Original Model

To start we show our original model that we started with last week, before adding and omitting variables (but including our dummy variables). This is shown below, in Model 1:

Model 1

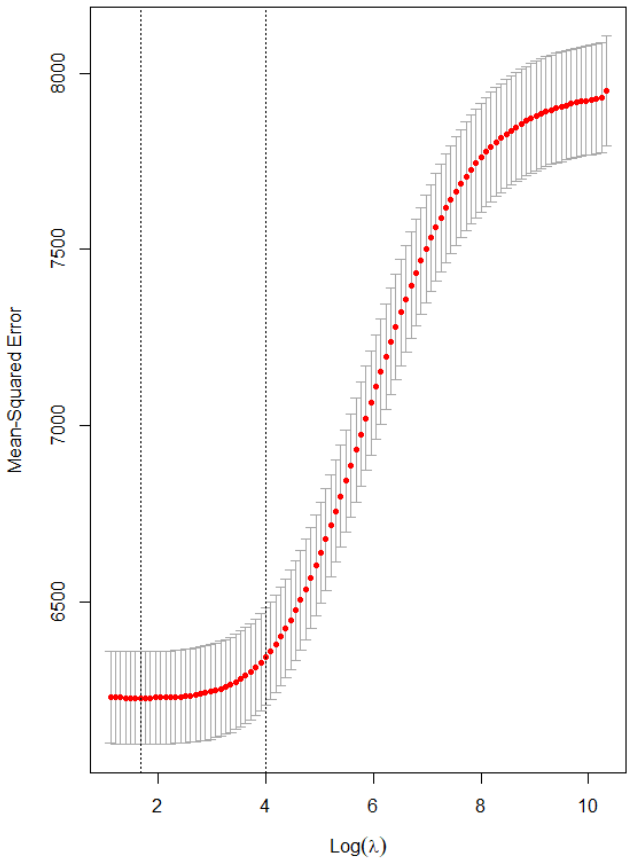
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             378.662981  18.998202  19.932  < 2e-16 ***
male                    -15.119220   2.687626  -5.625 2.00e-08 ***
raceeth.f2               60.050964  15.645886   3.838 0.000126 ***
raceeth.f3               -5.243507  14.366468  -0.365 0.715147
raceeth.f4               26.479657  14.220641   1.862 0.062682 .
raceeth.f5               41.436651  15.392327   2.692 0.007137 **
raceeth.f6               53.728960  20.454687   2.627 0.008660 **
raceeth.f7               62.814078  13.829361   4.542 5.76e-06 ***
preschool                -1.125057   3.008742  -0.374 0.708480
expectBachelors          56.805241   3.633769  15.633  < 2e-16 ***
motherHS                  3.618382   5.165591   0.700 0.483677
motherBachelors          11.064621   3.346652   3.306 0.000956 ***
motherWork               -1.394464   3.010339  -0.463 0.643233
fatherHS                 10.076380   4.723582   2.133 0.032980 *
fatherBachelors          17.959631   3.452365   5.202 2.09e-07 ***
fatherWork                3.931537   3.766085   1.044 0.296591
selfBornUS                2.568901   6.056676   0.424 0.671488
motherBornUS             -8.833656   5.096202  -1.733 0.083119 .
englishAtHome            12.077792   5.851679   2.064 0.039095 *
computerForSchoolwork    23.232637   4.944363   4.699 2.72e-06 ***
read30MinsADay           33.431804   2.918570  11.455  < 2e-16 ***
minutesPerWeekEnglish     0.016249   0.009192   1.768 0.077189 .
studentsInEnglish        -0.101253   0.196507  -0.515 0.606402
schoolHasLibrary         -2.783196   7.714895  -0.361 0.718304
publicSchool            -20.062051   5.696467  -3.522 0.000434 ***
urban                    -0.537920   3.383580  -0.159 0.873694
schoolSize                0.008123   0.001846   4.400 1.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.59 on 3377 degrees of freedom
Multiple R-squared:  0.2868,    Adjusted R-squared:  0.2813
F-statistic: 52.24 on 26 and 3377 DF,  p-value: < 2.2e-16
```

In the model above, it is shown that one of the betas is not 0. This means that we cannot reject the null hypothesis, and accept the alternative. The adjusted R-squared shows that it is .2813, which is low. The is means that 28.13% of variation in y in explained by x. The t tests have 12 independent variables that are over .05 (that don’t pass): raceeth.f3,raceeth.f4, preschool l, mothersHS, motherWork, fatherWork, SelfBornUS,motherBornUS,minutesperweekEnglish,StudentsInEnglish,schoolhaslibrary,and urban. Raceeth will be left in, since others pass. The other independent variables that do not pass the test, will be taken out.

Ridge Regression

Now we will use ridge regression to analyze the multicollinearity that is occurring in the model. In ridge regression, we are trying to shrink the additional term back to 0. When using Rstudio, the alpha is set to 0 with ridge regression. Below in Model 2 is the ridge trace, along with model 3 that displays its outputs.

Model 2



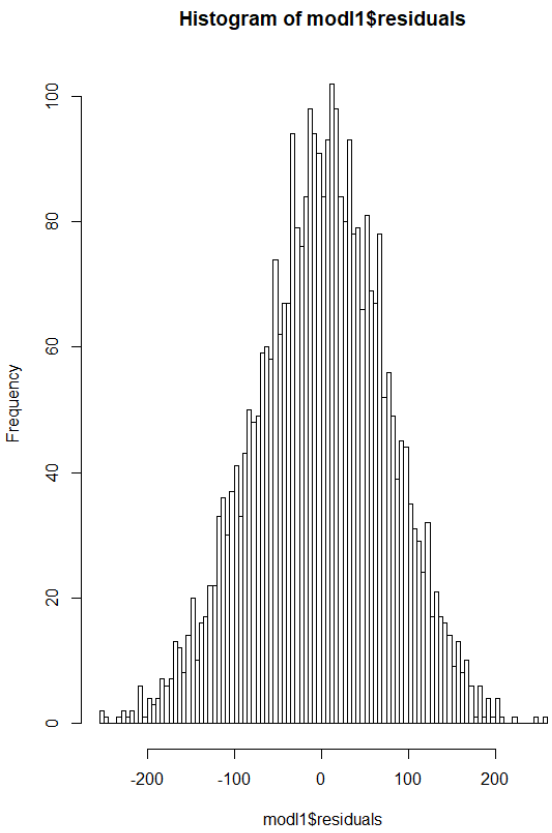
Model 3

	1
(Intercept)	406.97945534
preschool	-1.96184777
expectBachelors	55.78942101
motherHS	2.89330955
motherBachelors	11.58264394
motherWork	-2.14754663
fatherHS	13.87457855
fatherBachelors	22.69811312
fatherWork	7.79074400
selfBornUS	-1.21026972
motherBornUS	-1.72381365
fatherBornUS	4.47528737
englishAtHome	14.10088616
computerForSchoolwork	29.05907591
read30MinsADay	34.72650010
minutesPerWeekEnglish	0.01814557
studentsInEnglish	0.10095252
schoolHasLibrary	-6.57385074
publicSchool	-24.08741090
urban	-8.61277576
schoolsize	0.00654302

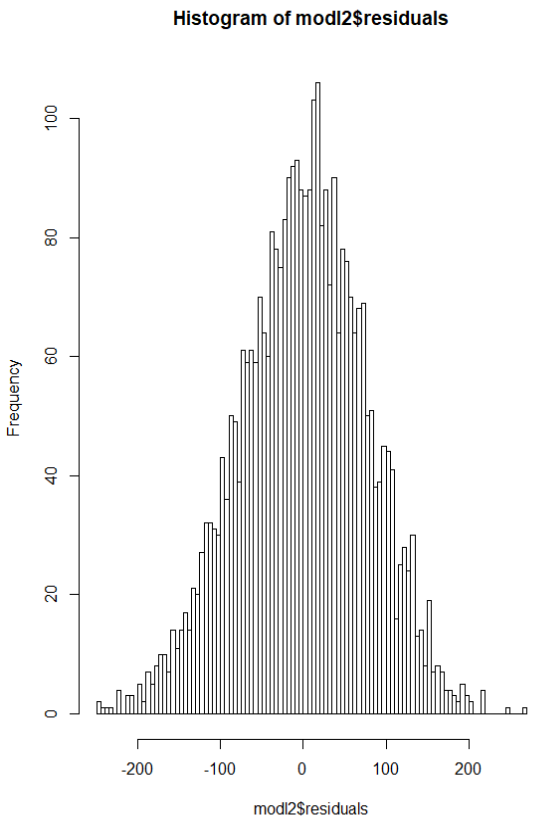
Above, in the ridge trace model (Model 2) we can see that when log lambda (the x axis) equals 0, we are just left with minimizing the sum of the square of the errors, and that gives us somewhere close to 6000 on the y-axis. As the lambda increases, we are giving more weight to the penalty function, and can see that when lambda hits 4, the errors start to increase. As we look at the pearson values in our model 3, we can see that there some values that we have we may consider taking out such as schoolSize, minutesPerWeekEnglish, selfBornUS, and motherBornUS due to potential multicollinearity.

After removing potential variables for multicollinearity, we can compare the distribution of our residuals from the original number of variables, to the model that took out variables that looked to be correlated. This is shown below in Models 4 and 5:

Model 4



Model 5

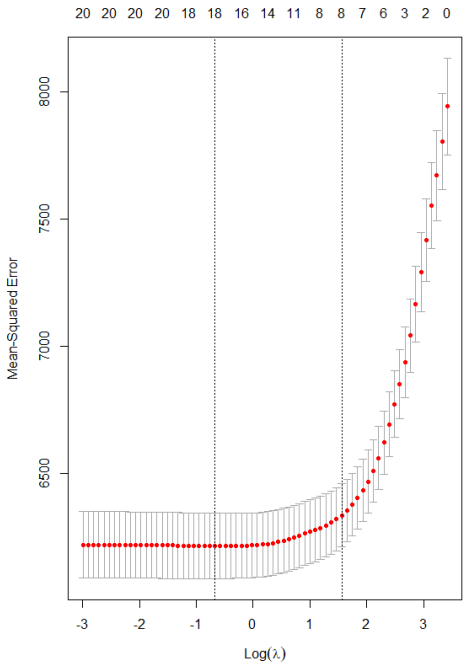


As shown above there are a couple of outliers on the right that are taken out. The outliers had the potential to be influential points in our testing. These models also show that that the distribution of the residuals is both normal. We will compare the final models of ridge and lasso regression further on.

Lasso Regression

Next, we are going to perform lasso regression. Lasso regression is a continuous feature selection technique, as opposed to stepwise which is discrete. Lasso also, reduces bias since it identifies the set of non-zero coefficients, and then fits an unrestricted linear model to the selected set of features.

Model 6



Model 7

	1
(Intercept)	404.079289062
preschool	-0.845054372
expectBachelors	57.832041114
motherHS	1.028096136
motherBachelors	10.766437099
motherWork	-0.939803785
fatherHS	13.712316635
fatherBachelors	23.538661747
fatherWork	6.578996366
selfBornUS	.
motherBornUS	.
fatherBornUS	2.423843148
englishAtHome	13.482495101
computerForSchoolwork	28.987923802
read30MinsADay	35.558025784
minutesPerWeekEnglish	0.015743625
studentsInEnglish	0.020690323
schoolHasLibrary	-3.579960664
publicSchool	-22.045266410
urban	-7.374966331
schoolSize	0.005988656

Above in Model 6, you can see that once Log lambda hits 2, the errors start to incerease. In Model 7, since lasso regression is a type of feature selection, you can see that it has taken self BornUS and motherBornUS out of the model. The reason there is a difference in the ridge trace from Ridge regression to Lasso regression is that ridge regression uses an abosulute value for its regulization term while Lasso penalizes high values in the coefficientents of beta.

Ridge and Lasso Regression Final Models

Model 8 (using ridge regrssion)

```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    370.97053   18.31930  20.250  < 2e-16 ***
male                                           -14.89027    2.69472  -5.526 3.53e-08 ***
raceeth.fAsian                                  67.66680   15.38409   4.398 1.12e-05 ***
raceeth.fBlack                                  -3.31930   14.40311  -0.230  0.81775
raceeth.fHispanic                               32.85739   14.15772   2.321  0.02036 *
raceeth.fMore than one race                     44.39038   15.42810   2.877  0.00404 **
raceeth.fNative Hawaiian/Other Pacific Islander 61.20516   20.32717   3.011  0.00262 **
raceeth.fWhite                                  64.35434   13.86863   4.640 3.61e-06 ***
preschool                                       -0.83918    3.01309  -0.279  0.78064
expectBachelors                                 57.90035    3.63854  15.913  < 2e-16 ***
motherHS                                         3.32162    5.15633   0.644  0.51950
motherBachelors                                 11.73981    3.34791   3.507  0.00046 ***
motherWork                                      -1.75781    3.01644  -0.583  0.56010
fatherHS                                         9.71295    4.72952   2.054  0.04008 *
fatherBachelors                                 19.56565    3.44275   5.683 1.43e-08 ***
fatherWork                                       3.79790    3.77300   1.007  0.31420
englishAtHome                                    6.91738    5.03032   1.375  0.16918
computerForSchoolwork                           25.04870    4.94268   5.068 4.24e-07 ***
read30MinsADay                                  32.99196    2.92494  11.280  < 2e-16 ***
studentsInEnglish                                0.07873    0.19351   0.407  0.68413
schoolHasLibrary                                -0.50539    7.71801  -0.065  0.94779
publicSchool                                   -10.16362    5.23818  -1.940  0.05243 .
urban                                            5.02952    3.15324   1.595  0.11080
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.84 on 3381 degrees of freedom
Multiple R-squared:  0.2812,    Adjusted R-squared:  0.2765
F-statistic: 60.12 on 22 and 3381 DF,  p-value: < 2.2e-16
```

Above in Model 8 is the final model using ridge regression. It is shown that one of the betas is not 0. This means that we cannot reject the null hypothesis, and accept the alternative. The adjusted R-squared shows that it is .2765, which is low. The is means that 27.65% of variation in y in explained by x. The t tests have 12 independent variables that are over .05 (that don’t pass): urban, publicSchool(we may leave that in there since it is close), studentsinEnglish, read30minsAday,motherword, and preshool. The independent variables that do not pass the t test, should be considered to be taken out.

Model 9 (using Lasso Regression)

```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   376.823161  18.546070  20.318  < 2e-16 ***
male                                          -15.120897   2.686558  -5.628 1.97e-08 ***
raceeth.fAsian                                 65.036782  15.347127   4.238 2.32e-05 ***
raceeth.fBlack                                 -4.766507  14.364879  -0.332 0.740048
raceeth.fHispanic                              29.111848  14.141373   2.059 0.039606 *
raceeth.fMore than one race                    42.281878  15.386103   2.748 0.006027 **
raceeth.fNative Hawaiian/Other Pacific Islander 58.399754  20.276229   2.880 0.003999 **
raceeth.fWhite                                 62.675485  13.831186   4.531 6.06e-06 ***
preschool                                      -1.153711   3.005516  -0.384 0.701103
expectBachelors                                57.040171   3.631399  15.707  < 2e-16 ***
motherHS                                        2.761941   5.141430   0.537 0.591170
motherBachelors                                10.813714   3.342714   3.235 0.001228 **
motherWork                                     -1.332118   3.010179  -0.443 0.658128
fatherHS                                        9.568994   4.715058   2.029 0.042490 *
fatherBachelors                                18.343506   3.442956   5.328 1.06e-07 ***
fatherWork                                      4.171663   3.761938   1.109 0.267546
englishAtHome                                   8.556240   5.025897   1.702 0.088767 .
computerForSchoolwork                          23.729118   4.935611   4.808 1.59e-06 ***
read30MinsADay                                 33.357660   2.918624  11.429  < 2e-16 ***
minutesPerWeekEnglish                           0.016047   0.009193   1.746 0.080969 .
studentsInEnglish                              -0.098197   0.196401  -0.500 0.617122
schoolHasLibrary                               -2.680441   7.714833  -0.347 0.728282
publicSchool                                  -20.643411   5.684240  -3.632 0.000286 ***
urban                                          -0.442937   3.382893  -0.131 0.895835
schoolSize                                      0.008342   0.001842   4.530 6.11e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.6 on 3379 degrees of freedom
Multiple R-squared:  0.2862,    Adjusted R-squared:  0.2811
F-statistic: 56.45 on 24 and 3379 DF,  p-value: < 2.2e-16
```

Above in Model 9 is the final model using ridge regression. It is shown that one of the betas is not 0. This means that we cannot reject the null hypothesis, and accept the alternative. The adjusted R-squared shows that it is .2811, which is low. The is means that 28.11% of variation in y in explained by x. The t tests have 12 independent variables that are over .05 (that don't pass): urban, publicSchool(we may leave that in there since it is close), studentsinEnglish, read30minsAday,motherword, and preshool. The independent variables that do not pass the t test, should be considered to be taken out.

Conclusion and Comparison About Using Lasso and Ridge Regression

The two final models using ridge and lasso regression differed. In general, ridge regression differs from lasso since it puts an absolute value over its regularization term. To get the best outcome, we would likely use a combination of ridge and lasso regression, as one is not superior to the other. In this case, lasso gave us a higher adjusted R squared, while we took out more variables with ridge regression.