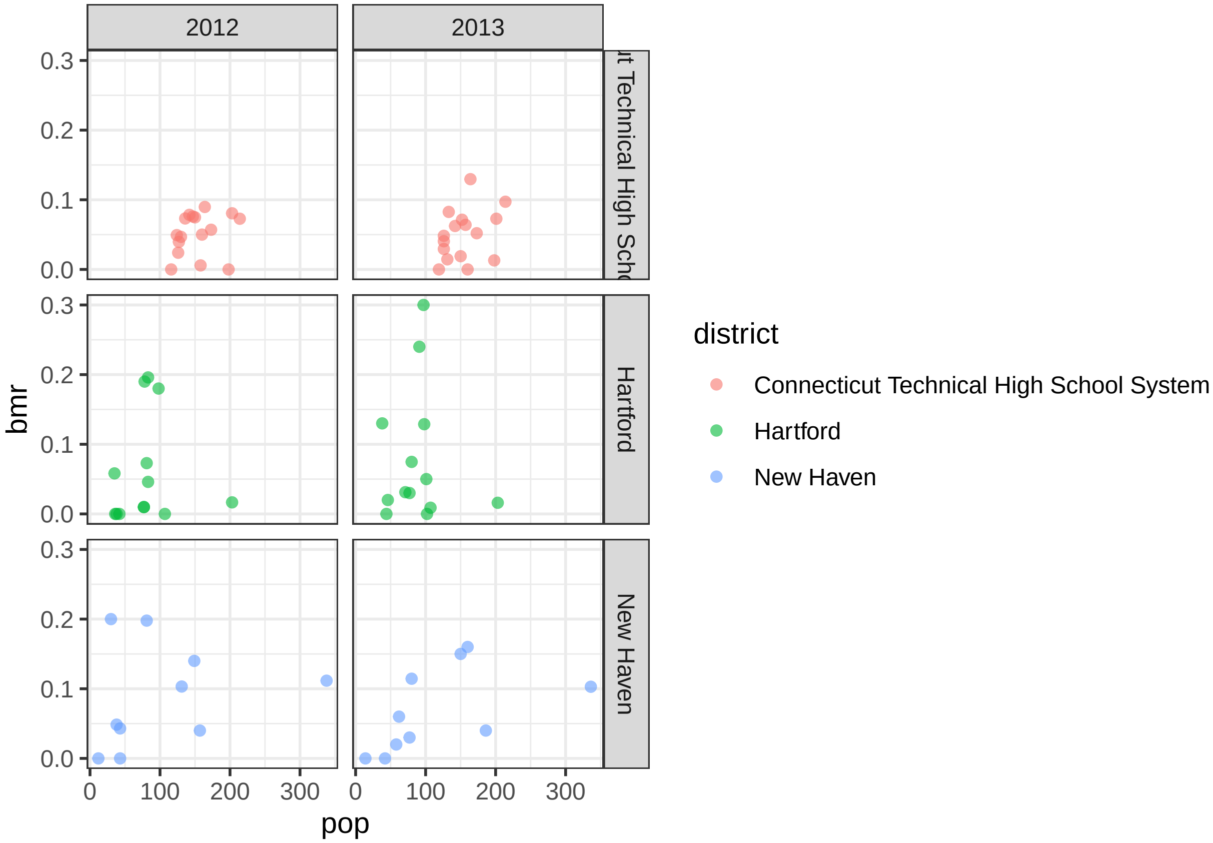
These coefficients SUCK

What if I repeated something with districts that have more than 5 schools

```r
popularDistrict <- data %>%
  group_by(district) %>%
  summarize(n=n()) %>%
  dplyr::filter(n>10) %>%
  left_join(data,by="district") %>%
  select(-n)
popularDistrict
```

```
## # A tibble: 78 x 8
##    district        school      year  t_takes part_rate perc_mb    bmr   pop
##    <chr>           <chr>       <chr>   <dbl>     <dbl>   <dbl>  <dbl> <dbl>
##  1 Connecticut Te~ Albert I P~ 2012       92        58       1 0.0058   158
##  2 Connecticut Te~ Albert I P~ 2013       88        55       0 0          160
##  3 Connecticut Te~ Bullard Ha~ 2012      127        64       0 0          198
##  4 Connecticut Te~ Bullard Ha~ 2013      129        65       2 0.013    198
##  5 Connecticut Te~ E C Goodwi~ 2012       38        30       8 0.024    126
##  6 Connecticut Te~ E C Goodwi~ 2013       62        49       6 0.0294   126
##  7 Connecticut Te~ E. T. Gras~ 2012       59        40      19 0.076    147
##  8 Connecticut Te~ E. T. Gras~ 2013       57        38       5 0.019    150
##  9 Connecticut Te~ Eli Whitne~ 2012       76        65       0 0          116
## 10 Connecticut Te~ Eli Whitne~ 2013       31        26       0 0          119
## # ... with 68 more rows
```

```r
ggplot(popularDistrict) +
  geom_point(aes(pop,bmr,color=district),alpha=3/5) +
  facet_grid(district~year) +
  theme_bw()
```


2012
2013
ut Technical High Scho
Hartford
New Haven
bmr
0.0
0.1
0.2
0.3
pop
0
100
200
300
district
Connecticut Technical High School System
Hartford
New Haven