Analyzing and Predicting the Relationships between Sesame Street Viewership and Test Scores among School Children

Angela Wang, QiHan Zhou, Matthew Murray, Michelle Mao, Sara Lemus, Sophie Dalldorf

12/13/2021

Introduction

Sesame Street is a hallmark feature of American television that has been educating young children since 1969. The publicly-funded show was created with the goal of supplementing the learning of low-income preschoolers who did not have the access to early education that their other peers did. However, Sesame Street has had far-reaching success across all segments of the American population, regardless of background. In fact, according to one 1996 survey, 95% of all American children had watched Sesame Street by the time they turned 3. A more recent 2018 estimate suggested that about 86 million Americans had watched the show as young children [1]. Such an influential TV series that has affected the childhoods of countless Americans warrants further investigation into the educational impacts of the program itself.

This project in particular makes use of a dataset that was compiled by researchers in the early 1970s meant to assess whether or not Sesame Street was achieving its goals of educating economically disadvantaged children. The researchers tested a variety of children from different backgrounds on critical skills like numbers and letters. They then re-tested these children after they had watched Sesame Street. Through our research, we seek to answer two questions related to this dataset about childhood education:

1. Can we predict the change between pre-test and post-test scores after children watch Sesame Street and analyze the most influential predictor variables?
2. Can we use information about the pre-test score and other demographic variables to predict what type of background the child is from?

There are a variety of situational and demographic factors that were collected by researchers when examining the relationship between watching Sesame Street and a child's performance on skill tests. For our first research question, we seek to not only incorporate some of these covariates to build an accurate predictive model, but also to gain some insight as to which of these covariates are associated with a higher difference between the child's performance on the test. Through this research goal, we ultimately hope to build a useful predictive model that helps us further understand the educational impact of Sesame Street on children of all backgrounds.

The second question investigates the relationship between various subject pre-test scores and children of different income and living situations. By understanding this, we can develop a more more nuanced understanding of how low income children may be affected by their background. The absence of early education for low-income students in America has been a problem for decades, and still persists today. One statistic suggests that a mere 18% of low-income students are enrolled in high-quality pre-K. Location also plays a role: 55% of children who live in rural areas lack access to high-quality education [2]. This lack of access to pre-K disproportionately affects students from disadvantaged backgrounds, as they are left less ready for kindergarten and may suffer further setbacks compared to their peers as they continue their education into elementary, middle, and high school. Thus, it is critical to further understand the relationship between the

background of children and how this may affect their ability to test well on critical subjects like learning and math.

We address our research questions using various machine modeling techniques. More specifically, for our first research question we will perform a comparison of least squares, ridge, and tree regression models to find the one with the lowest test MSE, which after performing our analysis ends up being ridge regression. For our second question, we will fit a multi-class SVM classifier to predict what background a child comes from.

Data

The data was collected in the early 1970s by researchers from Educational Testing Service, and the actual dataset itself was retrieved from a Columbia University database [3]. The researchers sampled children representative of 5 economically advantaged and disadvantaged populations in the US. The children were randomly assigned to either receive encouragement to watch Sesame Street or not to receive encouragement. The children in the encouragement group were given promotional materials about Sesame Street, and the ETS staff called and visited every week to ensure children were actively watching Sesame Street. Those who were not in this group did not receive any of this. Children were intially tested in 1970, and the post test evaluations were one year later.

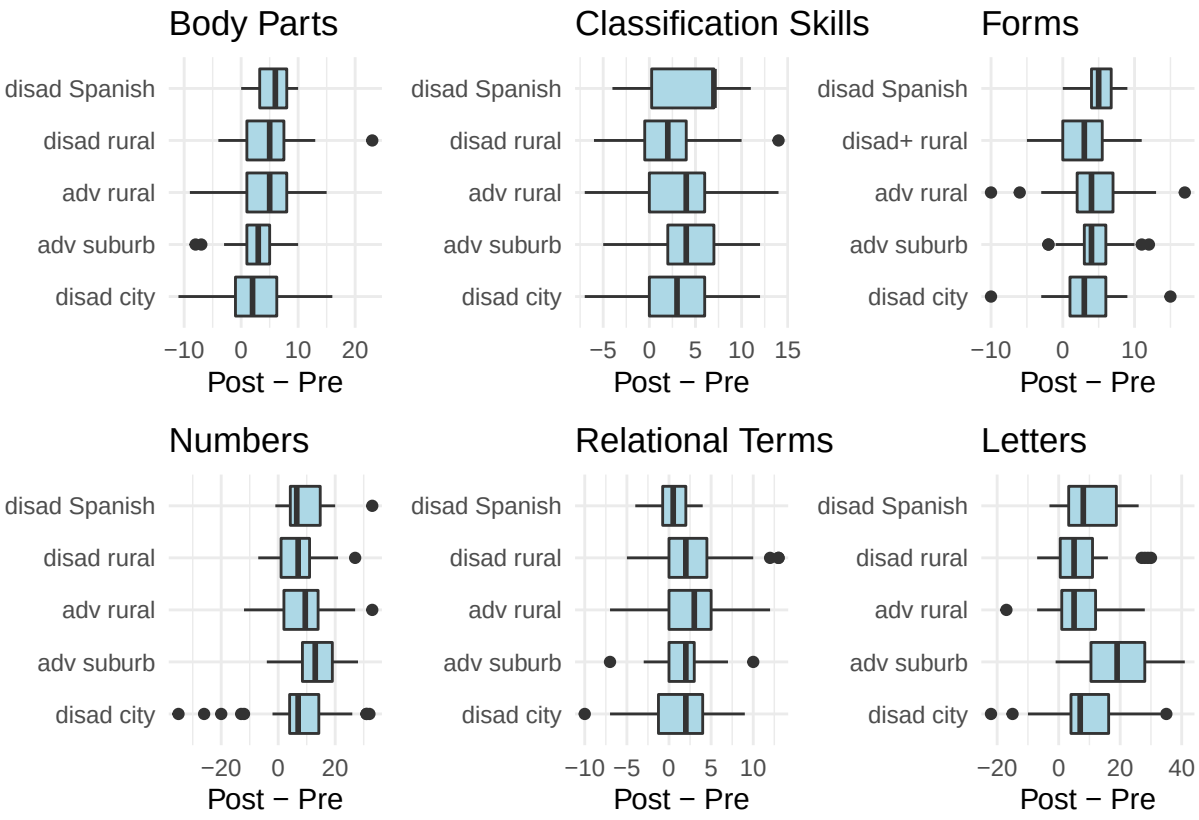
Our data contains 34 variables. The ID refers to a subject's identification number. The site refers to the age and background information of the child. A site value of 1 indicates a 3-5 year old disadvantaged child from the inner city. A site value of 2 represents a 4 year old advantaged child from the suburbs. A value of 3 represents an advantaged rural child. A site value of 4 indicates a disadvantaged rural child. Lastly, a value of 5 represents a disadvantaged Spanish speaking child. For the sex, a value of 1 indicates male, and a value of 2 indicates female. The age category is the child's age in months. The `viewcat` variable represents the frequency of viewing Sesame Street (1 = rarely, 2 = once/twice per week, 3 = 3-5 times a week, 4 = more than 5 times per week). The setting is where Sesame Street was viewed; a value of 1 indicates home and a value of 2 indicates school. The `viewenc` column refers to if the child was encouraged to watch or not (1 = child not encouraged, 2 = child encouraged). `Encour` is the same variable but with values 0 and 1, respectively. `Regular` is an indicator variable representing if a child is a regular viewer (0 = rarely watched, 1 = watched once per week or greater). We decided that it was not feasible to use data from multiple sources due to the fact that there is not any further information available about the participants in the study (other than the dataset from the Columbia University database).

The `prebody, prelet, preform, prenumb, prerelat,` and `preclasf` columns all describe pretest scores on varying types of assessments (body parts, letters, forms, numbers, relational terms, and classification skills, respectively). The columns labeled `postbody, postlet, postform, postnumb, postrelat,` and `postclasf` are the children's respective post-test scores. We created the following variables - `bodydiff, letDiff, formDiff, numbDiff, relatDiff, clasfDiff` - to represent the difference in post-test scores and pretest scores for each child. Lastly, `peabody` represents a score of "mental age" for vocabulary maturity from the Peabody Picture Vocabulary Test.

Our main focus will be on the new variables we created (`bodyDiff, letDiff, formDiff, numbDiff, relatDiff, clasfDiff`) and variables related to how often the children watch Sesame Street (`namely, viewcat and regular`). Lastly, we will look into the backgrounds of the children, including `site, sex,` and `age`.

We decided to plot the distribution of post-test - pre-test scores for each test using boxplots. First, we created separate plots for different values of the variable `site`, as we conjectured that a child's background would have an effect on his or her change in test scores. From these plots, it appears that children from some backgrounds show greater changes in test scores than others, but this change was something we knew that we could examine further in our models.

Plotting Differences in Unstandardized Scores By Students' Backgrounds



We also created separate plots for different values of the variable `viewcat` because we also assumed that the frequency at which children watch Sesame Street would affect their respective changes in test scores. One key takeaway from these plots is that accross all tests, the median value of post-test score - pre-test score is greater for children who watch Sesame Street very frequently (greater than 5 times a week) than for children who rarely watch Sesame Street. This trend suggests that Sesame Street has a positive educational impact on children by helping them improve their test scores. Again, however, this takeaway is something that we knew that we could examine further with our models.

Plotting Differences in Unstandardized Scores By Viewership Frequency

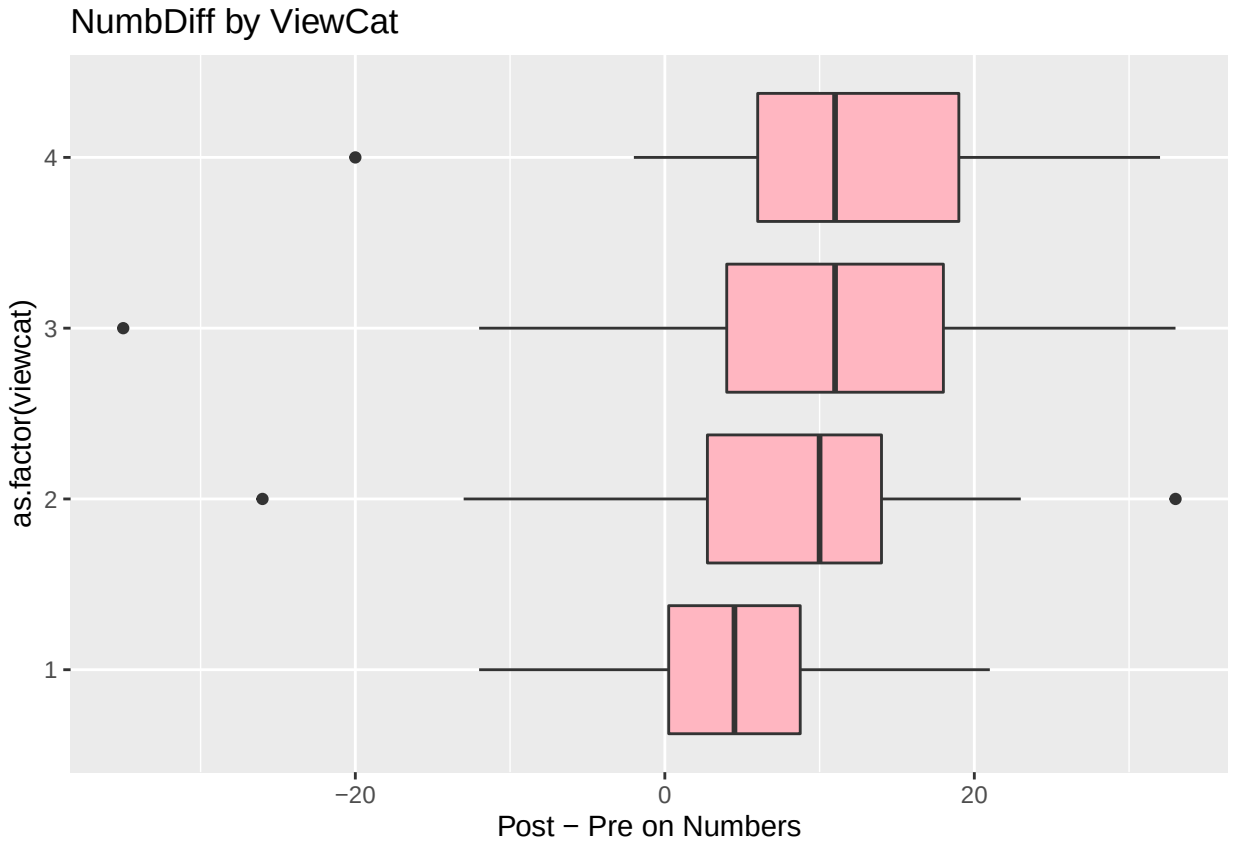
Methodology

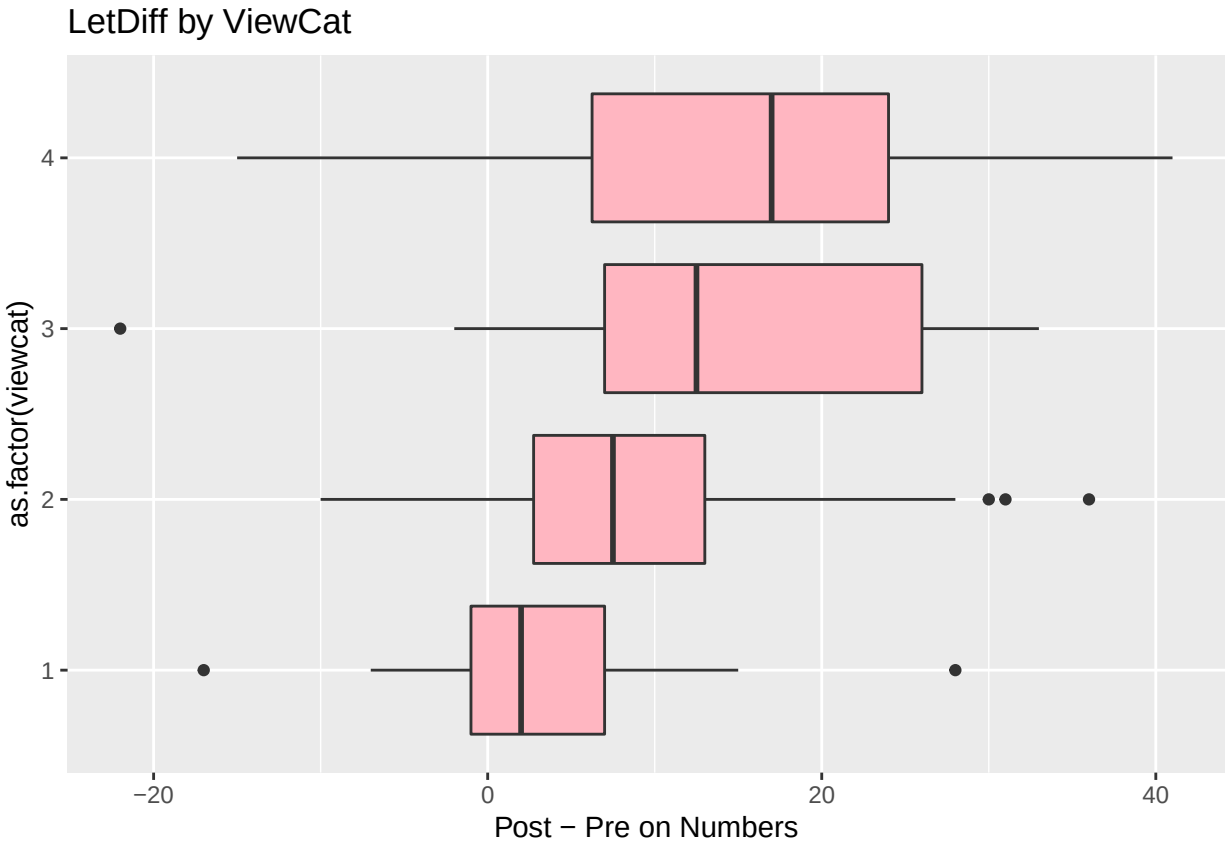
Research Question 1

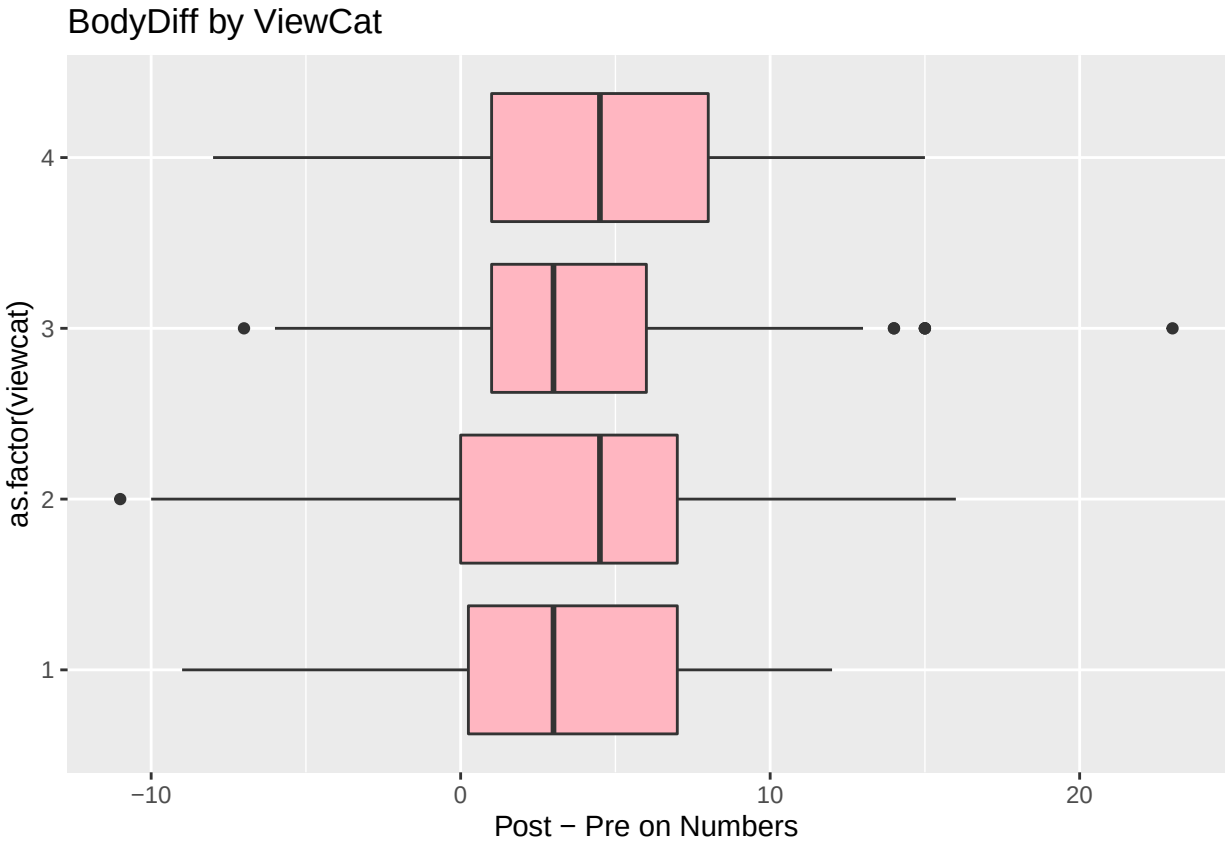
For our first research question, we seek to predict the difference in pre and post-test after children watch Sesame Street as well as identify the most influential covariates. More specifically, for all 6 tests we choose to let the difference in pre and post-test scores for that given test be the response variable, and look at the following variables as covariates: `site`, `sex`, `age`, `viewcat`, `setting`, `viewenc`. We choose to look at site (which is representative of the child's background) because, as mentioned in the introduction, Sesame Street is particularly aimed at helping underprivileged young children learn fundamental skills. We want to further study this potential association between the background of the child and the difference between their two test scores. We adjust for the age and sex of the child as well as whether or not they viewed the show at home or school, since both of these variables may potentially have an affect on their pre and post-test scores. We also include the frequency of which children watched Sesame Street (`viewcat`) as well as whether or

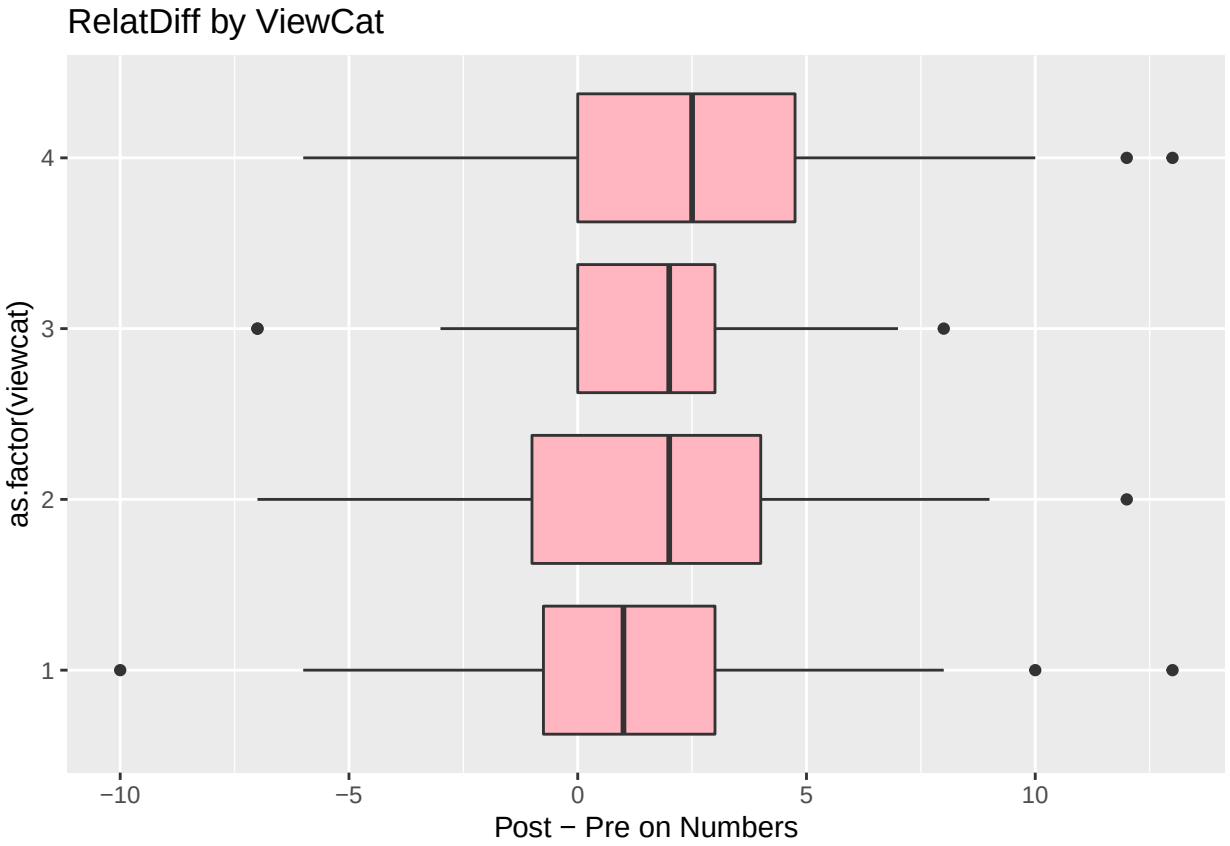
not they were encouraged to watch Sesame Street by the researchers. This is motivated by EDA plots that suggest that for all tests, children who were encouraged to watch have higher test differences than those who were not, and children who watched Sesame Street frequently had higher median test differences than children who watched rarely. Thus, we wanted to include these factors in our model since we believe they are important to include when trying to accurately model the difference between pre and post-test scores.

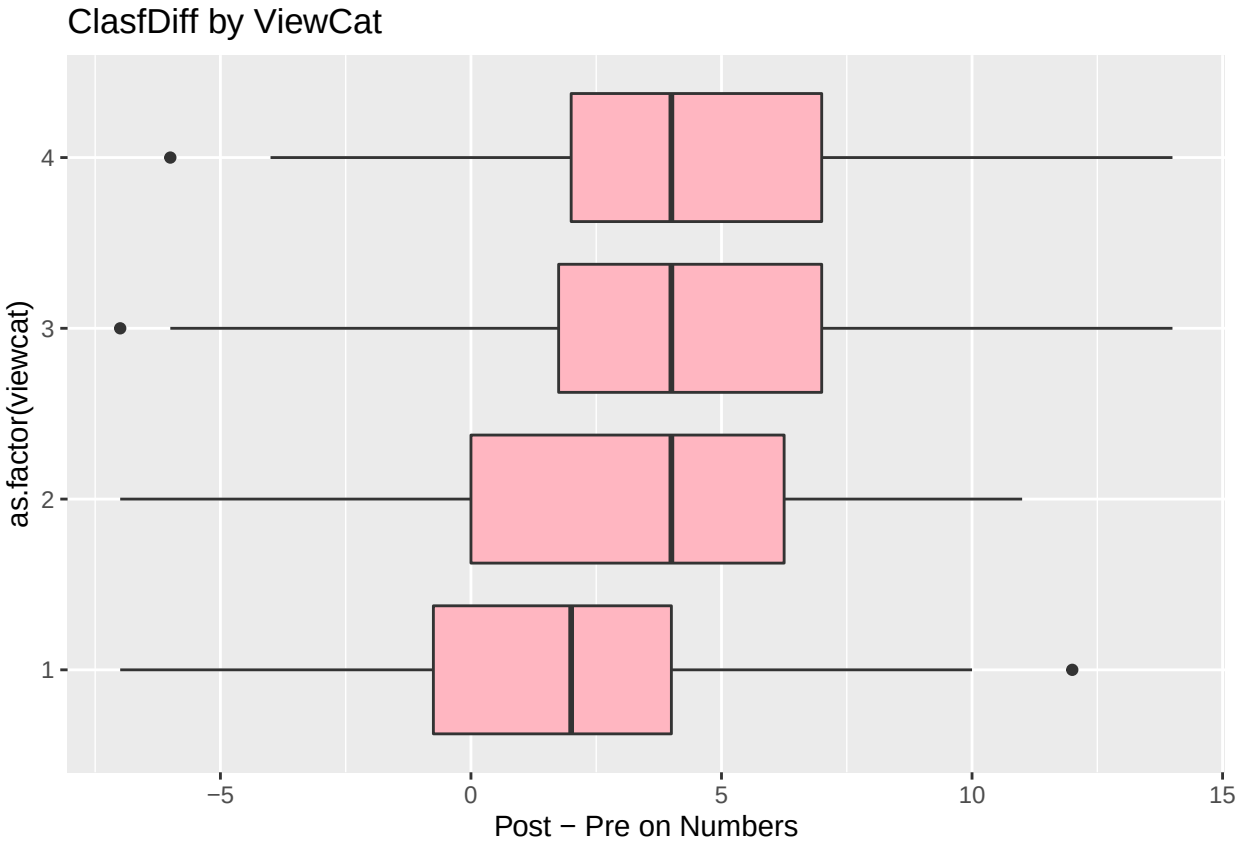
Before including these covariates in any model, we first encoded all of them as type categorical which is appropriate given that each number represents a level rather than a numerical value.

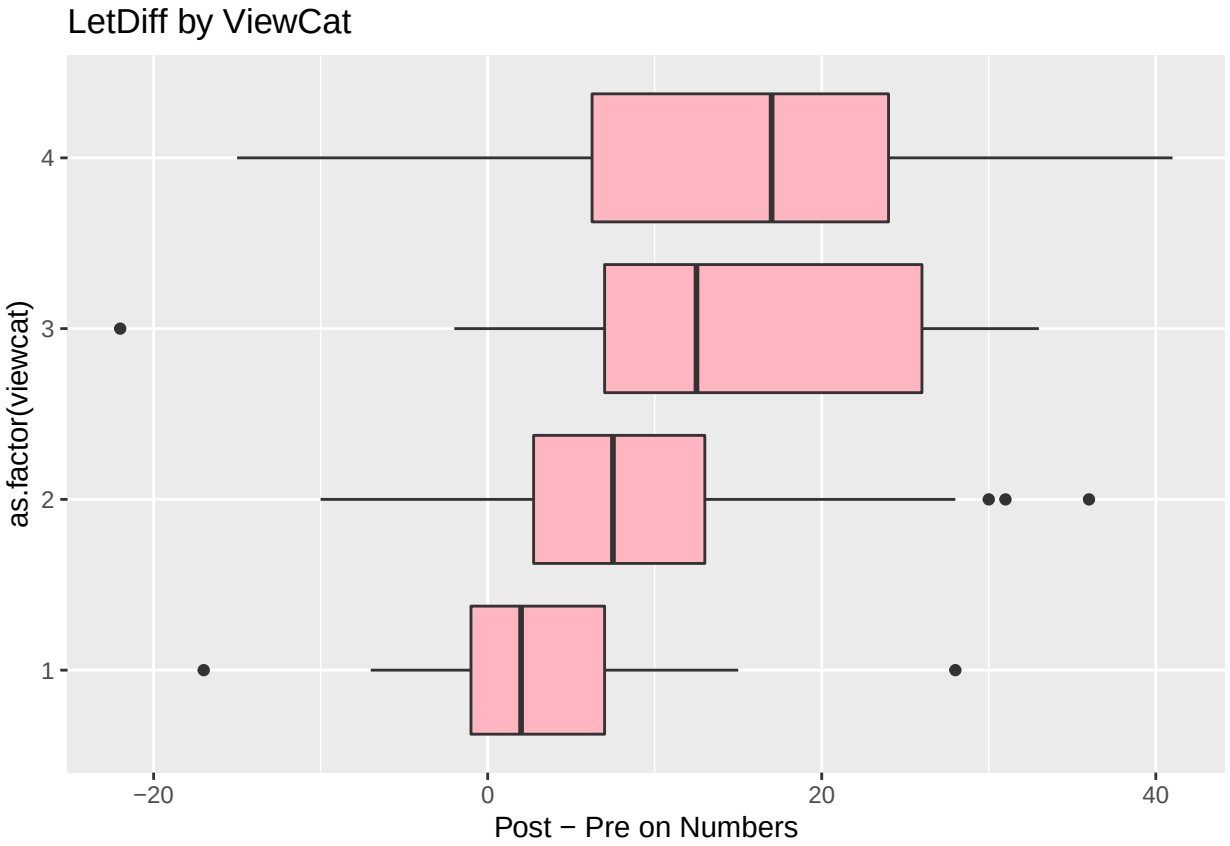



LetDiff by ViewCat
as.factor(viewcat)
4
3
2
1
−20
0
20
40
Post − Pre on Numbers


BodyDiff by ViewCat
as.factor(viewcat)
4
3
2
1
−10
0
10
20
Post − Pre on Numbers


RelatDiff by ViewCat
as.factor(viewcat)
4
3
2
1
−10
−5
0
5
10
Post − Pre on Numbers


ClasfDiff by ViewCat
as.factor(viewcat)
4
3
2
1
−5
0
5
10
15
Post − Pre on Numbers



After identifying the response variables and covariates, we must choose an appropriate model. There are many different ways to predict our continuous response variable, however we identified 3 potential models: least squares regression, ridge regression, and regression trees. All 3 of these models allow us to predict the test score differences, though have different ways of doing so and different degrees of interpretability. In order to pick a final model that best aligns with our research question, we choose to fit these 3 different models on our 6 different test score differences. We will then compare the predictive performance of each model as well as respective model properties to make an informed decision about which model is best to use for our results and conclusion. Next, we will discuss how we specifically fit all 3 model types to the data.

Least-squares Regression We first decide to consider least-squares regression models as a potential model due to the fact that they provide apt inference into the relationship between the covariates and the response variable. Our regression models follow the general form:

$$Pretest - Posttest = \beta_0 + \beta_1 \text{ (site = 2)} + \beta_2 \text{ (site = 3)} + \beta_3 \text{ (site = 4)} + \beta_5 \text{ (site = 5)} + \beta_6 \text{ (sex =female)}$$

$$+ \beta_7 (age) + \beta_8 \text{ (ViewCat = Once or Twice per Week)} + \beta_9 \text{ (ViewCat = Three to Five Times per Week)} +$$

$$\beta_{10} \text{ (ViewCat = 5 Times a Week or More on Avg)} + \beta_{11} \text{ (Setting = School)} + \beta_{12} \text{ (ViewEnc = Not Encouraged)}$$

We perform this regression for all 6 different test types.

Ridge Regression We also decide on selecting ridge regression as a candidate model due to the fact that ridge regression often provides performance improvements by shrinking slope coefficients. While shrinkage may introduce bias to a model, it decreases the variance and increases the precision of the slope coefficient estimates. The shrinkage is achieved by applying a shrinkage parameter, λ, to the Euclidean norm of a slope

coefficient. This slight increase in bias but significant decrease in variance usually decreases the MSE of a model.

We tuned our λ parameter using 10-fold cross validation. More specifically, we computed the cross-validation error rate for our model for a grid of λ values. Thereafter, we selected the λ value for which the cross-validation error is the smallest.

Initially, our least-squares and ridge regression models also took into account all possible two-way interaction terms between the variables. However, we noticed that both the Akaike Information Criterion (AIC) and adjusted-R^2 values were higher for models that did not include any interaction effects. That reason, coupled with the lack of apparent interaction effects in our exploratory data analysis and the fact that we wanted our linear models to be interpretable, is why we decided not to include any interaction effects in our final linear models.

Regression Trees Our third and final candidate model is the regression tree. Similar to linear regression, regression trees are not the most competitive in terms of predictive power or accuracy, but they are easy to interpret and can provide important inferences into the relationships between covariates and the response variable.

The process of building the trees was very simple. After initially building the tree, we then performed cross validation to determine if pruning the tree would improve the deviance (the sum of the squared errors) for the tree. Using the cross validation plot that was then created, we choose the number of nodes for the tree that minimized the deviance.

We also noticed that each of the the tests are scored on different scales. For example, the scores for the test on knowledge of body parts (noted by `bodyDiff`) range from 0-32, while those of the test on letters (noted by `letDiff`) range from 0-58. To be able to aptly compare the mean squared error (MSE) between models, we also decided to convert each response variable to the same range. More specifically, we scaled each variable to the arbitrary range [0, 30]. Lastly, we randomly split the data between testing and training, using 70% of the data for training and 30% of the data for testing.

Table 1: Test MSEs for 3 Candidate Models

Response	Least.Squares.Test.MSE	Ridge.Test.MSE	Regression.Tree.Test.MSE
Body Parts Diff	24.01	21.61	20.60
Letters Diff	14.31	14.20	15.41
Forms Diff	13.26	12.83	14.92
Numbers Diff	15.51	14.63	15.91
Relational Terms Diff	18.94	19.86	19.89
Classification Skills Diff	48.40	44.26	45.53

Selecting the Final Model The above table compares the predictive performance of all model types. We see that all 3 candidate models have fairly comparable performance on our test data set, with ridge regression seeming to have the lowest MSE for 4 out of the 6 tests. Least squares regression is the second best for predictive performance, scoring the lowest MSE for the relational terms test difference and having similar MSEs as ridge regression for most tests. Regression trees subsequently perform the worst.

Our goal for this question was not only to build a predictive model, but also to identify the variables that are influential in understanding the difference between a child's pre and post test score. Thus, we must also take into consideration the intepretability of all 3 model types, since we want to choose a model in which we can draw useful inferences from. While ridge regression gives the best predictive performance, it is less intepretable than linear regression due to its shrunken coefficients. While trees are the most interpretable model, they perform the worst compared to the other two across all tests. Thus, we choose to pick least

squares regression as our final model type, since it best allows us to make sufficient predictions and inferences regarding our response variable, difference in test scores.

After selecting least squares regression, we perform model diagnostics. All 6 models seemed to aptly satisfy the linearity condition, as the residual plots for each of the linear regression models have no discernible pattern or structure. The satisfaction of this condition ensures that there is in fact a linear relationship between the response variable and the predictors. The constant variance (homoscedasticity) condition of linear regression seems to be satisfied as well, as the vertical spread of the residuals is relatively constant across each of the plots. Lastly, the Normality assumption of linear regression seems to be aptly satisfied. For each of the models, the points fall along a straight diagonal line on the normal quantile (QQ) plot, indicating that the residuals follow a Normal Distribution.

Research Question 2

In order to address our second research question, predicting whether a child came from an disadvantaged background or not based on their pretest scores and demographic information, we utilized a support vector machine (SVM). As mentioned in the introduction, disadvantaged rural children often have worse access to educational opportunities and consequently suffer worse academic outcomes as well. Thus, we were interested in analyzing the strength of this association and seeing if we could predict the socioeconomic background of a child using their pre-test scores. We also wanted to adjust for other demographic factors such as age and sex, so we included these variables in our model as well. We expect younger children to have lower scores on their pre-test, and we also wanted to adjust for any possible effects of gender. Our model uses the sex, age, and all pre-test score variables to predict our response variable, site. Since our response variable is a categorical variable, a SVM is a valid choice to answer our research question. SVMs also have an advantage of using various kernels to place complex decision boundaries and capture relationships well. We also implemented a classification tree and a logistic model to answer this question as well (See Appendix for more details). The three models have relatively similar performance in terms of overall prediction accuracy. However, we decided to present the SVM model as our final model since we have spent a considerable amount of time researching different optimization methods to boost its performance. Our full model formula is:

$$SVM(site \sim female + male + age + bodyparts_{pretest} + letters_{pretest} +$$
$$forms_{pretest} + numbers_{pretest} + relationalterms_{pretest} + classificationskills_{pretest})$$

We split our data into a 70% training set and 30% testing set and analyzed the performance of our model on the test set. Since we are interested in high predictive power, we implemented a variety of different methods to improve our SVM models' performance. Standardizing the predictor variables in SVM and encoding categorical variables has been shown to improve performance for SVMs [4]. Thus, we used the standardized forms of the continuous variables in our model and encoded the female and male variables. We did indeed observe a small improvement in prediction accuracy across all models. However, one problem that particularly piqued our interests is that we are making no predictions for classes 4 or 5 for site. This could be due to sites 4 & 5 having a smaller number of observations than the other classes. We first tried using the Synthetic Minority Over-Sampling Technique (SMOTE) to increase observations within these classes; however, this technique had negligible effects on our accuracy. We then tried weighting the classes and used the formula below to assign weights to each class and specify "one versus one" comparison, which has been suggested to yield better prediction than "one versus all" [5].

$w_j = \frac{n}{kn_j}$, n is total number of data points, k is number of classes, n_j is the number of data in class j

We tested our model using linear, radial, sigmoid, and polynomial kernels and compared the predictive accuracy between these models. After the class weight assignment, the SVM models began to make prediction on class 4&5. However, doing so came at the cost of overall accuracy. Thus, more of the other observations

are being misclassified, but the few observations of sites 4&5 are correctly classified. To remedy this issue, we began to experiment with the class weights and increase the sites 4&5 weights by roughly 0.5 until we reached the highest predictive power. The model with the highest accuracy was the linear kernel SVM with weights 0.8 on site 1, 0.87 on site 2, 0.75 on site 3, 1.5 on site 4, and 3 on site 5 and a cross validation selected cost parameter of 0.1. Since this model has the highest predictive accuracy of all others that we had tried, we settled on this model as our final model. There are no explicit tests for SVM model diagnostics; however, there seems to be no major issues with our model fit.

Results

Research Question 1

As mentioned in methodology, we choose to model the relationship between the difference in test scores and our selected covariates using least squares regression. The below models are abbreviated and only include coefficients that were statistically significant at the $\alpha = 0.05$ level. Full model output can be found in the Appendix.

Body Parts Pre Score - Post Score $= 16.43 + 2.45\ (\text{site} = 3) + 2.97\ (\text{site} = 4) + 3.30\ (\text{site} = 5)$

Letters Pre Score - Post Score $= 10.52 + 3.09\ (\text{site} = 2) - 2.86\ (\text{site} = 3) + 4.85\ (\text{ViewCat} = 3 \text{ to } 5 \text{ Times Per Week}) +$
$4.77\ (\text{ViewCat} = 5 \text{ Times a Week or More on Avg})$

Forms Pre Score - Post Score $= 16.04 + 2.48\ (\text{ViewCat} = 5 \text{ Times a Week or More on Avg})$

Numbers Pre Score - Post Score $= 15.39 + 2.56\ (\text{site} = 2) + 3.20\ (\text{ViewCat} = 3 \text{ to } 5 \text{ Times Per Week}) +$
$3.64\ (\text{ViewCat} = 5 \text{ Times a Week or More on Avg})$

Relational Terms Pre Score - Post Score $= 18.49 + 2.42\ (\text{site} = 4) + 2.96 +$
$(\text{ViewCat} = 5 \text{ Times a Week or More on Avg})$

Classification Skills Pre Score - Post Score $= 10.95 + 4.06\ (\text{ViewCat} = 5 \text{ Times a Week or More on Avg})$

Interpretations

Looking at the least-squares models, we can note a few important features regarding predicting whether a child's test scores will increase. For `bodydiff`, or the difference between post and pre-tests scores for the recognition of body parts, we see that the intercept is 16.43. This large positive intercept may be explained by the fact that all 6 tests have more positive values for difference than negative values (meaning that more children had higher post test scores compared to their pre test scores). We also see that for all the levels of site that were statistically significant, our model estimated a positive coefficient value. Specifically, we see that if a child is `site3`, or an advantaged rural child, the difference between their scores goes up by 2.45 . For a disadvantaged rural child, the difference increases even more by 2.97 and for a disadvantaged Spanish speaking child, the difference goes up by 3.30. This indicates that for children from more disadvantaged backgrounds, the difference between the scores increases as compared to the two other sites.

For differences in letter recognition scores, we note the intercept is 10.52, the baseline difference. Again, holding all else constant, we see for 4 year old advantaged child from the suburbs (`site2`) this difference increases by 3.09. However, interestingly enough for an advantaged rural child (`site3`), the difference in test scores decreases by 2.86. It also seems that the frequency of watching Sesame Street has an effect on letter scores with `viewcat3` or watching 3-5 times a week increasing the difference by 4.85 and watching more than 5 times a week increasing the difference by 4.77.

For recognizing forms, the intercept is 16.04 and it seems that watching Sesame Street more than 5 times a week (`viewcat4`) , further increases this difference by 2.48. For numbers, the intercept is 15.39. We note that for an 4 year old advantaged child from the suburbs, the difference in number scores increases by 2.56. Watching Sesame Street 3-5 times a week (`viewcat3`) increases the difference by 3.20, and watching more than 5 times a week increases the difference by 3.64. For difference in relational term scores, the intercept is 18.49, the largest of all intercepts and the difference is further increased for a disadvantaged Spanish speaking child by 2.42. For those who watch more than 5 times a week, the difference increases by 2.96. For classification score differences, the intercept is 10.95. The only variable that increases this difference is the frequency of viewing, with those that watch Sesame Street more than 5 times a week increasing the difference by 4.06.

Research Question 2

As evidenced in the tables below, the linear kernel SVM has a higher accuracy than the other kernels we also tried out. The class weighted linear kernel SVM has a prediction accuracy of 0.403 on our test data with a confidence interval of 0.289 to 0.525. From the confusion matrix, we can also see that our model predicts well for classes 2 and 3 and struggles slightly with classes 1,4, and 5. Specifically, our model has tends to predict class 3 often when the true class is 4, and it tends to predict classes 3 and 5 as commonly as class 1 when the true class is 1.

Table 2: Classification Model Accuracy Table

Model	Accuracy
Linear Kernel SVM	0.4028
Radial Kernel SVM	0.3611
Sigmoid Kernel SVM	0.2639
Polynomial Kernel SVM	0.2917

Table 3: Class Weighted Linear Kernel SVM Prediction Results

Results	Values
Accuracy	0.403
95% CI	(0.289, 0.525)

Table 4: Class Weighted Linear Kernel SVM Confusion Matrix

Truth	Class1.Prediction	Class2.Prediction	Class3.Prediction	Class4.Prediction	Class5.Prediction
Class 1	6	1	6	2	5
Class 2	1	7	1	1	3
Class 3	0	1	11	1	3
Class 4	3	3	9	3	0
Class 5	0	1	1	1	2

Conclusion

Research Question 1

The findings from our least squares regression models is that the frequency at which a child watches Sesame Street (`viewcat`) and a child's background (`site`) are the most important variables in predicting the difference in test scores for a child. In 5 of the 6 least squares regression models, at least one level of `viewcat` is a significant predictor for the response variable. This speaks to two important insights. First, the importance of `viewcat` suggests that children who watch Sesame Street more frequently are more likely to see greater differences in test scores compared to children who rarely watch Sesame Street. More specifically, the greatest increase in differences are seen in children who watch Sesame Street at least 3-5 times per week or more than 5 times a week. This relationship suggests that Sesame Street does have an impact on children's learning of fundamental subject matter and that the show at least to some extent accomplishes its goals of affecting learning outcomes. The second important insight from our findings is that, at least to some extent, the background of a child is related to their difference between pre and post-test scores. At least one level of site was significant in 4 out of the 6 models that we found (body parts, letters, numbers, and relational). This implies that the background of a child is important when thinking about how children learn and perform on these tests.

We also cared about how well our model could accurately predict the difference in pre and post-test scores. Being able to predict these scores accurately is an indicator that we appropriately selected our covariates and were able to fit a model that alllows us to understand the relationship between these variables and the differences in test scores. When comparing performance across all 6 tests, we find that our models has the best predictive performance on the letters and forms tests, as these two had the lowest test MSEs. Our model had the worst performance on the Classification Skills test score difference, with an MSE value of 48. While we do feel like we satisfied our research goal of being able to build a predictive model, there is undoubtedly room for improvement when it comes to the predictive performance of our models, especially for the Classification Skills and Body Parts test.

One of the main limitations to our conclusions from our least squares models has to do with the relationship between our data and time. While our model does find statistically significant associations between covariates like a child's background and how often the child watches Sesame Street, we suspect that time may be a variable that is affecting the difference in between post-test and pre-test scores. Children can learn and develop a lot in a year, especially at such a young age (all children were between 3 and 5 years old). Thus, it is possible that time is a latent variable explaining some of this difference in testing scores. This hypothesis is supported by the fact that all of our least squares models had R-squared values below 0..4, suggesting that our models do not explain a lot of the variation in our response and there may be additional data that is affecting our response that was not included in our specification.

Research Question 2

Our goal for this research question was to build a model that could accurately predict the background that a child came from based on their age, sex, and pretest scores. Even though our model is not extremely accurate, we are still able to predict with 40% accuracy which background a child comes from using their pre-test scores and age and sex. This indicates that there is a fairly strong connection between a family's income level and living area and the performance of their children on critical subjects like letters and numbers. While our model does not perform very poorly on our data, there is still room for improvement. Of all the models we tried, the best accuracies were all right around 40%. This could be a flaw of the data as we have only 240 observations which is not very many. Additionally, there is some imbalance in our data set as we have less observations of children from sites 4 and 5, disadvantaged rural children and disadvantaged Spanish speaking children respectively. This may be a flaw of the data collection, and our data may not be representative of the general population. Additionally, there may have existed a clustering structure that caused the SVM model to struggle with separating the different levels of site. This phenomenon may be illustrated with the confusion matrix having some difficulty separating between certain classes. When the true site is 3,

indicating an advantaged child from a rural area, the model tends to predict site 4 as well which represents an disadvantaged rural child. The differences between children from these backgrounds may be quite small, so the model does not distinguish them quite well. This may be due to the fact that children living in rural areas have less access to education in general as mentioned in the Introduction. Therefore, socioeconomic differences between children living in these areas may not matter as much since high quality education is difficult to access for both high and low income families. We may be able to improve our model and gain greater predictive accuracy with more children in our data set and more data. Including more variables into our data set, such as child attendance in preschool or daycare, may also be useful for prediction and answering this research question.

Another possible inference from these results is that there may be more factors that explain pretest scores. For example, parent income level or highest degree of parent education may be a more useful response variable than a arbitrary determination of advantaged or disadvantaged. We do not have information on what the researchers determined to be advantaged or disadvantaged, so their definitions of these categories may be different than our opinions or those of the general public. Our data was also collected in the 70s which is a vastly different time period than the one we are currently in, so categories of advantaged and disadvantaged may be different now compared to back then. A more concrete response variable like the ones previously mentioned might provide more information. Future work may be focused on collecting information on these socioeconomic variables and performing analysis once again to see if our results differ. Additionally, there are also other optimization methods that we haven't explored. For example, one can try other feature transformations to make the data more separable or try to design a customized SVM kernel that is better tailored to the distribution of data for this dataset.

General Takeaways

For both research questions, we found that the site variable, which represents the socioeconomic background of a child, plays an important role on pre-test scores as well as the difference between pre- and post-test scores. This finding is consistent with our original beliefs as there is a well-known connection between worse educational opportunities and educational performance for low income and rural children. Our research supports this conclusion, indicating that more work is necessary to ensure that children of disadvantaged, low-income, and rural communities are able to access high-quality and affordable education from a young age to ensure proper learning of essential topics such as letters and numbers and ensure future academic success.

Appendix

References

[1] Sesame Street. (2021). Wikipedia. https://en.wikipedia.org/wiki/Sesame_Street

[2] The Need. (2021). First Five Years Fund. https://www.ffyf.org/why-it-matters/the-need/

[3] Index of /~gelman/arm/examples/sesame. (n.d.). Retrieved December 12, 2021, from http://www.stat.columbia.edu/~gelman/arm/examples/sesame/

[3] A Practical Guide to Support Vector Machines. (2003). National Taiwan University. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[4] One-vs-Rest and One-vs-One for Multi-Class Classification. (2020). Machine Learning Mastery. https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/

Research Question 1: Linear Models

Below are the R squared values for each test.

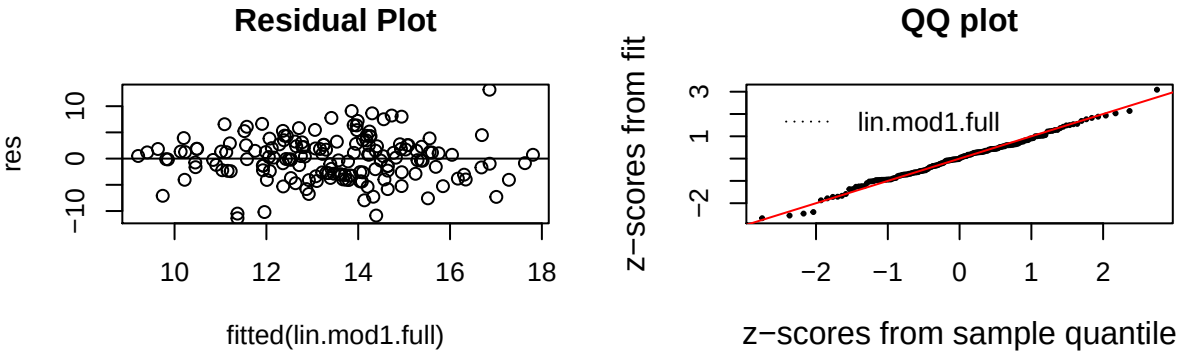
Table 5: Multiple R-Squared Values for Least-Squares Regression Models

Response	R.Squared
Body Parts Diff	0.159
Letters Diff	0.370
Forms Diff	0.063
Numbers Diff	0.139
Relational Terms Diff	0.107
Classification Skills Diff	0.094

Least Squares Model for Difference in Body Part Test Scores and Diagnostics

```
##
## Call:
## lm(formula = bodyDiff ~ (site + sex + age + viewcat + setting +
##     viewenc), data = training, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3707  -2.7927   0.0525   2.5545  13.1411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.42668    3.06383   5.361 2.92e-07 ***
## site2        0.58666    0.99737   0.588  0.55724
## site3        2.44827    0.96820   2.529  0.01244 *
## site4        2.96605    1.10488   2.685  0.00805 **
## site5        3.29854    1.45590   2.266  0.02485 *
## sex2        -1.00041    0.66732  -1.499  0.13586
## age         -0.08726    0.05558  -1.570  0.11845
## viewcat2    -0.60564    1.12908  -0.536  0.59244
```

```
## viewcat3     1.30568    1.09680   1.190  0.23568
## viewcat4     2.18215    1.12555   1.939  0.05434 .
## setting2    -1.02953    0.77713  -1.325  0.18718
## viewenc2    -0.37832    0.84111  -0.450  0.65349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.253 on 156 degrees of freedom
## Multiple R-squared:  0.1585, Adjusted R-squared:  0.09917
## F-statistic: 2.671 on 11 and 156 DF,  p-value: 0.00362
```



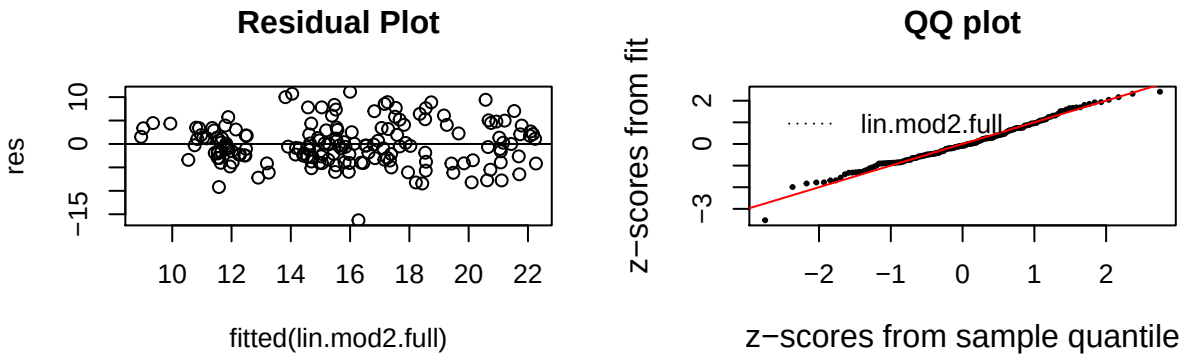
Ridge Regression Model for Difference in Body Part Test Scores

	Estimate
(Intercept)	16.012
site2	-0.270
site3	0.937
site4	0.952
site5	1.344
sex2	-0.544
age	-0.050
viewcat2	-0.774
viewcat3	0.348
viewcat4	0.976
setting2	-0.877
viewenc2	-0.426

Least Squares Model for Difference in Letters Test Scores and Diagnostics

```
##
## Call:
## lm(formula = letDiff ~ (site + sex + age + viewcat + setting +
##     viewenc), data = training, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2823  -2.9776  -0.3663   2.9275  11.1426
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 10.52353    3.32595   3.164  0.00187 **
## site2        3.08586    1.08270   2.850  0.00496 **
## site3       -2.85941    1.05103  -2.721  0.00726 **
## site4       -1.07861    1.19940  -0.899  0.36989
## site5       -0.38501    1.58046  -0.244  0.80786
## sex2         0.37911    0.72441   0.523  0.60148
## age          0.05050    0.06034   0.837  0.40390
## viewcat2     1.59616    1.22568   1.302  0.19474
## viewcat3     4.85135    1.19064   4.075 7.32e-05 ***
## viewcat4     4.77123    1.22184   3.905  0.00014 ***
## setting2     0.70346    0.84362   0.834  0.40563
## viewenc2    -1.56705    0.91307  -1.716  0.08810 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.617 on 156 degrees of freedom
## Multiple R-squared:   0.37,  Adjusted R-squared:  0.3256
## F-statistic:  8.33 on 11 and 156 DF,  p-value: 2.043e-11
```



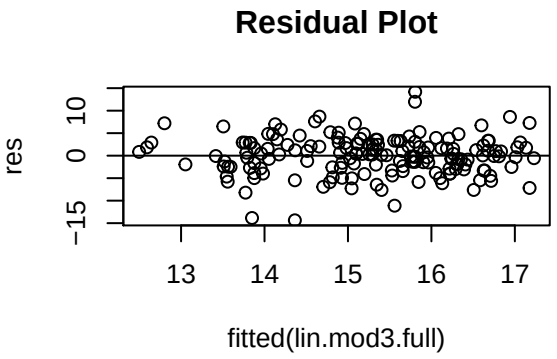
Ridge Regression Model for Difference in Letters Test Scores

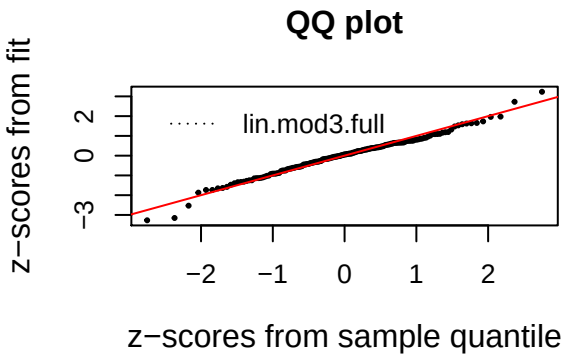
	Estimate
(Intercept)	12.087
site2	3.090
site3	-2.557
site4	-1.344
site5	-0.478
sex2	0.314
age	0.041
viewcat2	0.543
viewcat3	3.532
viewcat4	3.469
setting2	0.534
viewenc2	-1.693

Least Squares Model for Difference in Forms Test Scores and Diagnostics

```
##
## Call:
## lm(formula = formDiff ~ (site + sex + age + viewcat + setting +
```

```
##     viewenc), data = training, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3637  -2.5344   0.2658   2.7054  14.1942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.04193    3.16388   5.070 1.11e-06 ***
## site2        0.97029    1.02994   0.942   0.3476
## site3        0.73495    0.99982   0.735   0.4634
## site4        0.92962    1.14096   0.815   0.4164
## site5        1.00499    1.50344   0.668   0.5048
## sex2         0.09439    0.68911   0.137   0.8912
## age         -0.05234    0.05740  -0.912   0.3632
## viewcat2     0.99101    1.16595   0.850   0.3966
## viewcat3     1.65614    1.13262   1.462   0.1457
## viewcat4     2.47768    1.16230   2.132   0.0346 *
## setting2    -0.08994    0.80251  -0.112   0.9109
## viewenc2    -0.42935    0.86858  -0.494   0.6218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.392 on 156 degrees of freedom
## Multiple R-squared:  0.06302,    Adjusted R-squared:  -0.003048
## F-statistic: 0.9539 on 11 and 156 DF,  p-value: 0.4909
```



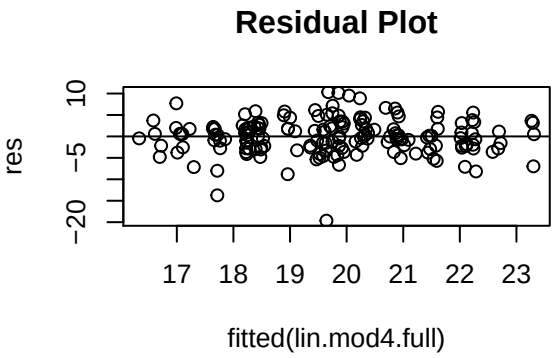


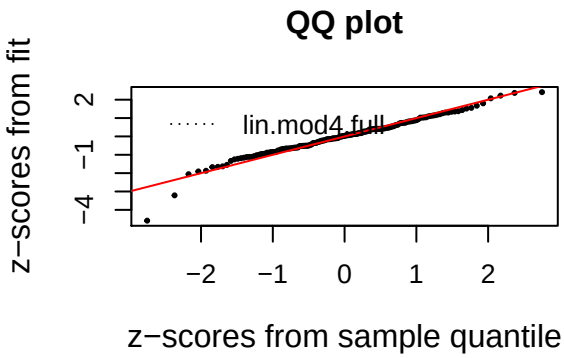
Ridge Regression Model for Difference in Forms Test Scores

	Estimate
(Intercept)	15.524
site2	0.084
site3	0.027
site4	-0.064
site5	0.026
sex2	0.020
age	-0.005
viewcat2	-0.027
viewcat3	0.039
viewcat4	0.144
setting2	-0.058
viewenc2	-0.100

Least Squares Model for Difference in Numbers Test Scores and Diagnostics

```
##
## Call:
## lm(formula = numbDiff ~ (site + sex + age + viewcat + setting +
##     viewenc), data = training, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6433  -2.6010   0.1375   2.3541  10.3247
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.388591   3.084422   4.989  1.6e-06 ***
## site2        2.561427   1.004073   2.551  0.01170 *
## site3        0.768616   0.974707   0.789  0.43157
## site4        1.151651   1.112304   1.035  0.30210
## site5        1.946294   1.465685   1.328  0.18615
## sex2         0.013992   0.671801   0.021  0.98341
## age          0.009505   0.055956   0.170  0.86534
## viewcat2     1.841715   1.136668   1.620  0.10719
## viewcat3     3.199614   1.104174   2.898  0.00430 **
## viewcat4     3.636674   1.133111   3.209  0.00161 **
## setting2     0.596707   0.782354   0.763  0.44679
## viewenc2     0.589407   0.846765   0.696  0.48742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.281 on 156 degrees of freedom
## Multiple R-squared:  0.139,  Adjusted R-squared:  0.07825
## F-statistic: 2.289 on 11 and 156 DF,  p-value: 0.01273
```





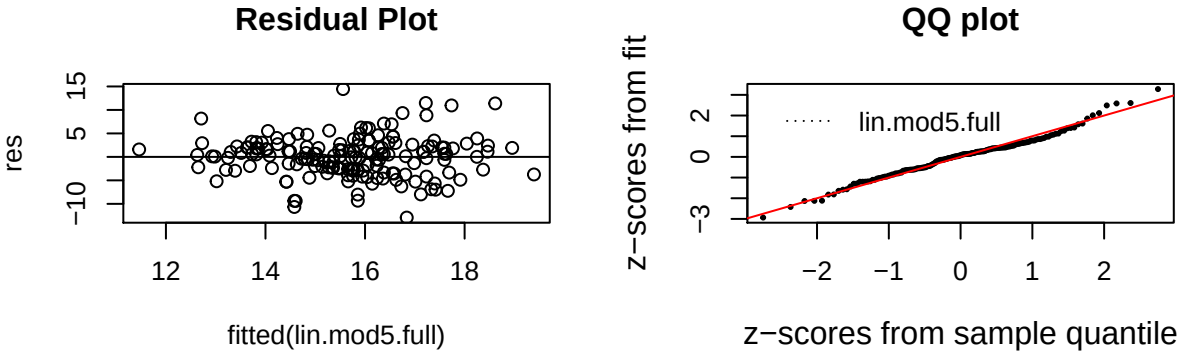
Ridge Regression Model for Difference in Numbers Test Scores

	Estimate
(Intercept)	18.982
site2	0.993
site3	-0.100
site4	-0.285
site5	0.199
sex2	0.034
age	0.005
viewcat2	-0.114
viewcat3	0.493
viewcat4	0.800
setting2	0.145
viewenc2	-0.114

Least Squares Model for Difference in Relational Terms Test Scores and Diagnostics

```
##
## Call:
## lm(formula = relatDiff ~ (site + sex + age + viewcat + setting +
##     viewenc), data = training, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9246  -2.7114   0.3439   2.2185  14.4415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.48969    3.17203   5.829  3.1e-08 ***
## site2       -0.05900    1.03259  -0.057   0.9545
## site3        1.66242    1.00239   1.658   0.0992 .
## site4        2.41955    1.14390   2.115   0.0360 *
## site5       -0.40561    1.50731  -0.269   0.7882
## sex2         0.31932    0.69088   0.462   0.6446
## age         -0.09729    0.05755  -1.691   0.0929 .
## viewcat2     1.05190    1.16895   0.900   0.3696
## viewcat3     0.96616    1.13554   0.851   0.3962
## viewcat4     2.95954    1.16529   2.540   0.0121 *
## setting2    -0.60666    0.80457  -0.754   0.4520
```

```
## viewenc2     0.19462    0.87082   0.223   0.8234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.403 on 156 degrees of freedom
## Multiple R-squared:  0.1072, Adjusted R-squared:  0.04421
## F-statistic: 1.702 on 11 and 156 DF,  p-value: 0.07739
```



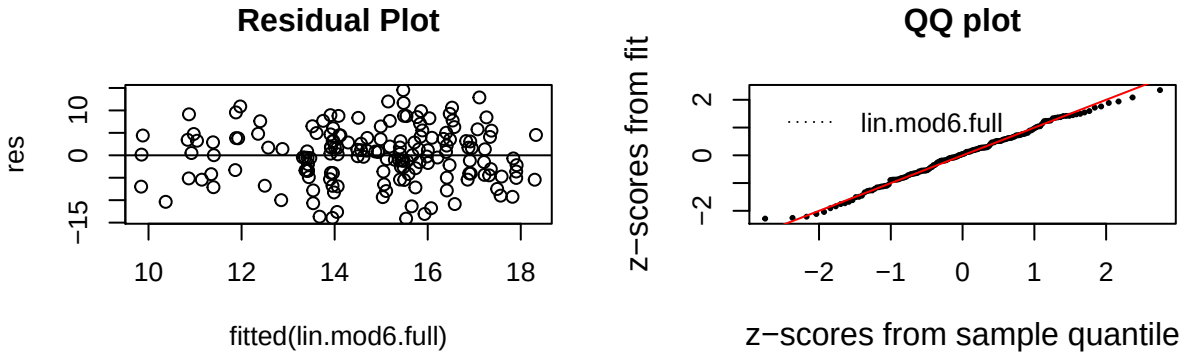
Ridge Regression Model for Difference in Relational Terms Test Scores

	Estimate
(Intercept)	17.933
site2	-0.264
site3	0.587
site4	0.667
site5	-0.563
sex2	0.220
age	-0.050
viewcat2	-0.074
viewcat3	-0.124
viewcat4	0.951
setting2	-0.431
viewenc2	-0.115

Least Squares Model for Difference in Classification Skills Test Scores and Diagnostics

```
##
## Call:
## lm(formula = clasfDiff ~ (site + sex + age + viewcat + setting +
##     viewenc), data = training, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1077  -4.0442   0.2665   3.8807  14.5260
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.951998   4.455915   2.458   0.0151 *
## site2        1.799749   1.450536   1.241   0.2166
## site3        0.420798   1.408112   0.299   0.7655
```

```
## site4        2.040728   1.606891   1.270   0.2060
## site5        1.497017   2.117404   0.707   0.4806
## sex2         1.500233   0.970519   1.546   0.1242
## age         -0.008915   0.080838  -0.110   0.9123
## viewcat2     3.041609   1.642090   1.852   0.0659 .
## viewcat3     3.072479   1.595147   1.926   0.0559 .
## viewcat4     4.060404   1.636951   2.480   0.0142 *
## setting2     0.477618   1.130229   0.423   0.6732
## viewenc2    -1.088979   1.223281  -0.890   0.3747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.185 on 156 degrees of freedom
## Multiple R-squared:  0.09426,    Adjusted R-squared:  0.03039
## F-statistic: 1.476 on 11 and 156 DF,  p-value: 0.1455
```



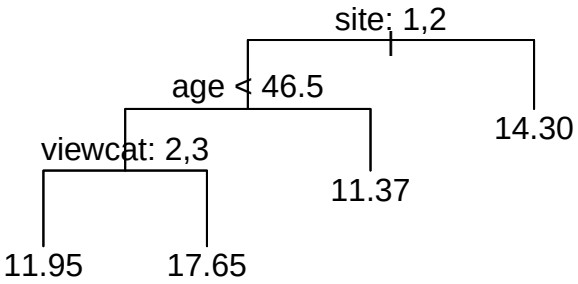
Ridge Regression Model for Difference in Classification Skills Test Scores

	Estimate
(Intercept)	14.769
site2	0.151
site3	-0.022
site4	-0.041
site5	0.021
sex2	0.116
age	-0.001
viewcat2	0.041
viewcat3	0.035
viewcat4	0.151
setting2	-0.015
viewenc2	-0.151

Research Question 1: Regression Tree Models

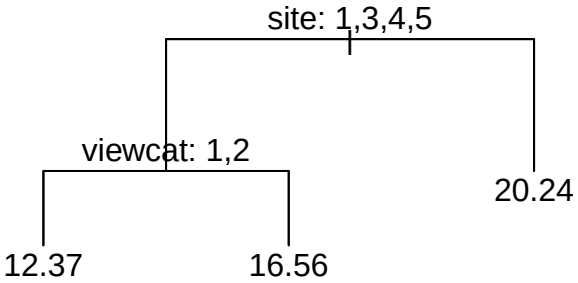
Model for Difference in Body Parts Test Scores

Variables used in tree construction:`site`, `age`, `viewcat`, `sex`, `viewenc`. Terminal nodes: 13. Residual mean deviance: 14.16.



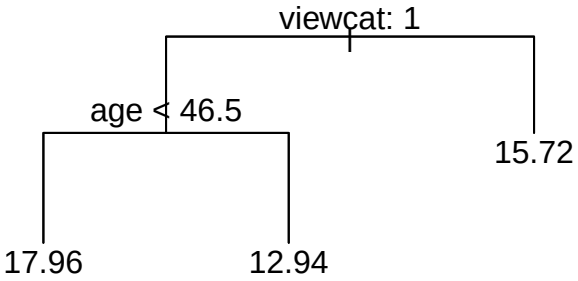
Model for Difference in Letters Test Scores

Variables used in tree construction:`site`, `age`, `viewcat`, `setting`, `viewenc`. Terminal nodes: 12. Residual mean deviance: 18.63.



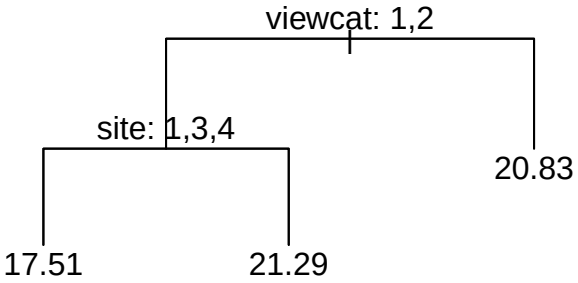
Model for Difference in Forms Test Scores

Variables used in tree construction:`site`, `age`, `viewcat`, `setting`. Terminal nodes: 10. Residual mean deviance: 15.88.



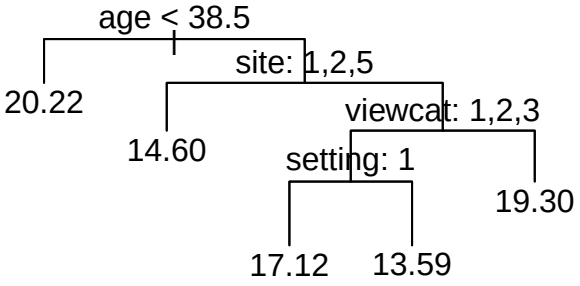
Model for Difference in Numbers Test Scores

Variables used in tree construction:`site`, `age`,`sex`, `viewcat`, `setting`, `viewenc`. Terminal nodes: 18. Residual mean deviance: 13.64.



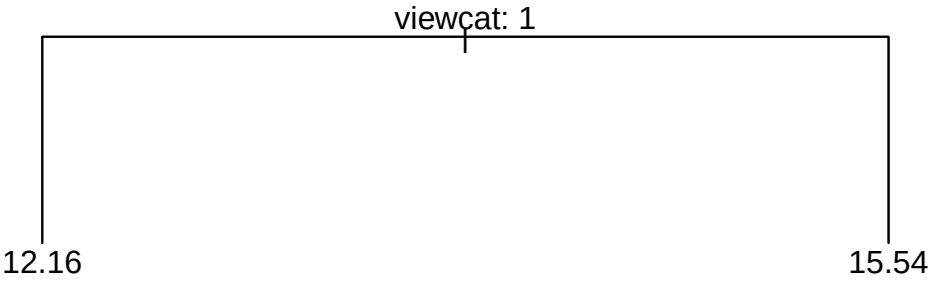
Model for Difference in Relational Terms Test Scores

Variables used in tree construction:`site`, `age`, `viewcat`, `setting`, `sex`. Terminal nodes: 10. Residual mean deviance: 15.3.



Model for Difference in Classification Skills Test Scores

Variables used in tree construction:`site`, `age`, `viewcat`, `sex`. Terminal nodes: 14. Residual mean deviance: 29.66.



Research Question 2

Accuracy Table of Random Forest & Logistic Regression

Table 6: Classification Model Accuracy Table

Model	Accuracy
Random Forest	0.4444
Logistic Regression	0.4307

Random Forest

```
## Confusion Matrix and Statistics
##
##     pred
## true  1  2  3  4  5
##    1  6  5  7  2  0
##    2  4  7  2  0  0
##    3  3  0 10  3  0
##    4  3  1  8  6  0
##    5  1  0  2  2  0
##
```

```
## Overall Statistics
##
##                Accuracy : 0.4028
##                  95% CI : (0.2888, 0.525)
##     No Information Rate : 0.4028
##     P-Value [Acc > NIR] : 0.5447
##
##                   Kappa : 0.2215
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
## Sensitivity           0.35294  0.53846   0.3448  0.46154       NA
## Specificity           0.74545  0.89831   0.8605  0.79661  0.93056
## Pos Pred Value        0.30000  0.53846   0.6250  0.33333       NA
## Neg Pred Value        0.78846  0.89831   0.6607  0.87037       NA
## Prevalence            0.23611  0.18056   0.4028  0.18056  0.00000
## Detection Rate        0.08333  0.09722   0.1389  0.08333  0.00000
## Detection Prevalence  0.27778  0.18056   0.2222  0.25000  0.06944
## Balanced Accuracy     0.54920  0.71838   0.6026  0.62907       NA
```

Logistic Regression

```
## # weights:  60 (44 variable)
## initial  value 270.385569
## iter  10 value 224.661284
## iter  20 value 205.001057
## iter  30 value 193.994082
## iter  40 value 186.948903
## iter  50 value 186.213534
## iter  60 value 186.148958
## iter  70 value 186.146812
## iter  70 value 186.146811
## iter  70 value 186.146811
## final  value 186.146811
## converged
```

```
## Call:
## multinom(formula = factor(site) ~ ., data = train.log)
##
## Coefficients:
##   (Intercept)         age    viewcat     setting    viewenc     prebody
## 2 -2.33155723 -0.02851059  0.1671689  -0.1723663 -0.4205104  0.18169812
## 3  0.05755516  0.20107446  0.0726002  -1.4396594 -1.5937567 -0.10132305
## 4  0.52070488  0.14041265 -1.1560807  -0.2744569 -1.3458286 -0.08902983
## 5 10.77382380  0.20246633 -0.1290428 -16.3009311 -1.5011369 -0.09094212
##         prelet     preform       prenumb    prerelat    preclasf
## 2 -0.021574850 -0.12985947 -0.0781747329  0.16740257  0.10875790
## 3 -0.009618286 -0.32080214 -0.0307718697  0.05942666 -0.08062801
## 4 -0.023880439  0.05801188 -0.0008763081 -0.11001706 -0.01601207
## 5  0.149877507 -0.31580541 -0.1020246788  0.34854595 -0.12639855
```

```
##
## Std. Errors:
##   (Intercept)        age   viewcat   setting   viewenc    prebody     prelet
## 2    2.539808 0.05877122 0.2735599 0.6038785 0.6013929 0.07166225 0.03755288
## 3    2.552263 0.06296057 0.2800905 0.6392345 0.6272709 0.07157933 0.05105309
## 4    2.743404 0.06311748 0.3228949 0.6489907 0.6712331 0.07501958 0.04404500
## 5    2.206224 0.08353394 0.4579698 2.2063318 0.9368912 0.09992158 0.07033801
##     preform    prenumb  prerelat   preclasf
## 2 0.1144522 0.05108224 0.1361291 0.08630272
## 3 0.1256428 0.05602075 0.1440701 0.09338212
## 4 0.1173569 0.05356710 0.1389164 0.09183664
## 5 0.1771026 0.08446858 0.2087433 0.14707062
##
## Residual Deviance: 372.2936
## AIC: 460.2936
```

```
## Confusion Matrix and Statistics
##
##     pred
## true  1  2  3  4  5
##    1  3  4  8  3  0
##    2  8  7  2  1  0
##    3  2  3 12  2  0
##    4  0  1  4  6  1
##    5  0  0  2  1  2
##
## Overall Statistics
##
##                Accuracy : 0.4167
##                  95% CI : (0.3015, 0.5389)
##     No Information Rate : 0.3889
##     P-Value [Acc > NIR] : 0.3557
##
##                   Kappa : 0.2396
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
## Sensitivity           0.23077  0.46667   0.4286  0.46154  0.66667
## Specificity           0.74576  0.80702   0.8409  0.89831  0.95652
## Pos Pred Value        0.16667  0.38889   0.6316  0.50000  0.40000
## Neg Pred Value        0.81481  0.85185   0.6981  0.88333  0.98507
## Prevalence            0.18056  0.20833   0.3889  0.18056  0.04167
## Detection Rate        0.04167  0.09722   0.1667  0.08333  0.02778
## Detection Prevalence  0.25000  0.25000   0.2639  0.16667  0.06944
## Balanced Accuracy     0.48827  0.63684   0.6347  0.67992  0.81159
```