Bias Score Development

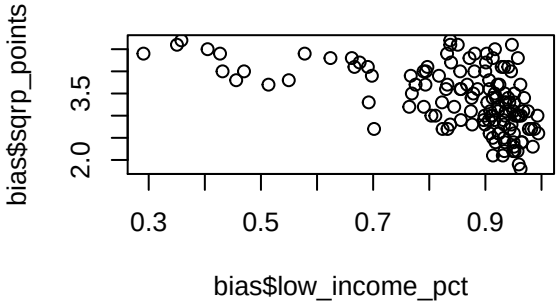
Introduction

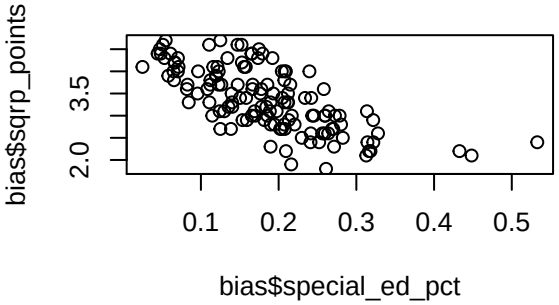
This document outlines the process of developing the SQRP bias score and deciding how to calculate it.

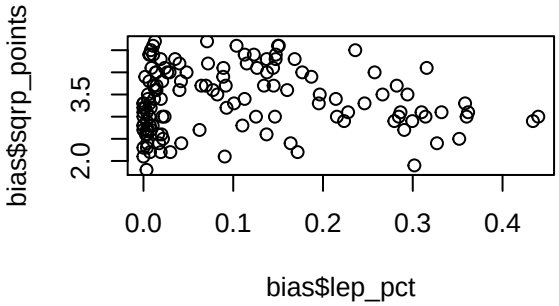
Preliminary data vizualization

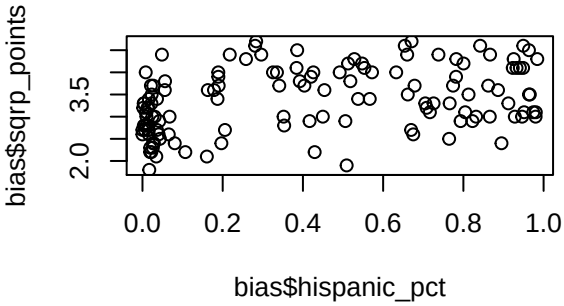
Scatterplots of our options for predictor variables show clear negative, linear relationships between SQRP points and special education percent, mobility rate percentile, and chronic truancy percentile.

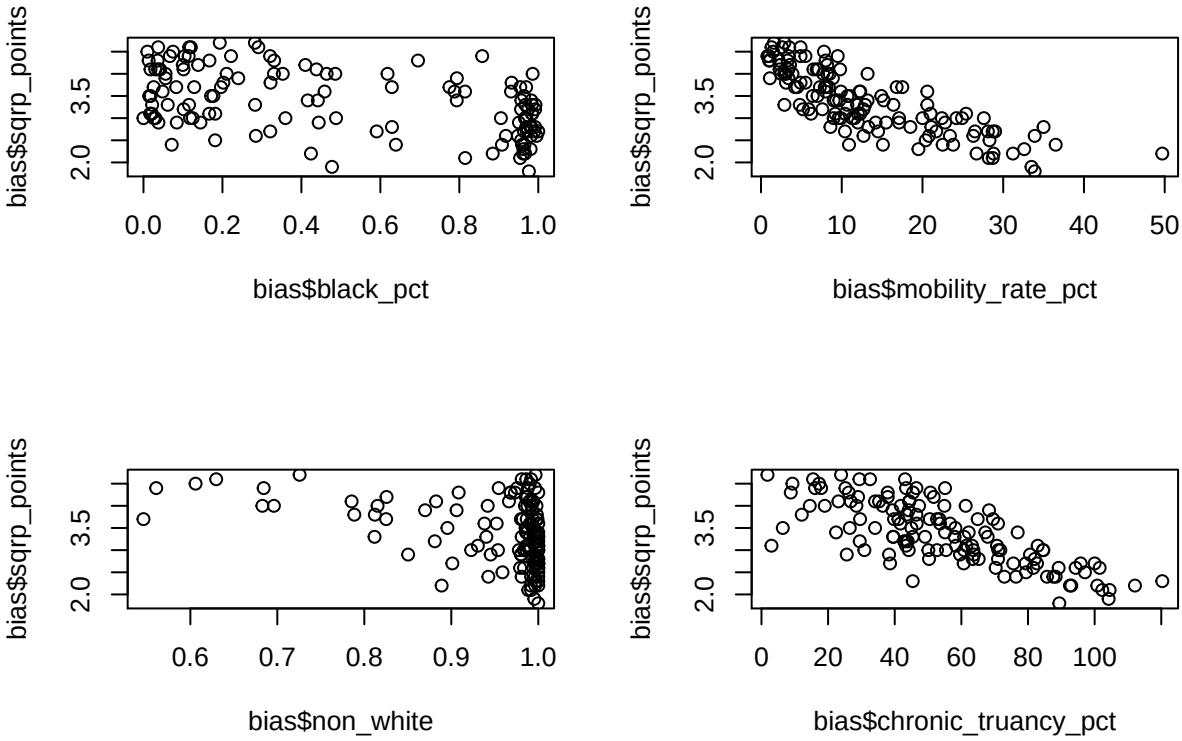
One concern with the chronic truancy percentile is that there are values up to 120.











It's important to note that these variables are closely related to each other. Let's look at the correlation among them.

```
options(digits=3)
cor(bias[c(37:38, 40:46)])
```

```
##                     mobility_rate_pct chronic_truancy_pct black_pct
## mobility_rate_pct               1.000               0.742     0.439
## chronic_truancy_pct             0.742               1.000     0.516
## black_pct                       0.439               0.516     1.000
## hispanic_pct                   -0.355              -0.395    -0.926
## lep_pct                         0.016              -0.127    -0.654
## low_income_pct                  0.477               0.515     0.254
## non_white                       0.312               0.420     0.406
## special_ed_pct                  0.724               0.640     0.367
## asian_pct                      -0.230              -0.318    -0.382
##                     hispanic_pct lep_pct low_income_pct non_white
## mobility_rate_pct         -0.3548  0.0160         0.4768    0.3123
## chronic_truancy_pct       -0.3948 -0.1267         0.5147    0.4202
## black_pct                 -0.9259 -0.6537         0.2536    0.4064
## hispanic_pct               1.0000  0.7320         0.0664   -0.0483
## lep_pct                    0.7320  1.0000         0.2947    0.0931
## low_income_pct             0.0664  0.2947         1.0000    0.8636
## non_white                 -0.0483  0.0931         0.8636    1.0000
## special_ed_pct            -0.2583  0.0836         0.5112    0.3368
## asian_pct                  0.0637  0.1125        -0.5627   -0.6884
```

```
##                     special_ed_pct asian_pct
## mobility_rate_pct           0.7245   -0.2301
## chronic_truancy_pct         0.6401   -0.3182
## black_pct                   0.3673   -0.3822
## hispanic_pct               -0.2583    0.0637
## lep_pct                     0.0836    0.1125
## low_income_pct              0.5112   -0.5627
## non_white                   0.3368   -0.6884
## special_ed_pct              1.0000   -0.2824
## asian_pct                  -0.2824    1.0000
```

Now we will use regsubsets to generate possible linear regressions using most of the variables graphed above. Mobility rate and chronic truancy will not be included, in part due to the concerns about the validity of the chronic truancy percentile, as well as because they are not as "simple" a descriptor as the other variables.

```
sub1 <- regsubsets(x=bias[,c(40:45)], y=bias[,39], names=names(bias)[c(39:45, 47)],
                   method="exhaustive",nvmax=9)
tbl <- data.frame(summary(sub1)$outmat)
tbl$adjr2 <- summary(sub1)$adjr2
tbl$cp <- summary(sub1)$cp
tbl$bic <- summary(sub1)$bic
tbl$rsq <- summary(sub1)$rsq
tbl
```

```
##          black_pct hispanic_pct lep_pct low_income_pct non_white
## 1  ( 1 )
## 2  ( 1 )         *
## 3  ( 1 )         *                   *
## 4  ( 1 )                     *       *             *
## 5  ( 1 )         *                   *             *         *
## 6  ( 1 )         *           *       *             *         *
##          special_ed_pct adjr2    cp    bic   rsq
## 1  ( 1 )              * 0.475 48.63  -77.0 0.479
## 2  ( 1 )              * 0.532 30.29  -88.3 0.539
## 3  ( 1 )              * 0.585 13.34 -100.4 0.594
## 4  ( 1 )              * 0.599  9.56 -101.2 0.611
## 5  ( 1 )              * 0.614  5.80 -102.2 0.628
## 6  ( 1 )              * 0.613  7.00  -98.2 0.631
```

Testing candidate models

Using the resutls from above, we will now train regression models using three, four, and five variables.

```
fit3 <- lm(formula = sqrp_points ~ black_pct + special_ed_pct
                 + lep_pct, data=bias)
summary(fit3)
```

```
##
## Call:
## lm(formula = sqrp_points ~ black_pct + special_ed_pct + lep_pct,
##     data = bias)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0140 -0.2866 -0.0678  0.3020  0.9970
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.802      0.117   40.94  < 2e-16 ***
## black_pct        -0.987      0.163   -6.04  1.5e-08 ***
## special_ed_pct   -3.897      0.570   -6.84  2.8e-10 ***
## lep_pct          -2.157      0.513   -4.20  4.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.458 on 129 degrees of freedom
## Multiple R-squared:  0.594,  Adjusted R-squared:  0.585
## F-statistic:   63 on 3 and 129 DF,  p-value: <2e-16
```

```
fit4 <- lm(formula = sqrp_points ~ low_income_pct + special_ed_pct
                + lep_pct + hispanic_pct, data=bias)
summary(fit4)
```

```
##
## Call:
## lm(formula = sqrp_points ~ low_income_pct + special_ed_pct +
##     lep_pct + hispanic_pct, data = bias)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9713 -0.2620 -0.0386  0.2737  0.9715
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.844      0.246   19.67  < 2e-16 ***
## low_income_pct   -1.259      0.323   -3.89  0.00016 ***
## special_ed_pct   -3.420      0.595   -5.74  6.4e-08 ***
## lep_pct          -1.757      0.555   -3.16  0.00195 **
## hispanic_pct      0.991      0.185    5.37  3.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.45 on 128 degrees of freedom
## Multiple R-squared:  0.611,  Adjusted R-squared:  0.599
## F-statistic: 50.3 on 4 and 128 DF,  p-value: <2e-16
```

```
fit5 <- lm(formula = sqrp_points ~ low_income_pct + special_ed_pct
                + lep_pct + black_pct + non_white, data=bias)
summary(fit5)
```

```
##
## Call:
## lm(formula = sqrp_points ~ low_income_pct + special_ed_pct +
##     lep_pct + black_pct + non_white, data = bias)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9920 -0.2518 -0.0239  0.2839  0.9752
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.552      0.515    6.89  2.3e-10 ***
## low_income_pct   -2.163      0.656   -3.30   0.0013 **
## special_ed_pct   -2.950      0.619   -4.77  5.0e-06 ***
## lep_pct          -1.903      0.584   -3.26   0.0014 **
## black_pct        -1.109      0.190   -5.84  4.2e-08 ***
## non_white         3.106      0.965    3.22   0.0016 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.442 on 127 degrees of freedom
## Multiple R-squared:  0.628,  Adjusted R-squared:  0.614
## F-statistic: 42.9 on 5 and 127 DF,  p-value: <2e-16
```

In all of these models, all of the predictor variables are highly significant.

Cross validation was used to compare the mean squared error, in order get a better idea of which of these models is best. For the sake of brevity, this output is not included; the MSE of all three models is nearly equal.

As the MSE is almost equal for these three models, it's not an useful metric in this decision.

Given the high correlation among the predictor variables, it's also important to check for multicollinearity. We will use the variance inflation factor (VIF) to measure this. VIF has a minimum of 1 and no upper bound. A VIF higher than 4 is cause for concern.

```
ols_vif_tol(fit3) #all VIF under 3
```

```
##        Variables Tolerance  VIF
## 1      black_pct     0.393 2.54
## 2 special_ed_pct     0.682 1.47
## 3        lep_pct     0.452 2.21
```

```
ols_vif_tol(fit4) #all VIF under 3
```

```
##        Variables Tolerance  VIF
## 1 low_income_pct     0.675 1.48
## 2 special_ed_pct     0.603 1.66
## 3        lep_pct     0.372 2.68
## 4   hispanic_pct     0.361 2.77
```

```
ols_vif_tol(fit5) #several VIF 3-6+
```

```
##        Variables Tolerance  VIF
## 1 low_income_pct     0.158 6.33
## 2 special_ed_pct     0.538 1.86
## 3        lep_pct     0.325 3.08
## 4      black_pct     0.271 3.69
## 5      non_white     0.182 5.49
```

Choosing a model

The model with five predictor variables has VIFs above 4, so we will no longer use it. Further assumption checking and visualization of the other two models was performed. Again, for the sake of brevity, this output is not included.

A final decision

There are no strong reasons to exclude either of our two candidate models. Both do show that two points (with index 19 and 167) may have undue leverage in the models (as indicated by Cook's distance). However, excluding them from the dataset and rerunning the analysis does not change the output in any significant way, so we are not concerned with leaving them in.

Which model, then, should we use? We are actually going to turn to a third option: using percent low income, percent special education, and percent English language learners. As we will see, under the current policy, percent low income is not a significant predictor in this model. However, we are seeking to measure how well the demographics of a school predict its SQRP points, and not fit the best model possible. Furthermore, using these three predictors eliminates the question of why some race percentages (e.g., percent of students who are Hispanic) were included, and others no. Let's run through some final assumption checks on this model, though keep in mind that in the web app, we will change the values of the dependent variable (the SQRP points).

```
cor(bias[c(42, 43, 45)])
```

```
##                lep_pct low_income_pct special_ed_pct
## lep_pct         1.0000          0.295         0.0836
## low_income_pct  0.2947          1.000         0.5112
## special_ed_pct  0.0836          0.511         1.0000
```

```
fit_final <- lm(formula = sqrp_points ~ low_income_pct + special_ed_pct
                + lep_pct, data=bias)
summary(fit_final)
```

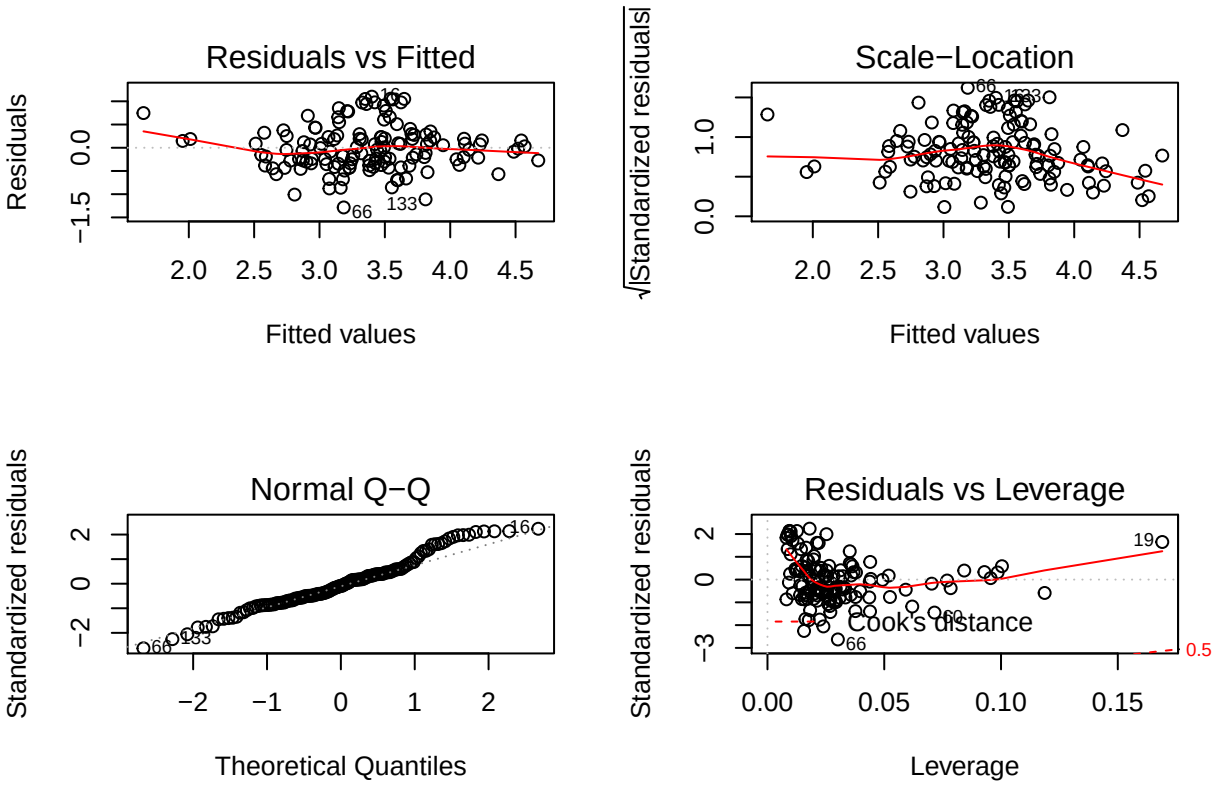
```
##
## Call:
## lm(formula = sqrp_points ~ low_income_pct + special_ed_pct +
##     lep_pct, data = bias)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2853 -0.3360 -0.0436  0.2316  1.0995
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.240      0.259   20.23  < 2e-16 ***
## low_income_pct   -1.233      0.356   -3.46  0.00073 ***
## special_ed_pct   -4.770      0.595   -8.02  5.7e-13 ***
## lep_pct           0.531      0.392    1.35  0.17816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.496 on 129 degrees of freedom
## Multiple R-squared:  0.524,  Adjusted R-squared:  0.513
## F-statistic: 47.3 on 3 and 129 DF,  p-value: <2e-16
```

```r
layout(matrix(c(1,2,3,4),2,2))

gvmodel_final <- gvlma(fit_final)
summary(gvmodel_final)
```

```
##
## Call:
## lm(formula = sqrp_points ~ low_income_pct + special_ed_pct +
##     lep_pct, data = bias)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2853 -0.3360 -0.0436  0.2316  1.0995
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.240      0.259   20.23  < 2e-16 ***
## low_income_pct   -1.233      0.356   -3.46  0.00073 ***
## special_ed_pct   -4.770      0.595   -8.02  5.7e-13 ***
## lep_pct           0.531      0.392    1.35  0.17816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.496 on 129 degrees of freedom
## Multiple R-squared:  0.524,  Adjusted R-squared:  0.513
## F-statistic: 47.3 on 3 and 129 DF,  p-value: <2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = fit_final)
##
##                      Value p-value                Decision
## Global Stat        2.29944   0.681 Assumptions acceptable.
## Skewness           1.90909   0.167 Assumptions acceptable.
## Kurtosis           0.00335   0.954 Assumptions acceptable.
## Link Function      0.03102   0.860 Assumptions acceptable.
## Heteroscedasticity 0.35597   0.551 Assumptions acceptable.
```

```r
plot(fit_final)
```




```
ols_vif_tol(fit_final) #All VIF under 1.5
```

```
##        Variables Tolerance  VIF
## 1 low_income_pct     0.675 1.48
## 2 special_ed_pct     0.734 1.36
## 3        lep_pct     0.907 1.10
```

Other avenues pursued

Other methods of analysis were considered. For example, in order to capture the racial makeup of a school in a more concise way, we considered using principal components. However, the low number of predictors meant that little was gained with this approach. It also makes interpretation and visualization of the results more difficult. Therefore, we decided to use simple linear regression instead.

Ordinal logistic regression was also considered, in which the dependent variable would be the school's rating instead of points. This had two major drawbacks: one, there is no equivalent of an R2 value defined for OLR. While "psuedo-R2" measures have been developed, we would lose transparency and ease of interpretation. Second, this approach could dampen the impact of changes to SQRP weights, as schools could move around within the bounds of a level without any apparent changes to the bias score.

For the sake of brevity, the code associated with exploring these options is not included.