Statistical Linear Models Final Project

Anna Figge, Jaden Sides

December 2025

1 Abstract

The SAT is known as a metric to judge student performance for college admissions, but it has been demonstrated widely that it fails to accurately judge student ability and is correlated with numerous demographic factors. But beyond judging student performance, SAT scores often have impacts on people's lives, from what college students go to to how much funding schools get. We investigate the effect of school and district level factors, both demographic data and funding data, on the average SAT math and writing scores of those schools, using several different kinds of linear models. Through this, we identified significant correlation between some demographic factors, including the percentage of students on free/reduced price meals, as well as school-based factors, such as school size, and SAT English and Math scores. We also fit multiple linear models to the data and performed model selection, achieving an RMSE of 34.9 for Math and RMSE of 33.0 for Writing SAT scores on a test set, while observing that the relative performance of our models differ greatly depending on assignment of training, testing, and validation sets.

2 Introduction

The widespread use of the SAT as a metric for student learning and potential for future success is well-known. On the individual scale, it can determine college admissions, which influences access to a variety of careers. And on the school level, SAT scores are often used to judge student and teacher, which could further influence student performance through resource allocation.

Many previous analyses have been on the scale of individuals. For instance, in 2013, Dixon-Román et. al. published a paper analyzing SAT scores between Black students and white students, demonstrating how the impact of income on SAT scores differs between students of varying racial backgrounds using structural equation modelling (Dixon-Román et. al., 2013). Everson and Millshap also used structural equation modelling in their 2004 investigation of school-based factors in SAT scores, only examining school factors and leaving out demographic factors (Everson and Millshap, 2004). We wish to examine a combination of school and demographic factors, using quantitative school-level data rather than categorical individual data.

In this project we seek to explore the relationship between demographic information and the average SAT score of high schools in California. In particular, we examine variables related to demographics such as race distribution, English learners, and students on free/reduced-price meals, and school characteristics such as per-student budget, budget distribution, teacher salary, and school size.

Our research questions are: 1. What school-level demographic and organizational data is significantly correlated with SAT score? 2. How can we most effectively model the average SAT score of a high school using demographic and organizational data?

Through this, we hope to better understand what characteristics of a high school are most predictive of SAT score, which can be used to better understand the uneven distribution of SAT scores.

3 Dataset

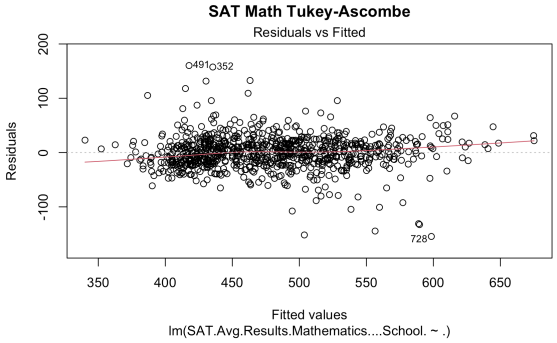
We used data from the Ed-Data partnership which stores school and district level data from California public schools reported by the California District of Education. The data is found on the web at https://www.ed-data.org/Comparisons?compType=schools. We used data from the 2015-2016 school year, as this was the most recent year that provided SAT score data. We chose the average school leveled math and writing SAT scores as our response variables, and picked as predictor variables the census day enrollment (number of students at the school), % of English learners, % of members of each racial demographic (the categories included American Indian or Alaska Native, Asian, Black or African American, Filipino, Hispanic or Latino, Native Hawaiian or Pacific Islander, White, Two or More Races, or None Reported; because these columns are nearly but not perfectly collinear we assume that people were allowed to select multiple races). We also used the % of students on free or reduced price meals, the student teacher ratio, average number of years teaching of teachers at that school, and suspension rate. We additionally downloaded the county column to use for mixed effect analysis and the school district column to join on the district data.

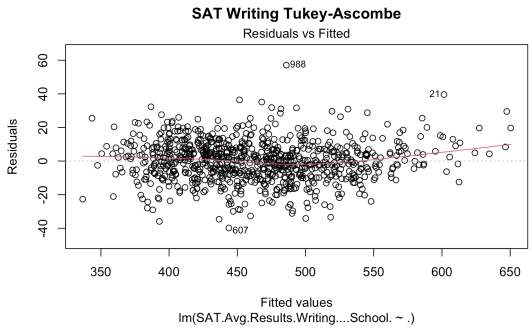
We wanted to include financial data, which was not reported at school level, so we used Ed-Data's school district data (found at https://www.ed-data.org/Comparisons?compType=districts) to download the expenditures per average daily attendance, the percentage of funding used for each category of spending, and the average teacher salary at the school district level.

To process this data, we filtered only high schools, as these are the only category of schools where average SAT score is meaningful. Then, we joined the district level data with the school level data using the district column of the school data such that each school is paired with the financial data from its district (which means schools in the same district will have identical values for all financial variables). We then did some data cleaning (making sure to interpret NaNs properly and removing special characters from column names). Some average SAT scores were reported as 0, which is outside the range of valid SAT scores, so we set these to NaN. We then dropped all schools with NaN values for any of the variables, which left us with 1065 data points. We then randomly split the data (using a seed of 790 generated by Google's random number generator) into a training, validation, and test set using 80%, 10%, and 10% of the data respectively.

Briefly examining the data, we see that the school average SAT scores form a roughly normal distribution

with mean around 475; about what we would expect from SAT scores. Also in this investigation, as a preliminary measure before more in depth analysis, we ran a multiple regression predicting each SAT score type(simple lm call, on only the training data) with all predictor variables and confirmed from the residual plots that the SAT math and SAT writing scores are mostly linear with relation to the predictors and that the errors are homoskedastic the training dataset.





4 Methods

4.1 Coefficient Analysis

To properly analyze the coefficients of each variable, we did not perform variable selection, as that could introduce omitted variable bias and a selection step would interfere with the validity of statistical testing. We did however use VIF to remove collinear predictors, since collinearity destabilizes the estimate of variable coefficients in multiple regression. The VIF of predictor j is defined as $\text{VIF}_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination of predicting predictor j from all other predictors. In general, a VIF over about 10 is considered problematic. When we computed VIF on the full predictor set, we found that the racial percentages were so collinear that the highest VIF was over 100,000. We removed the variable with the highest VIF and recalculated the VIF of each predictor on the remaining variable set, repeating until the maximum VIF was less than 10. Only one variable needed to be removed, the percentage of Hispanic or Latino students, and this resulted in all remaining VIFs being less than 3.

Then, we performed two OLS multiple linear regressions to predict average math SAT score and average writing SAT score from all of the remaining variables. This method assumes the prediction function is linear of the form $y_i = f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ where p is the number of predictor variables. It estimates

β by minimizing the least squares objective, such that

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}]) = (X^\top X)^{-1} X^\top y$$

where n is the number of observations in the training dataset, X is the design matrix, and $y = [y_1, \cdots, y_n]^\top$ is the vector of outcomes. A nice feature of OLS is that the $\hat{\beta}$ it outputs is guaranteed to be unbiased.

We also performed a mixed effect linear regression, treating county as a random effect. Treating county as a random effect essentially means that the model has categorical coefficients for county but does not try to estimate these coefficients, instead incorporating them into the error structure. This allows the model to account for county level confounders without attempting to explicitly estimate differences by county (which would lead to an unnecessarily inflated number of coefficients given that this question is not of interest to us). Essentially, for counties $1, \cdots, K$ where (x_{ij}, y_{ij}) is observation j within county i, we assume $y_{ij} = \mu + \beta x_{ij}^\top + \tilde{\epsilon}_{ij}$ where $\tilde{\epsilon}_{ij} = \alpha_i + \epsilon_{ij}$ for a county specific α_i and a general error ϵ_{ij}. Then under the assumptions that the random intercepts (α_i) are independent of the covariates and each county has common variance, and the errors without county are homoskedastic, this imposes a particular symmetrical block diagonal structure on the covariance matrix ($\Sigma(X)$), the parameters of which can be estimated using restricted maximum likelihood in a GLS estimator, allowing us to use the GLS estimate $\hat{\beta} = (X^\top \Sigma(X)^{-1} X) X^\top \Sigma(X)^{-1} y$. Using this covariance structure and the GLS rather than OLS estimator gives more accurate confidence intervals; neglecting mixed effects increases the probability of false positives in coefficient estimates (Yu et al, 2022), and using random effects tends to lead to tighter confidence intervals that incorporating county as a fixed effect (a categorical variable in the model the coefficients of which must be estimated). Note that when we used the mixed effect model to predict on the validation and test sets, we used only the fixed effect coefficients for any counties that were not present in the training data.

We computed this mixed effect model on the training data using the VIF pruned variables to predict both math and writing SAT scores. We then used an ANOVA test between the trained mixed effect model and the trained OLS model (which excludes county) for each SAT score type and found that the mixed effect model was significantly better than the OLS model for predicting both math and writing SAT scores ($p < 0.001$ for both). Due to the GLS, a likelihood ratio test rather than the more standard F test was performed, but the interpretation is the same. Additionally the ICC (the ratio of the variance between counties to the total variance) was around 0.1 for both models, and an ICC greater than 0 is a sign that a mixed effect

model should be used (Yu et al, 2022). Therefore, we did analysis of the coefficients based on the confidence intervals and p values reported by the mixed effects model, which we presume to be more accurate.

However, we cannot simply use the reported p values for the null hypothesis that any given coefficient is 0. Since there are 24 variables after VIF filtering (25 coefficients including the intercept), we run into the multiple testing problem; if we use a confidence threshold of 0.05 we have a very high chance of making at least one type 1 error. To solve this, we use two methods of adjusting p values for multiple hypothesis testing, the Holm adjustment and the Hommel adjustment.

The Holm adjustment first ranks all m hypotheses by their p values from smallest (p_i) to largest (p_m), and takes α to be the significant threshold. Then for the ith ordered hypothesis it computes the adjusted significance threshold $\alpha_i' = \frac{\alpha}{m-i+1}$. It rejects the ith null hypothesis if $p_i < \alpha_i'$. It repeats until it finds a hypothesis where $p_i > \alpha_i'$, and then it stops and fails to reject the null hypothesis for that hypothesis and all hypotheses with higher p values. This guarantees that the type 1 error rate is still α (Chen et al, 2017).

The Hommel adjustment lets H be the global hypothesis that all m null hypotheses are true. It similarly orders the p values; and it says that H will be rejected (ie we assume that at least one null hypothesis is false) if $p_i < i\alpha/m$ for at least one $i \in \{1,2,\cdots,m\}$. Then we let j be the largest subset of the m hypotheses for which the global hypothesis H built on that subset is true (ie j is the size of the largest subset of hypothesis for which $p_i > i\alpha/j$ for all $i \in \{1,2,\cdots,j\}$). If j does not exist, all null hypotheses are rejected; otherwise all the null hypothesis with $p_i < \alpha/j$ are rejected and we fail to reject the rest (Chen et al, 2017)..

Our mixed effect multiple regressions gave us for each coefficient the p-value that that coefficient was nonzero (based on HC2 standard errors). We then calculated the adjusted p values using both Holm and Hommel adjustments (with $\alpha = 0.05$) using the p.adjust function of R's stats package. We found that the Holm and Hommel adjustments, while they did not agree on the exact values of the adjusted p-values, agreed entirely on which coeficients were significantly different from 0 with the $\alpha = 0.05$ threshold. Thus, we considered the set of coefficient with Holm and Hommel adjusted p values below 0.05 to be significantly different than 0 while failing to reject the null hypothesis that the coefficient = 0 for all other coefficients.

4.2 Model Selection

To optimize the performance of our predictions of SAT score, we tried several different types of model. One was the basic OLS with all predictors (except for the VIF pruning) as described above. Another was OLS using only variables selected by forward feature selection. We also used a LASSO and group LASSO

model. Finally, we used the mixed effects model described above. For all of these, we trained two models on the training dataset, one to predict SAT math scores and one to predict SAT writing scores, and then calculated the MSE of each of those models on the validation set. We report the test set performance of only the model which did best on the validation set.

4.3 Forward Feature Selection

Forward feature selection (FFS) is a method for choosing a sparser set of variables to use for modeling. While using fewer variables will always increase the RSS of a model, it can reduce out of sample error by removing spurious correlations. To account for this, we use Mallows Cp criterion which is defined as $C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2)$ where n is the number of observations, p is the number of variables, and $\hat{\sigma}$ is the estimated variance of the errors. This essentially penalizes models with more predictors to approximate out of sample performance. FFS uses this criterion and a greedy algorithm to estimate the best predictor set for the model to use. This greediness means that it is likely to miss the best possible set of variables, unlike best subset selection, but with our number of variables (25) there are 2^{25} subsets which makes best subset selection computationally infeasible.

FFS starts with an intercept only model and computes and saves the C_p for that model. Then, it tries all of the models with 1 predictor by adding all the remaining predictors one at a time, and saves the model with the best RSS (which is also guaranteed to be the model with the best C_p, as p is the same for all these models), calculating and saving that model's C_p. It repeats this process, finding the best variable to add to the existing set of variables and calculating the C_p, until a C_p is calculated for every possible number of variables. Then the set of variables that produced the model with the lowest C_p is chosen.

We performed FFS using the regsubsets function from the leaps package. We performed separate feature selection for the SAT math and SAT writing models, and fit the feature selection on only the training data. Because FFS can be unstable in the presence of highly collinear predictors, VIF pruning was performed first as described above (dropping the % Hispanic predictor).

4.4 LASSO

We used the LASSO method to get a model that is sparser than the full linear model which allows for more interpretable results than the full OLS model, and prunes potentially noisy variables that could harm

out of sample prediction. The equation of LASSO is as follows:

$$\hat{\beta}_{\lambda}^{L} = \text{argmin}_{\beta=(\beta_0,\beta_1,\ldots,\beta_p)^\top} \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}])^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

We can recognize that the first summation is equivalent to that of OLS. Where LASSO differs from OLS is in the second summation, where high values of β are penalized based on their L_1 norm. The regularization that comes from this penalty allows LASSO to set predictors equal to zero, leading to a sparser model.

Through selecting lambda, we are able to control the amount of regularization performed by our model. A higher value of λ corresponds to more regularization and a sparser model. The opposite is true for lower lambda, with a λ of 0 being equivalent to the OLS model. Often, the smallest λ that produces a crossvalidated MSE within one standard error of the λ that produces the minimum MSE is selected, as it yields an error close to the minimum, produces a model at least as sparse (and interpretable), and may result in lower errors on a validation and test set due to helping to prevent overfitting.

We chose λ using 5-fold cross validation on the training data set testing values between e^{-3} and e^{4}. The model we analyzed uses the λ selected by the one standard error rule. We also note that the LASSO model is robust to collinear variables, so it is fed all variables not the VIF pruned set.

4.5 Group LASSO

Of our variables, 8 describe the distribution of school funding, and 9 describe the distribution of student race and ethnicity. In our original LASSO model, there was no regard for the grouping of these variables, and some could be selected while others were discarded. This prompts us to use of a LASSO procedure that takes these groupings into account to improve the interpretability of our model.

One such method is the group LASSO. Where LASSO applies an L_1 penalty to coefficients, resulting in a sparser model, group LASSO is a more generalized model, which causes either all predictors or no predictors in each group to be set to zero. It does so by combining non-sparse regularization (similar to Ridge Regression) within groups with LASSO outside of it, still producing a sparse model, but one where variables within a group are selected together. As such, group LASSO is often used in situations where variables should only be considered if all within a group must be considered, such as levels of proteins belonging to the same pathway or in our case, to consider the overall distribution of spending rather than just distinct categories. The equation of group LASSO is as follows:

$$\hat{\beta}_{\lambda}^{L} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k} ||\boldsymbol{\beta_j^k}||_2$$

where β is the vector of predicted coefficients equal of size equal to the number of predictors, K is the number of groups, β^k is the vector of predicted slopes within group k, and p_k is the size of group k. We can see from this equation that group LASSO is very similar to LASSO, differing only in the penalty term. Where LASSO's penalty term consists of the L_1 norm of the vector of predicted coefficients, group LASSO's consists of the L_1 norm of the vector of the adjusted L_2 norm of the group's coefficients. As such, group LASSO penalizes coefficients within a group using the L_2 norm, similar to Ridge Regression, not producing sparsity within the group, while using the L_1 norm to penalize the magnitude of group coefficients, producing an overall sparse model. As such, group LASSO presents a middle-ground between Ridge Regression and LASSO. This can be seen by noticing that if all groups are of size 1, group LASSO is the same as LASSO, and if there is one group, it is the same as Ridge Regression.

Just as with LASSO, λ determines the amount of regularization, and thus the level of sparsity in our final model. We selected λ using 5-fold cross validation, and the mean squared error was reported for the model selected using the λ with the 1se rule. However, due to the cross-validation library only automatically testing values that produced a very sparse model and not covering λ values that produced anywhere near a minimum, we needed to manually input a range for lambda, for which we used λ of the form e^k for $k \in [1,11]$ at an interval of .2. This range appears to encompass λ values that produce an error that approaches the minimum. While the λ that truly produces the minimum error appears to be slightly less than e^1, testing values that are smaller than this takes significant time to compute and does not meaningfully change the 1 standard error rule lambda.

4.6 Box-Cox

Because the Tukey-Ascombe plot (Figure **??**) shows that the linear model doesn't entirely hold at the edges of the range, we wanted to test one model that uses a nonlinear transformation function. We chose the Box-Cox transformation because Box-Cox is flexible and we do not have significant knowledge about the structure of the true prediction function.

The Box-Cox model uses restricted maximum likelihood to select a parameter λ, and the outcomes are

transformed using the function

$$g_\lambda(z) = \begin{cases} \frac{z^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(z) & \lambda = 0 \end{cases}$$

Once the outcomes are transformed, a linear model f is fit using the same design matrix and the transformed outcome variable. To predict outcome values, we use the function $g^{-1}f(X)$, where X is the design matrix.

We created a Box-Cox model using the MASS package, selecting a λ value using restricted maximum likelihood. The VIF pruned variables were used because Box-Cox is sensitive to collinearity.

5 Results and Discussion

5.1 Coefficient Analysis

As discussed in the methods section, we used a mixed effects model with county as a random effect and all available variables except % Hispanic students (which was pruned due to collinearity) fitted on the training data to get multiple regression coefficients, p values, and confidence intervals for the relationship between math and writing SAT scores and the predictor variables, and then used p value adjustment with a significance threshold of 0.05 to isolate the significant variables. A table of the confidence intervals, estimates, and Holm adjusted p values (which were essentially indistinguishable) of the coefficients of the statistically significant variables for the SAT math score regression are displayed in Figure 1, and a similar table for the coefficients of the statistically significant variables for the SAT writing score regression are displayed in Figure 2.

	Lower Bound	Estimate	Upper Bound	Holm Adjusted P-Value
Intercept	182.6526777	302.4194199	422.1861621	0.0000174
Census Day Enrollment	0.0023508	0.0057369	0.0091229	0.0175051
% English Learners	-1.2896902	-0.9491049	-0.6085196	0.0000013
% Asian	1.7551518	1.9700282	2.1849046	0.0000000
% Black/African American	-1.0183552	-0.7379792	-0.4576033	0.0000063
% White	0.5895478	0.8085272	1.0275066	0.0000000
% on Free/Reduced Price Meals	-0.8129119	-0.6089159	-0.4049200	0.0000002

Figure 1: The confidence intervals and adjusted p values for all significant coefficients for predicting math SAT scores in the mixed effect model.

We see that the percentage of English learners, Asian students, Black/African American students, White students, and students on reduced price meals are significant variables for both average school SAT math scores and average school SAT writing scores. The percentage of district expenditures spent on instruction is also a significant variable for SAT writing scores, and the number of students is also a significant variable

	Lower Bound	Estimate	Upper Bound	Holm Adjusted P-Value
Intercept	143.0965701	254.2204077	365.3442454	0.0001705
% English Learners	-1.1653758	-0.8462145	-0.5270533	0.0000054
% Asian	1.1709779	1.3719837	1.5729895	0.0000000
% Black/African American	-0.8172895	-0.5544573	-0.2916251	0.0007636
% White	0.4831966	0.6879984	0.8928002	0.0000000
% on Free/Reduced Price Meals	-0.9449910	-0.7542969	-0.5636028	0.0000000
District level Instruction Expenditure %	0.6794577	1.7421456	2.8048334	0.0254969

Figure 2: The confidence intervals and adjusted p values for all significant coefficients for predicting writing SAT scores in the mixed effect model.

for SAT math scores.

For the variables in common between the SAT math and SAT writing models, the direction of the effect on SAT score is the same between models. For example, (while keeping all other variables constant), increasing the % of English learners at a school decreases both predicted average SAT math and predicted average SAT writing scores.

Note that we have shown correlation - not causation! This is NOT an experimental study. It would be incredibly incorrect to use these results to say, for example, that White students are better at math because increasing the number of White students at a school increases average SAT math scores. Our dataset does not include many confounding variables that could be involved, such as average household income, and only examines average school scores - not individual scores. Additionally, the causal relationship could go the other way (for instance, schools with more white students could be more likely to give additional SAT tutoring). Further, SAT score is not a measure of ability, so our results are insufficient to predict academic performance or overall ability.

We see that the effect of % of White students, % of African American students, % of English learners, and % on reduced price meals is about the same on average math SAT scores and average writing SAT scores as the confidence intervals overlap. However, we can see that the 95% confidence intervals of the effect of % Asian students on average math SAT scores and average writing SAT scores do not overlap, and in fact the % of Asian students has a significantly greater impact on math SAT scores than writing SAT scores.

Additionally, we can say that (again keeping all other factors constant), schools with more students (census day enrollment) have on average higher average SAT math scores to a statistically significant degree, and schools that put a greater percentage of their budget into instruction expenditure have higher average SAT writing scores to a statistically significant degree.

5.2 Feature Selection

The features selected by each of our models are displayed in this table, with green cells representing included predictors and an x denoting variables removed due to high collinearity.

	Forward Selection		LASSO		Group LASSO	
	Math	Writing	Math	Writing	Math	Writing
Census Day Enrollment						
English Learners %						
American Indian or Alaska Native %						
Asian %						
Black or African American %						
Filipino %						
Hispanic or Latino %	x	x				
Native Hawaiian or Pac Islander %						
Two or More Races %						
None Reported %						
White %						
Free/Reduced Meals %						
Per Pupil Ratio: Teacher						
Avg Years Teaching						
Suspension Rate						
Current Exp of Educ per ADA						
Instruction Exp %						
Instruc-related Svcs Exp %						
Pupil Services Exp %						
Ancillary Services Exp %						
Community Services Exp %						
Enterprise Exp %						
General Administration Exp %						
Plant Services Exp %						
Teacher Salary-Avg						

Figure 3: Selected variables for forward selection, LASSO, and Group LASSO

5.2.1 FFS

As described above, forward feature selection uses a greedy algorithm to select a set of variables with presumed optimal performance by minimizing the C_p. Figure 4 shows the C_p over number of variables curve for the math SAT score variable selection; the number of variables selected was 13. Figure 5 shows the C_p over number of variables curve for the writing SAT score variable selection; the number of variables selected was 17.

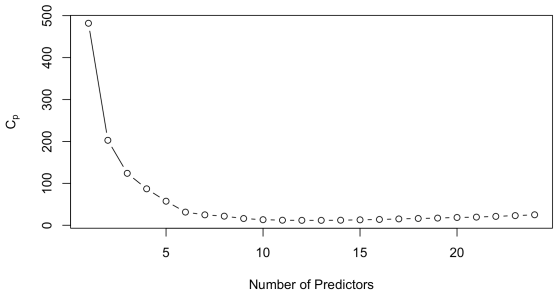


Figure 4: C_p over number of variables, Math

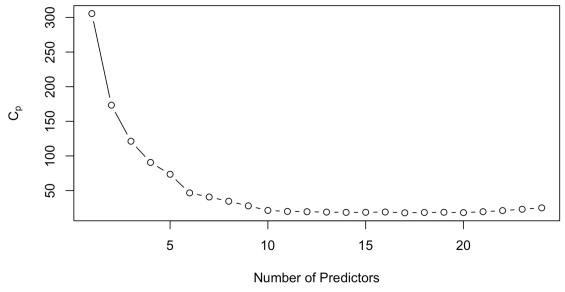


Figure 5: C_p over number of variables, Writing

Note that we can't really ascribe any statistical significance to these selected variables; they're just the ones that optimize C_p under the forward selection algorithm described in the methods section.

5.2.2 LASSO

As described in the methods section, LASSO uses cross-validation to select a λ regularization parameter, which is used to penalize large coefficients and set some to zero.

Below, we can see the value of the mean squared error across a variety of λ values, with the leftmost dotted line denoting the λ producing the minimum MSE and the rightmost dotted line denoting the largest λ producing an MSE within 1 standard error of the minimum MSE. We used the 1 standard error λ, as it results in higher regularization and a sparser model for the cost of a small increase in the mean squared error. Note that λ was selected using the MSE on the training set using 5-fold cross-validation, hence why the y-axis values differ from those presented as our MSE on the validation set in this section.

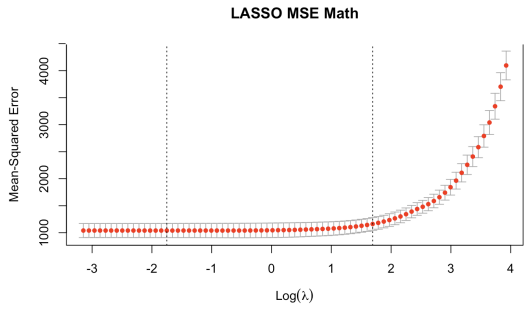


Figure 6: LASSO log(lambda) vs. MSE, Math

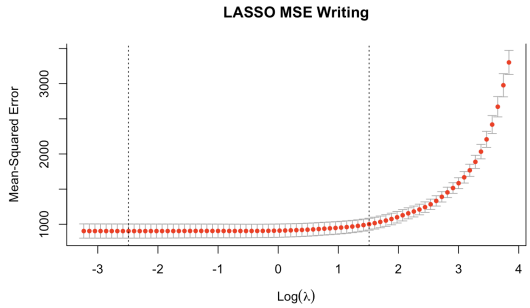


Figure 7: LASSO log(lambda) vs. MSE, Writing

5.2.3 Group LASSO

As described in the methods section, group LASSO uses cross-validation to select a λ regularization parameter, which is used to penalize groups of large coefficients and set all coefficients of some groups to zero.

Below, we can see the value of the mean squared error across a variety of λ values, with the leftmost dotted line denoting the λ producing the minimum MSE and the rightmost dotted line denoting the largest producing an MSE within 1 standard error of the minimum MSE. Note that the range of lambdas does not fully encompass the global minimum. The cross validation library took significant time to load, and based on our plot of coefficients, all variables were selected for $\lambda < e$, meaning that the global minimum is very similar to the minimum on the range we tested and the λ within one standard error would be nearly the same. Similar to our process for LASSO, we used the λ within 1 standard error, as it results in higher regularization and a sparser model for

the cost of a small increase in the mean squared error. Again, λ was selected using the MSE on the training set, hence why the y-axis values are larger than those presented as our MSE on the validation set in this section.

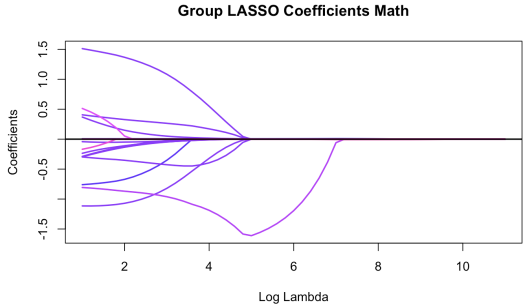


Figure 8: Group LASSO $\log(\lambda)$ vs. Predicted Coefficients, Math

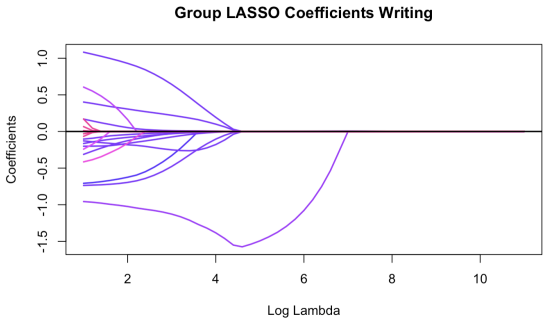


Figure 9: Group LASSO $\log(\lambda)$ vs. Predicted Coefficients, Writing

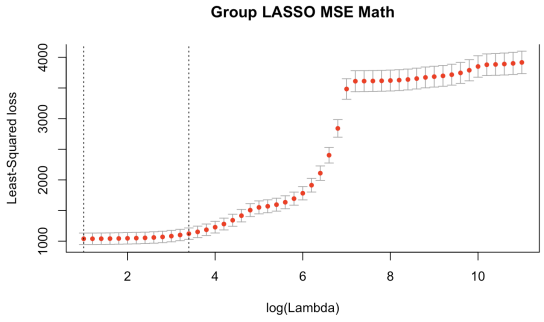


Figure 10: Group LASSO $\log(\lambda)$ vs. MSE, Math

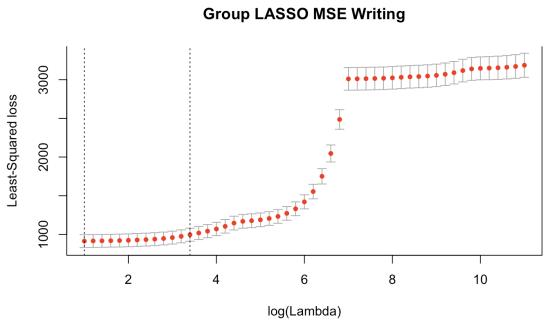


Figure 11: Group LASSO $\log(\lambda)$ vs. MSE, Writing

5.2.4 Box-Cox

Using restricted maximum likelihood, we selected $\lambda = 0.42$ for math and $\lambda = 0.10$ for writing. These produced the following residual plots:

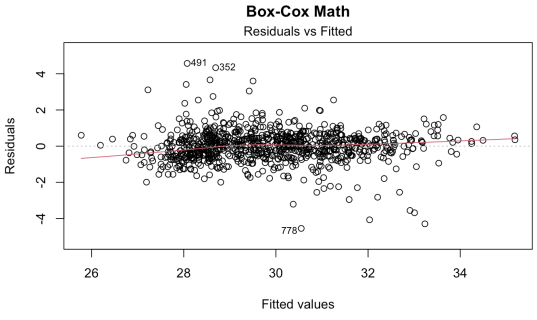


Figure 12: Box-Cox Residuals Plot, Math

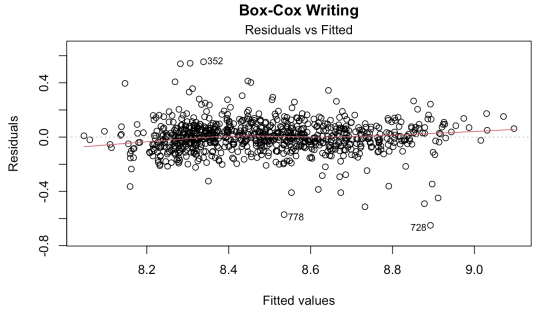


Figure 13: Box-Cox Residuals Plot, Writing

Notice that the linear model seems to hold slightly more with the outcome transformed, particularly at the edges of the range of SAT scores, than in the linear model. While the MSE of the Box-Cox model is slightly worse than the OLS, this still means there's a chance it could provide more accurate predictions for schools with predicted scores near the upper edge of the SAT score range (where the linear model holds less well).

5.3 Model Selection

The MSE of each of our models when run on the validation set is displayed below.

Model	MSE Math	MSE Writing
OLS	742.4	636.2
Forward Selection	744.6	639.5
LASSO	899.8	801.3
Group LASSO	829.3	789.89
Box-Cox	753.7	638.2
Mixed Effects	741.0	626.6

Table 1: MSE values for various models

We notice that the model producing the lowest MSE on the validation set for both the Math and Writing SAT data is the mixed effect model, which is an OLS with county data added as a random effect. The full OLS and forward-selected model perform almost as well as the mixed effects model, meaning that we could probably use OLS for ease of computation or the forward selected model for greater interpretability with only a slight increase in error.

Box-Cox performed somewhat worse than the forward selected model for math and about the same for writing, though as described it did produce a model for which the linear assumption holds slightly better. The improved performance for writing makes sense since the linear model holds better for math than for writing. This does not seem to be grounds to use the Box-Cox model, as the linear approximation is pretty good for the vast majority of the range of SAT scores we see in the training data and we do not expect to see many average SAT scores outside this range.

The LASSO and group LASSO did not perform well on the validation set, and performed similarly to each other. This is likely due to sensitivity to randomization in our data; the λ selected using cross-validation on the training data may not be the one that is actually best on our validation set. Indeed, the MSE of the LASSO and group LASSO with the λ that produces the minimum cross-validated MSE on the training set is comparable to that of the OLS and forward selection models.

This sensitivity is something that we would need to investigate moving forward. In particular, when

we changed the seed used to assign the train, validation, and test sets, we got very different values for the MSE of various models. For some seed values, the group LASSO and LASSO performed very well, and on others it did not. Similarly, for some seed values the MSE on the train set was lower than that of the validation set (which was not the case for this seed), and for some, the one standard error λ values performed better than the minimum selected λ values. In order to produce reliable results, we would need to perform cross-validation in order to mitigate this sensitivity to randomization.

6 Test of Best Model

Since the mixed effect model had the best performance on the validation set for both SAT math and SAT writing scores, we report its performance on the test set. The MSE for the predictions of average school level math SAT scores was 1217.09, and the MSE for the predictions of average school level writing SAT scores was 1091.205. The large difference between MSE in the validation and test set is evidence that the model is highly sensitive to data values, the same issue with the random seed called out above. Crossvalidation would be a good solution.

To put this in units of SAT, we take the square root; the RMSE of school level SAT math score predictions was 34.9, and the RMSE of school level SAT writing predictions was 33.0. Considering that the range of values for SAT was around 300 to 700, this is relatively good predictive accuracy.

7 AI Use Statement

In this assignment, we did not use generative AI tools, besides brief glances at the AI results that are automatically at the top of Google search.

Firstly, we do not trust AI tools enough to cite them. Large language models are predictive and trained on a variety of data, much of which could not be verifiably cited. As such, we used scientific papers, class notes, and code documentation to obtain any information treated as factual in the paper as it is traceable and verifiable.

Our goal for being in this class is to learn to perform statistical analysis, and we did not determine generative AI use to substantially contribute to that goal. While it could have reduced time spent coding, we did learn to debug R code and how to navigate common difficulties in popular data analysis packages, which we feel contributed to our learning in this class. We are also both comfortable reading documentation of existing packages, and in order to feel confident in our work, we would need to read the documentation regardless of AI use to fully understand what packages are doing. As such, it has been faster to simply use documentation and sources such as Stack Exchange from the start, ensuring that we are not enticed by code that seems to work or produce a desirable result without fully understanding what it's doing. By coding all of our results ourselves, we feel we increased our learning.

Additionally, the social and environmental impacts of AI are significant. Without strong justification for AI use, which we do not have as described above, we do not consider the environmental tradeoffs to be worthwhile, especially since these consequences largely fall on people with less access to wealth and resources than us. And due to social consequences of AI use (class stratification from job loss, lower investment in creative fields, and widespread impacts on loneliness and critical thinking), we personally do not want to contribute to the consumer market for such tools.

8 Works Cited

Chen SY, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. J Thorac Dis. 2017 Jun;9(6):1725-1729. doi: 10.21037/jtd.2017.05.34. PMID: 28740688; PMCID: PMC5506159.

Dixon-RomÁN, E. J., Everson, H. T., Mcardle, J. J. (2013). Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance. Teachers College Record: The Voice of Scholarship in Education, 115(4), 1-33. https://doi.org/10.1177/016146811311500406 (Original work published 2013)

Everson, Howard T., and Roger E. Millsap. 2004. "Beyond Individual Differences: Exploring School Effects on SAT Scores." Educational Psychologist 39 (3): 157–72. doi:10.1207/s15326985ep3903_2.

Fetler, M. E. (1991). Pitfalls of Using SAT Results to Compare Schools. American Educational Research Journal, 28(2), 481-491. https://doi.org/10.3102/00028312028002481 (Original work published 1991)

Yang, Yi, Zou, Hui. (2014). A fast unified algorithm for solving group-lasso penalize learning problems. Statistics and Computing. 25. 10.1007/s11222-014-9498-5.

Yu S, Guindani M, Grieco SF, Chen :, Holmes TC, Xu X. Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. Neuron. 2022; 110(1):21-35. doi: 10.1016/j.neuron.2021.10.030.