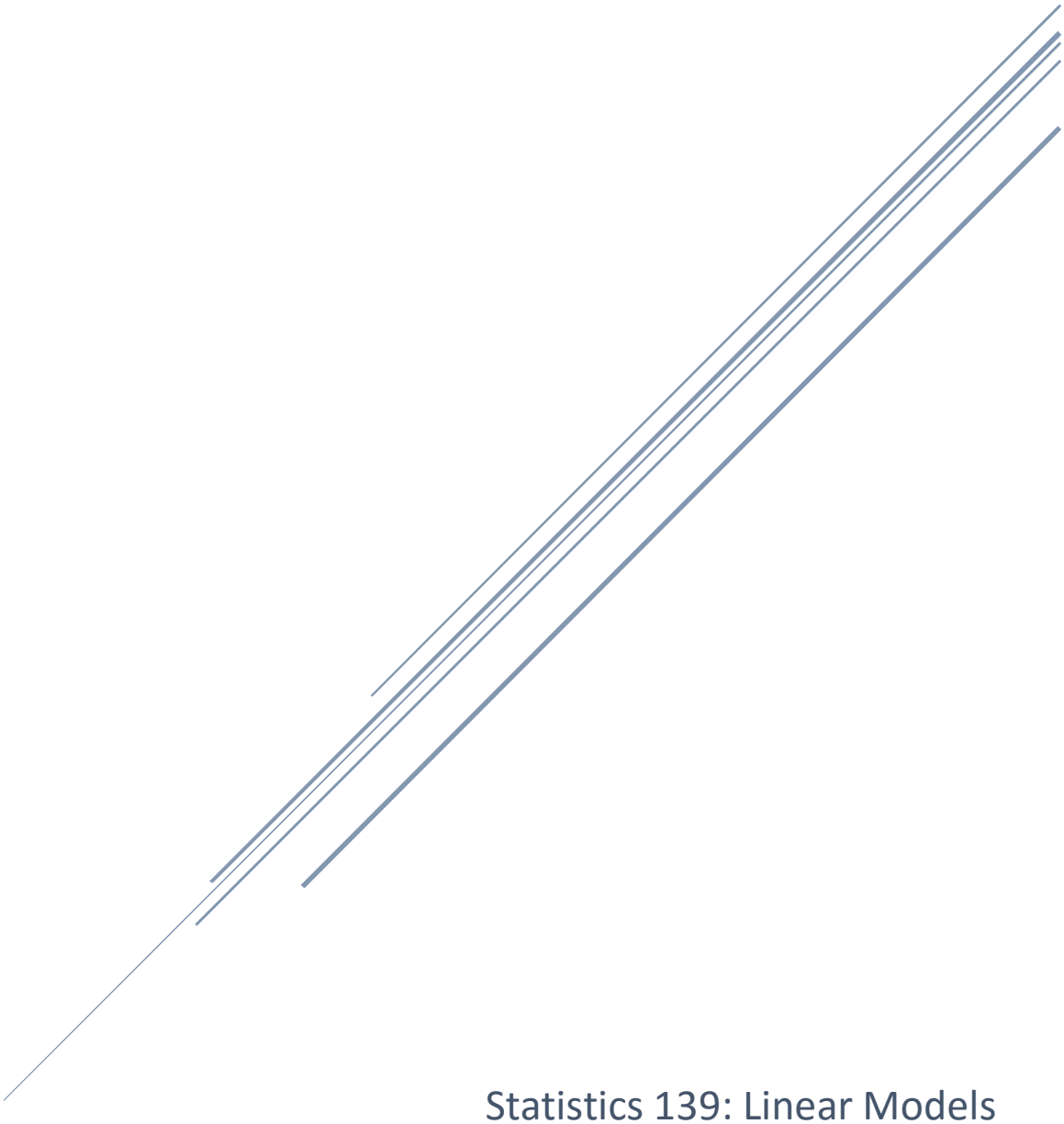
LITERACY RATES IN INDIA

An Analysis of Overall Literacy Rates Across India's 36 States and Union Territories



Statistics 139: Linear Models
Professor Kevin Rader | Fall 2017

Literacy Rates in India

By: Mirai Shah, Paul Stainier, Ramtin Talebi, and Rajet Vatsa

I. Introduction and Motivation

"Education is the most powerful weapon which you can use to change the world." - Nelson Mandela

Education is a cornerstone of empowerment; it gives people agency and enables those who are disenfranchised to break out of cycles of poverty and low expectations. Unfortunately, in many developing countries around the world, illiteracy remains a prevalent issue. For India in particular, approximately one in four people are considered illiterate. We hope to use our final project to gain a better understanding on the status of education in India.

India's education system is complex, with a host of problems the country has yet to address. For instance, schools in India are either government-backed or private. While government schools must meet certain requirements for their curricula, study materials, syllabi, examinations, etc., non-aided private schools have full autonomy over their admissions and curriculum standards. As a result, school performance is generally worse at private schools than in government-aided schools. Furthermore, while more than 95 percent of children attend primary school, just 40 percent of adolescents attend secondary school (grades 9-12). The country's mean number of years of schooling is 5.12 years, which is significantly below the average for all developing countries (7.09 years). On top of this, illiteracy rates are heightened by India's stratified social system. Scheduled Castes and Scheduled Tribes - historically disadvantaged groups which comprise about 25 percent of the Indian population - have significantly worse dropout rates than India's already high average dropout rate.[1]

We aim to incorporate such factors (i.e. percentage of scheduled castes and scheduled tribes, percentage of private schools, etc.) in our analysis of literacy rates across India. Based on a district-by-district analysis of the hundreds of potential predictors in the *Education in India* database (https://www.kaggle.com/rajanand/education-in-india), we will identify sets of attributes that are most strongly associated with and predictive of overall literacy rates.

Overall, our project's hypotheses are three-fold:

1. The state a district is in is an important factor in predicting that district's overall literacy rate
2. Attributes such as Percentage of Scheduled Tribes, Percentage of Private Schools, and Percentage of Schools with Computers are statistically significant predictors of overall literacy rate
3. Predictors that are useful in predicting overall literacy rates are also useful in predicting male and female literacy rates.

II. Data and Methods

Data Transformations

[1] Ghosh, A. (2014) **Indian School Education System: An Overview.** *The British Council*

In preparing our dataset, our first step was to determine the predictors we thought would be relevant for predicting literacy rates in each district. The original dataset had 625 observations and 819 variables, many of which were merely subcategories of overall district variables. We thus only considered the aggregate variables, such as total private schools rather than the number of private primary schools; we did this, in part, to ensure that the number of predictors would be substantially less than the number of observations. We settled on 25 variables, the complete list of which can be seen on page 1 of the Appendix. Much of the data was recorded in raw numbers (i.e. total number of private schools per district), and so to make the observations comparable across districts, we created proportion variables. For example, we created variables measuring the percentage of a district's population between 0 and 6 years-old, the number of schools per capita in each district, and the percentage of private schools in each district. The complete list of variables converted to percentages can be found on pages 2 and 3 of the Appendix.

Once our predictors were ready for analysis, we examined the distributions of our three response variables: overall literacy rate, male literacy rate, and female literacy rate. As can be seen in the histograms and QQ-plots in **Figure 1**, overall and female literacy rates were normally distributed across districts, but male literacy rate was slightly left-skewed. Thus, we transformed male literacy rate by cubing it, resulting in a distribution that was much more normal.

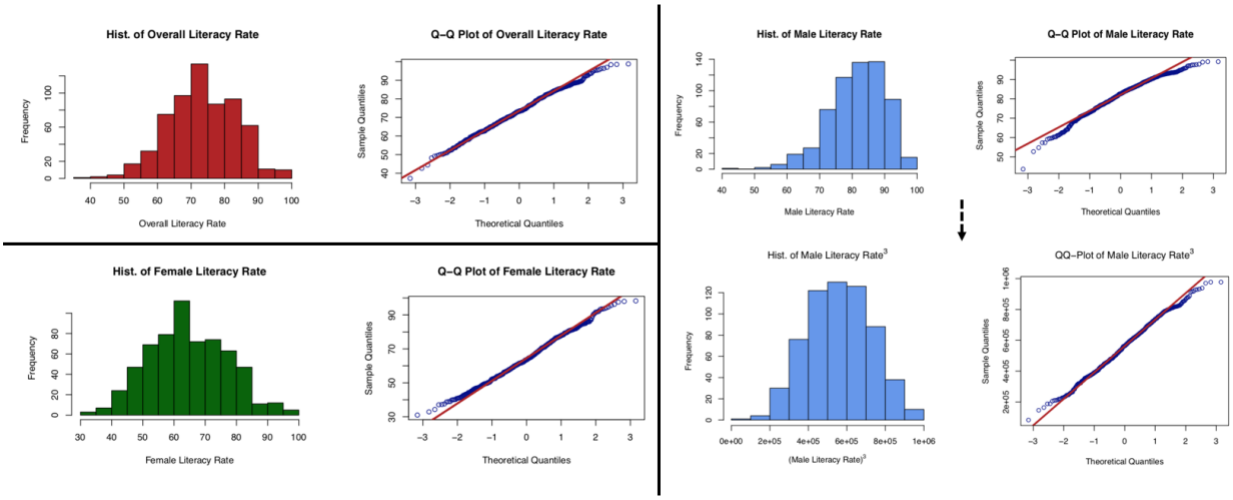


Figure 1: Transformed Response Variables

Note that in the majority of our ensuing analyses, our response variable of interest was overall literacy rate. In other words, each of the regression models we developed were built to determine attributes that were predictive of overall literacy rates across India. In the final step of our project, we then applied the predictors from our best model for the overall literacy rate response to the male and female literacy rate responses to determine if the predictors that were deemed to be most "useful" in predicting overall literacy rate were also "useful" in predicting male and female literacy rates.

Next, we examined the normality of our predictor variables. For this step, we plotted histograms of each of the predictor variables at the district level. Based on the histograms, we decided to transform 14 variables, most of which were right-skewed. We log-transformed the majority of the right-skewed predictors and squared one of the left-skewed predictors (Percentage of Government Schools per District). **Figure 2** contains histograms of representative predictors both before and after their

transformations. The histograms for the complete list of untransformed predictors are on pages 8 and 11 of the Appendix, while the histograms for the complete list of transformed predictors are on page 15 of the Appendix.

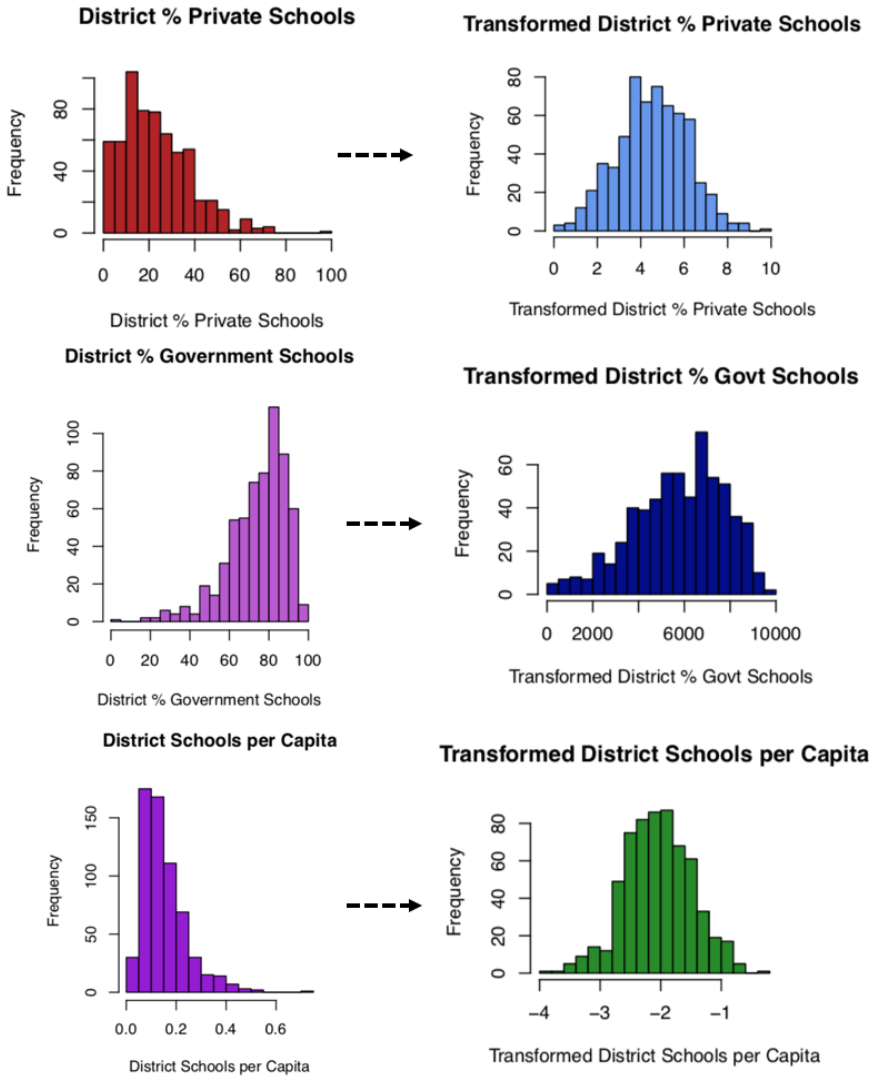


Figure 2: Sample Predictor Variables Requiring Transformations

In plotting our histograms, another observation we made was that the variables measuring the percentage of all-boys schools and all-girls schools were unrealistically high (very close to 100% in many cases). Given that the sum of these percentages was often greater than 100% (the total number of schools per district), we decided to exclude these variables from our final list of predictors.

Data Exploration

In order to examine multicollinearity, we created a heat map of the correlation matrix of all our predictors. As can be seen in **Figure 3**, there are several hot spots of higher correlation between

predictors, most of which are entirely predictable. For example, there is a cluster of high correlation between the Percentages of Population between ages 0 and 6, 6 and 10, and 11 and 13. In addition, there is a strong negative correlation between Percentage of Private Schools and Percentage of Government Schools, and a strong correlation between the Percentage of Schools with Electricity and the Percentage of Schools with Computers. However, one of the surprising sets of correlations is a negative one between the Percentage of Schools with Electricity and the proportion of younger ages in a population; that is, when there are higher proportions of people between ages 0 and 13, the percentage of schools with electricity is lower. This observed multicollinearity may make the interpretation of the coefficients associated with these variables more difficult. However, our main focus is the predictive power of our model rather than the coefficients associated with individual predictors. Because multicollinearity does not affect the reliability of our model, we decided to keep all these variables in our set of predictors. We did not want to arbitrarily remove one collinear variable and keep the others without knowing which may have a higher degree of predictive power.

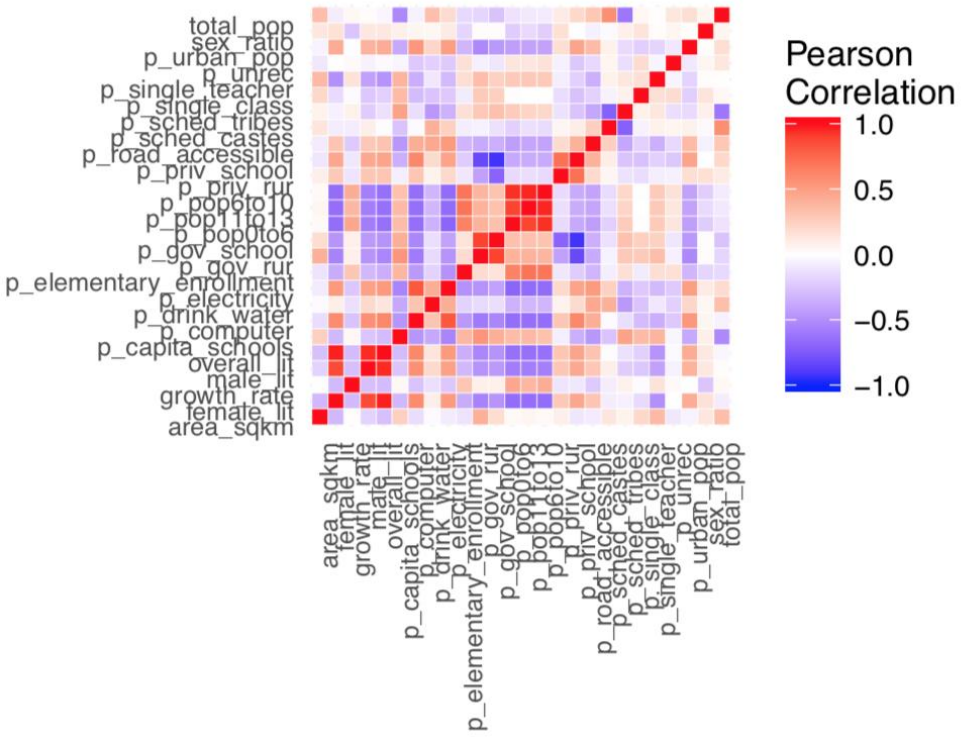


Figure 3: Correlation Matrix of Predictors

Finally, we examined the linearity of the relationships between the predictor variables and our primary response variable, overall literacy rate. We saw several strong linear relationships between our predictor variables and overall literacy rate, as exemplified by **Figure 4**. Based on these scatterplots, we also identified predictors that might have quadratic relationships with the overall literacy rate response variable. These included District Area, Percentage of Rural Government Schools, and Percentage of Rural Private Schools. An example of one such predictor with its quadratic transformation is shown in **Figure 5**. The scatterplots of all the predictors vs. overall literacy rate can be seen on pages 21 and 24 of the Appendix.

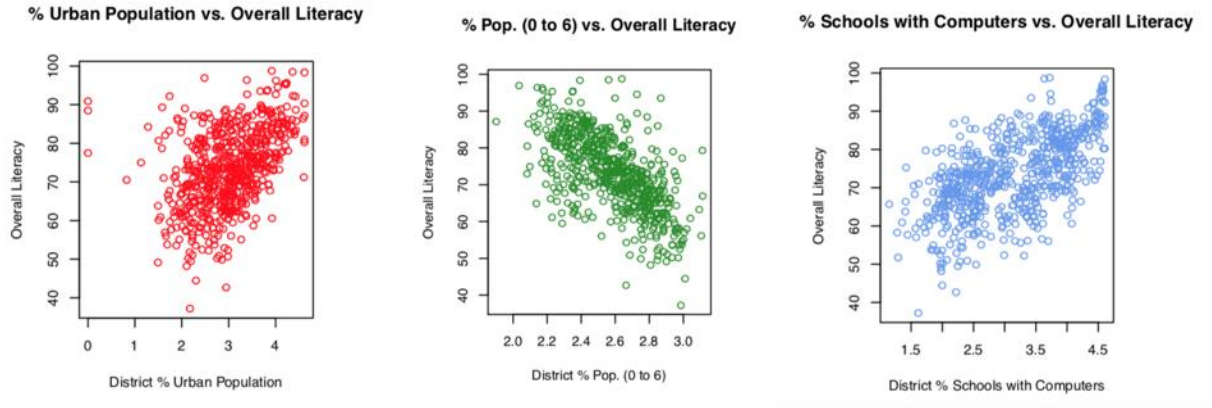


Figure 4: Sample Scatterplots

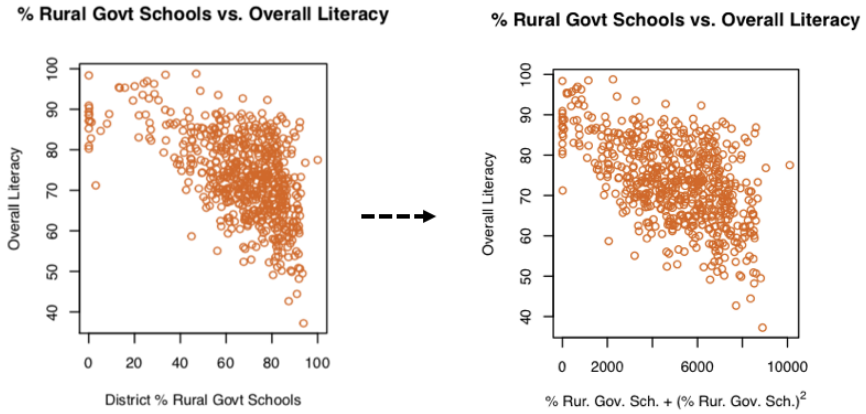


Figure 5: Potential Quadratic Relationship

Once we had a strong understanding of our variables and their individual relationships with the overall literacy rate response, we were ready to explore and compare six different linear regression models.

Regression Models and Preliminary Analysis

Model 1: We began by constructing a model that regresses the main effects of all 23 continuous predictors on the overall literacy rate response variable; the only predictor in our final dataset excluded from this model is the state code factor variable. There are 10 predictors which have a significant association with overall literacy rate: Percentage of Urban Population, District Area, Percentage of Population (Age 0 to 6), Percentage of Population (Age 11 to 13), Schools per Capita, Percentage of Rural Government Schools, Percentage of Private Rural Schools, Percentage of Single-Class Schools, Percentage of Single-Teacher Schools, and Percentage of Schools with Drinking Water. The coefficients associated with each of these predictors corroborate our intuition. For instance, if the percentage of single-class schools increases by one unit (in log-transformed space), our mean overall literacy rate decreases by 1.7, while if the log-transformed number of schools per capita increases by one unit, the mean overall literacy rate increases by 3.5.

Model 2: Once we had our baseline model, we decided to add in the state code factor variable. This larger model, which we labeled as Model 2, removed 35 degrees of freedom from our initial model. Moreover, in this model, 13 of the continuous predictor variables were significant, along with 15 of the state codes. Variables that we predicted to have a strong association with overall literacy rate based on the linearity scatter-plots (i.e. % of urban population per district, % of population from ages 0 to 6 per district, and % of schools with computers per district) were all statistically significant predictors.

It is unsurprising that we have a larger number of continuous predictor variables that have a significant association with overall literacy rate once we control for the state code factor. In Model 1, it is plausible that the sheer differences amongst states (i.e. cultural norms and attitudes, state government politics, etc.) explain a portion of the response variable, thus masking the effects that some of the predictors in our model have on overall literacy rate. In order to control for these unknown factors and identify the true associations between the predictor variables and the response, it makes intuitive sense to control for the state code factor.

ESS *F*-Test: To further our analysis of the improvement of Model 2 compared to Model 1, we conducted an extra sum-of-squares (ESS) *F*-test. The null hypothesis of this hypothesis test states that the added predictor variables in Model 2 compared to Model 1 have no explanatory effect on the overall literacy rate response. Meanwhile, the alternative hypothesis holds that some combination of the added predictors is useful in better explaining the response. In our ESS *F*-test, which we modeled as an ANOVA test between Models 1 and 2, we obtained a p-value that was extremely significant at an α-level of 0.05. As a result, we concluded that Model 2 is a statistically significantly better explanatory model than Model 1.

Model 3: In our third regression model, we performed a stepwise sequential selection in both directions, starting with Model 2, and using BIC as our model selection criterion. We decided on BIC because of its stricter penalty term for adding more predictors; AIC can often be too liberal in choosing predictors and opt for more complex models, while BIC chooses the most parsimonious model. In our sequential selection method, we set the intercept-only model as the "lower bound" and a model with certain interaction and quadratic terms as the "upper bound". The quadratic terms included were those identified earlier, namely District Area, Percentage of Rural Government Schools, and Percentage of Rural Private Schools; however, after stepwise selection, no quadratic terms remained. We included interaction terms for all predictors interacting with state code factors, as well as several others that made intuitive sense. For example, we thought that the Per Capita Number of Schools might have an influence on the effect that the Percentage of Single Teacher Schools or Percentage of Single Classroom Schools has on overall literacy rates. A full list of the interaction terms included in the "upper bound" model is on pages 28 and 29 of the Appendix.

It is important to note, however, that the only interaction term included in Model 3 after the stepwise sequential selection was an interaction between the Percentage of Scheduled Tribes and the Percentage of Private Schools per district. Both of these predictor variable main effects have a significant negative association with the overall literacy rate response. However, the significant interaction term suggests that at different percentages of Scheduled Tribes per district population, there is a more positive association between the proportion of private schools and overall literacy rates. Perhaps this is because in communities with higher under-privileged populations, any schooling (even if traditionally under-resourced private schooling), is positively associated with

overall literacy. Conversely, in communities with lower under-privileged populations, where the majority of students attend better-resourced government schools, the presence of more private schools is associated with poorer overall literacy outcomes.

Model 4: Model 4 focused on addressing the lack of independence in the data points in our previous models. Districts in the same state are dependent in that they are all affected by state government policies and state-specific cultural factors. Although Model 2 and Model 3 both control for state-level information by including the state code factor variable, 35 additional predictors are added as a result, causing the ratio of data points to predictors to decrease and thus potentially overfitting the data. Therefore, another way to address this violation of independence – a core assumption of multiple regression – is to build a mixed effects model, as we did in Model 4.

Our mixed effects model consisted of a random intercept based on state code and the 14 continuous variables selected in Model 3. A random intercept for state allowed us to resolve the hierarchical structure of the data by assuming different baseline values for each district, based on the state to which it belongs. The results from our mixed effects model were promising: the 14 continuous variables that were selected in Model 3 were all extremely significant (with p-values below 0.001), the coefficients were similar in value to those in Model 3, and the number of predictors significantly decreased, which prevents the chance of overfitting our data.

Models 5 and 6: Models 5 and 6 use Lasso and Ridge Regression respectively to penalize β coefficients in order to highlight the most important variables. Lasso shrinks parameters unevenly by driving the values of some coefficients to zero, thus providing a relatively sparse solution. Conversely, Ridge shrinks parameters more evenly and keeps all coefficients in the model. Because both of these methods are useful for minimizing the effects of multicollinearity, we decided to implement them here. As we saw from our heat map above, certain predictors are strongly positively or negatively correlated with one another; even after stepwise sequential selection, some of the most significant predictors are highly correlated with one another, such as Percentage of Government Schools and Percentage of Private Schools. Thus, we built our Lasso and Ridge Regressions based off of the main effects and interaction terms in Model 3 in order to shrink coefficients of collinear predictors and theoretically improve the interpretability of our model.

After tuning our regularization parameter, λ, in **Figure 6**, we find that the mean sum of squared errors is minimized when λ is relatively small for both Ridge and Lasso, implying that the constraints on the coefficients are not too strict; had λ been larger, more coefficients would have been shrunk towards zero because we would be paying a larger penalty for non-zero coefficients.

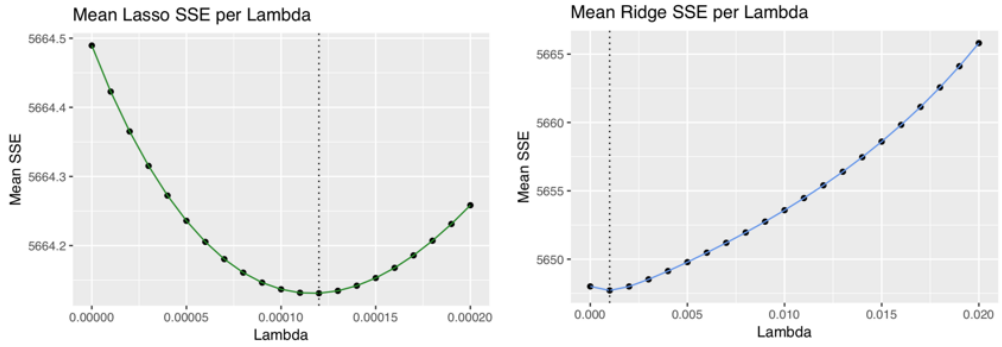


Figure 6: Mean Lasso and Ridge SSEs per Lambda

III. Results and Analysis

After generating the above regression models, we analyzed whether our models passed the assumptions necessary for linear regression, namely linearity, constant variance, and normality. Independence of residuals is the fourth requirement for linear regression, but we addressed violations of this assumption above by including state code as a predictor or as a random intercept in Models 2-6. We found the other three assumptions to be generally held throughout our different models.

Based on the Fitted vs. Residuals plot for each model, Models 1-4 only contained slight deviations from linearity (nothing major enough to warrant adjustments), with our graphs centered around zero, as expected. Moreover, these plots also displayed homoscedasticity, as the regressions had generally constant variance regardless of predicted value. By observing the QQ-plots for each model, we next saw that all four were mostly normal, with only a small degree of left skew. **Figure 7** displays the diagnostic plots for Model 3 and Model 4. Had our models violated some of these assumptions, we would have considered using alternative regression approaches; for instance, fanning of our residuals could have been corrected with weighted least squares regression or robust standard errors, while strong non-normality of residuals could have been improved via resampling methods like bootstrapping. Nevertheless, in this case, the models roughly satisfied all assumptions, allowing us to forego these alternate approaches.

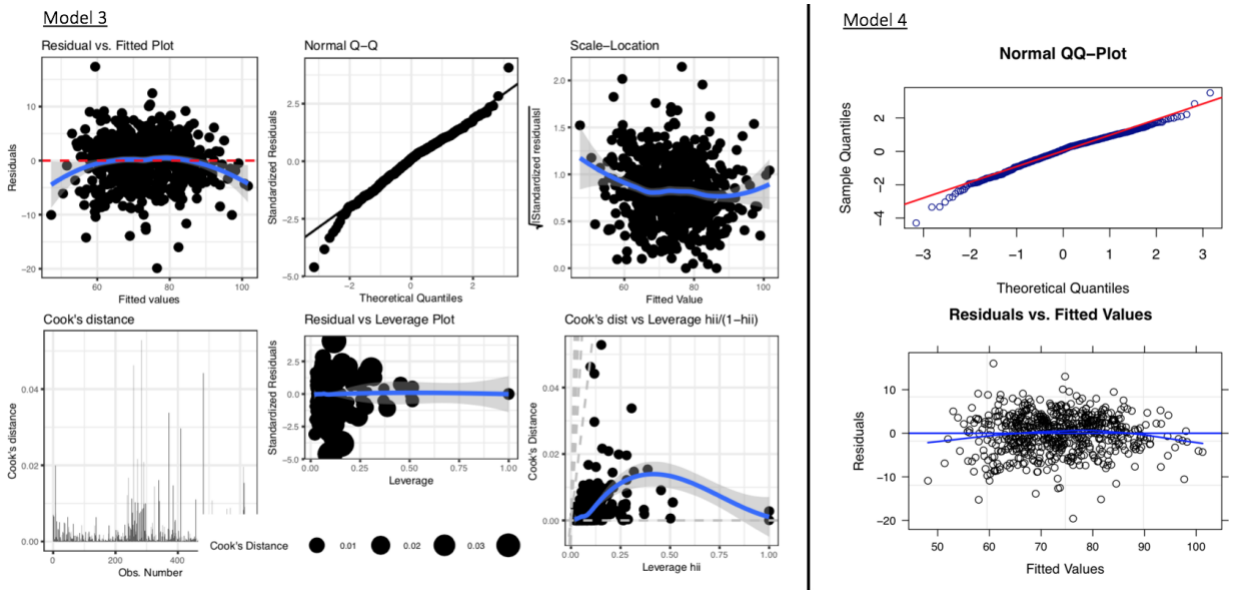


Figure 7: Diagnostic Plots to Check Assumptions

We then analyzed our models for the presence of significant influential points. In examining plots for Cook's Distance, the regression models displayed no points with a Cook's Distance anywhere near 1 (the maximum Cook's Distance was below 0.1). As an example, refer to **Figure 7**, which displays plots for the influential points in Model 3. See pages 35-37 in the Appendix for the remaining diagnostic plots.

After verifying the assumptions of our models, we employed cross-validation techniques to explore which of our models was best for predicting overall literacy rates. When setting up our cross-

validation, we had to ensure not to put all the data from any one state into either the testing or training set, as state is a predictor in most of our models. This issue came up because some of the states have a small number of districts, and three of them only have one: Dadra & Nagar Haveli, Lakshadweep, and Chandigarh. We thus removed these three states from the dataset and then set up our training set to include a maximum of 65% of the districts from every state. This resulted in a training set of size 404, and a test set of size 218. For each of the Models 1 through 4, we fit our model to the training set and then tested it on the remaining 218 districts; we repeated this process 200 times, with new training and testing sets in every iteration. We then calculated the sum of squared errors (SSE) for each iteration, and then took the mean SSE, the results of which are shown in **Figure 8**.

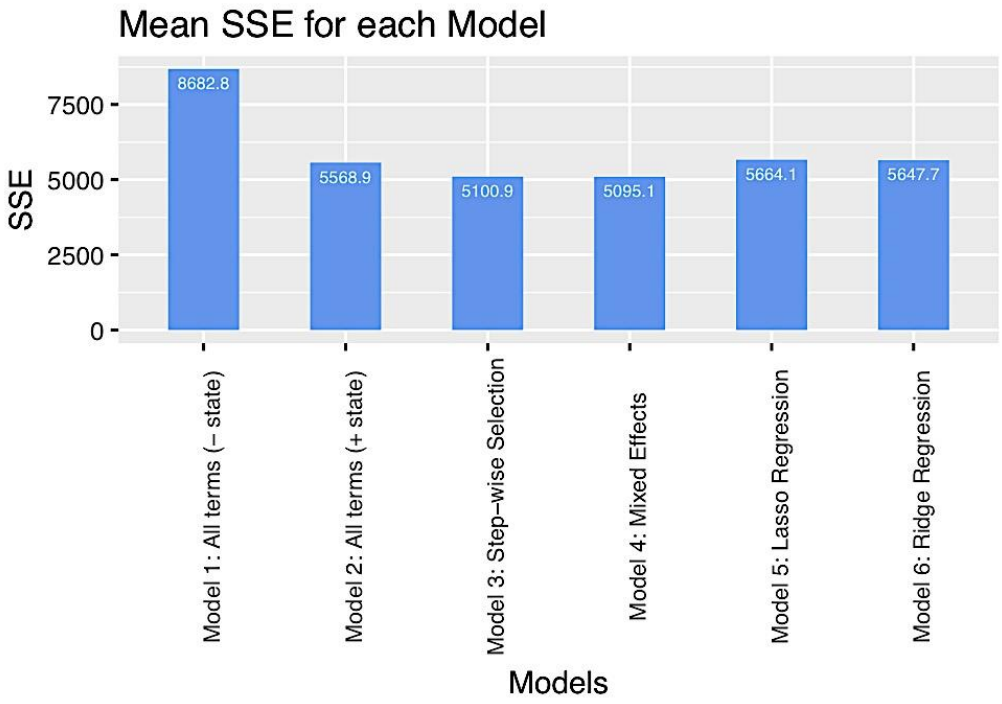


Figure 8: Cross-Validation SSE Results

As can be seen, Model 4, the mixed effects model, had the lowest SSE, with Model 3, the stepwise-selection model performing very similarly. The SSEs for both of these models were likely very similar because (beyond the state code factor), both models had the exact same predictors. Moreover, both models accounted for the state code factors in their results. Model 4 might have performed slightly better than Model 3 simply because it consisted of less predictors and hence had less risk of overfitting the data. Interestingly, the Lasso and Ridge models performed worse than Model 2, which had all main predictors including the state code factor variable. One potential reason for this was that we started our Lasso and Ridge regressions with Model 3, which had already undergone a stepwise selection process, rather than including all potential predictor main effects and interaction terms. As a majority of the multicollinearity present in the full model (with almost all main effects and interaction terms) was likely filtered out in the stepwise selection, Lasso and Ridge are less useful in further simplifying the regression model; we see this in the fact that regularization parameters are quite small. Moreover, even if Lasso and Ridge removed any remaining multicollinearity, the predictive power of these models would not necessarily improve; as discussed

above, multicollinearity does not affect reliability of a model. Finally, it makes sense that Model 1 performed the worst, as it does not control for the inherent variability among states.

IV. Discussion and Conclusions

Throughout this project, we faced a few obstacles, as highlighted in this section. One of the main challenges we encountered had to do with our predictor variables, some of which we were unable to transform to distributions that were satisfactorily normal. For example, the Percentage of Schools with Computers histogram ended up resembling a bimodal distribution, and the Percentage of Single Classrooms Schools histogram remained substantially right skewed, even after being transformed. In addition, there were a few variables that were extremely left skewed, and we were not able to transform them at all. Thus, we left these variables - including Percentage of Road-Accessible Schools, Percentage of Schools with Drinking Water, and Percentage of Schools with Electricity - untransformed in our model. Similarly, Percentage of Unrecognized Schools was right-skewed with a large number of zeros, and even transforming after adding one to each data point did not resolve the right skew.

Another key challenge we faced was in the addition of interaction terms in the "upper bound" of our stepwise sequential selection (Model 3). Originally, our approach was for the "upper bound" to be a full model with all predictor variable main effects and interaction terms. This resulted in Model 3 containing several significant interaction terms that did not make intuitive sense; moreover, this model was likely overfit. Our task then, was to identify a set of interaction terms to include in the "upper bound" that were intuitively plausible. This was challenging as there was no automatic method to determine these optimal interaction terms. Thus, we ultimately hard-coded in the interaction terms deemed to be most plausible.

Finally, we struggled to split our training and testing sets during cross-validation because certain states consisted of only one or two districts. Originally, we randomly split the data into training and testing sets, and therefore, not all states were represented in each set; this was problematic because most of our models included each state as a predictor. Thus, we used a stratified sampling method in the R library *splitstackshape* to fix this issue. Additionally, we had to remove data points that belonged to states with only one district, because these could not possibly be represented in both the training and testing sets. Fortunately, we only had to remove three data points, thus not affecting the robustness of our model.

Overall, our analyses showed that the mixed effects model (Model 4) was the best predictive model of overall literacy rates. The summary table of Model 4 is shown in **Figure 9**, with 14 significant continuous predictors and a random intercept based on state. A key finding from our regression analyses, which was consistent with our original hypothesis, was the importance of factoring in the state a district was located in. Based on our ESS *F*-test, Model 2 (which included the state code factors), was a better explanatory model than Model 1. Moreover, Models 3 and 4, which were the best predictive models, both accounted for variability among states. This intuitively makes sense given that districts in a certain state are all similarly affected by the state's government, culture, and economic well-being.

Additionally, we hypothesized that predictors such as Percentage of Schools with Computers, Percentage of Scheduled Tribes, and Percentage of Private Schools would all have a significant association with overall literacy rates. The summary tables from Model 3 and Model 4 corroborated

this hypothesis. When we controlled for other effects, the coefficients for each of these predictors were statistically significant.

```
Linear mixed model fit by REML t-tests use Satterthwaite approximations
  to degrees of freedom [lmerMod]

Formula:
overall_lit ~ total_pop + p_urban_pop + p_sched_tribes + area_sqkm +
    p_pop0to6 + p_pop11to13 + p_capita_schools + p_gov_school +
    p_priv_school + p_unrec + p_single_class + p_drink_water +
    p_computer + p_sched_tribes:p_priv_school + (1 | state_code)
   Data: overall_data

REML criterion at convergence: 3801.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2941 -0.6061  0.1015  0.6629  3.5077

Random effects:
 Groups     Name        Variance Std.Dev.
 state_code (Intercept) 47.01    6.856
 Residual               20.83    4.564
Number of obs: 625, groups:  state_code, 36

Fixed effects:
                            Estimate Std. Error         df t value  Pr(>|t|)
(Intercept)                2.041e+02  1.335e+01  6.100e+02  15.294   < 2e-16 ***
total_pop                  1.859e+00  5.199e-01  6.079e+02   3.575  0.000378 ***
p_urban_pop                1.487e+00  3.717e-01  6.023e+02   4.000  7.11e-05 ***
p_sched_tribes            -1.923e-01  2.873e-02  6.016e+02  -6.694  4.97e-11 ***
area_sqkm                 -1.879e+00  3.576e-01  6.089e+02  -5.254  2.06e-07 ***
p_pop0to6                 -2.033e+01  2.253e+00  5.966e+02  -9.022   < 2e-16 ***
p_pop11to13               -1.102e+01  2.163e+00  5.890e+02  -5.094  4.73e-07 ***
p_capita_schools           6.187e+00  9.631e-01  6.043e+02   6.424  2.69e-10 ***
p_gov_school              -4.217e-03  8.074e-04  6.041e+02  -5.223  2.43e-07 ***
p_priv_school             -4.637e+00  1.068e+00  6.070e+02  -4.341  1.66e-05 ***
p_unrec                   -6.545e-01  1.457e-01  6.057e+02  -4.491  8.48e-06 ***
p_single_class            -1.622e+00  4.275e-01  6.100e+02  -3.795  0.000163 ***
p_drink_water             -1.839e-01  3.637e-02  5.995e+02  -5.057  5.67e-07 ***
p_computer                 2.500e+00  7.506e-01  6.078e+02   3.331  0.000918 ***
p_sched_tribes:p_priv_school  3.799e-02  7.283e-03  5.568e+02   5.216  2.59e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9: Summary Table for Mixed Effects Model (Model 4)

The final exploration that we performed was whether the "best" set of predictors for the overall literacy rate response were also useful in predicting the male and female literacy rate responses. Our initial hypothesis stated that the "best" set of predictors for overall literacy rate would indeed be useful in predicting the other two responses. To determine this, we generated mixed effects models for the male and female literacy rate responses with the same set of predictors from Model 4. We then conducted a stepwise backward selection to eliminate any predictors that were not deemed to be useful in predicting male and female literacy rates. As shown in the table in **Figure 10**, for the female literacy rate mixed effects model, all of the predictors from Model 4 were kept in as useful predictors. The same result was seen for the male literacy rate mixed effects model, which was

consistent with our original hypothesis. If we were to continue this project, we would conduct further explorations into which additional variables might be useful in predicting male literacy rates and female literacy rates. We would start with the original 25 predictors in our dataset and then iterate through the same six models we used for overall literacy rate to determine if we arrived at a similar set of useful predictors as in this paper.

```
Random effects:
           Chi.sq Chi.DF elim.num p.value
state_code 327.51      1     kept < 1e-07

Fixed effects:
                                Sum Sq   Mean Sq NumDF  DenDF F.value elim.num Pr(>F)
total_pop                     423.4428  423.4428     1 609.63 15.9991     kept  1e-04
p_urban_pop                   559.3667  559.3667     1 600.42 21.1348     kept  0e+00
p_sched_tribes                698.0812  698.0812     1 605.41 26.3759     kept  0e+00
area_sqkm                    1148.0272 1148.0272     1 609.91 43.3764     kept <1e-07
p_pop0to6                    2553.1952 2553.1952     1 594.89 96.4685     kept <1e-07
p_pop11to13                   643.7911  643.7911     1 587.84 24.3246     kept  0e+00
p_capita_schools              708.8471  708.8471     1 607.65 26.7827     kept  0e+00
p_gov_school                  558.9578  558.9578     1 602.31 21.1193     kept  0e+00
p_priv_school                 312.1568  312.1568     1 605.36 11.7944     kept 0.0006
p_unrec                       337.8552  337.8552     1 603.82 12.7653     kept 0.0004
p_single_class                214.5792  214.5792     1 609.65  8.1075     kept 0.0046
p_drink_water                 377.7068  377.7068     1 597.45 14.2711     kept 0.0002
p_computer                    133.9555  133.9555     1 609.65  5.0613     kept 0.0248
p_sched_tribes:p_priv_school  641.1879  641.1879     1 572.58 24.2263     kept  0e+00


Least squares means:
     Estimate Standard Error DF t-value Lower CI Upper CI p-value

 Differences of LSMEANS:
     Estimate Standard Error DF t-value Lower CI Upper CI p-value

Final model:
lme4::lmer(formula = female_lit ~ total_pop + p_urban_pop + p_sched_tribes +
    area_sqkm + p_pop0to6 + p_pop11to13 + p_capita_schools +
    p_gov_school + p_priv_school + p_unrec + p_single_class +
    p_drink_water + p_computer + p_sched_tribes:p_priv_school +
    (1 | state_code), data = female_data)
```

Figure 10: Summary Table for Female Literacy Mixed Effects Model (Model 8)

In conclusion, our project was successful in confirming the three hypotheses with which we had started:

1. That the state a district is in is an important factor in predicting that district's overall literacy rate
2. That factors such as Percentage of Scheduled Tribes, Percentage of Private Schools, and Percentage of Schools with Computers are statistically significant predictors of a district's overall literacy rate
3. That predictors that are useful in predicting overall literacy rates are also useful in predicting male and female literacy rates.