Investigating Literacy Rates in US Counties

Sophie Moore & Eliza Jacobson

March 2021

Contents

Introduction

While populations are increasing and technology is simultaneously breaking down barriers between us, the ability to communicate and interact with others is critical. As a political science major and an anthropology/geography major, we are students who are interested in people, politics, culture and the environment. Using statistics in social sciences is essential in making policy conclusions, analyzing social and behavior changes, and answering cultural questions, which is why we chose to focus our final project on literacy rates. Literacy rates can be used to assess health, social progress and economic achievement. Numerous studies have found correlations between high literacy rates and better economic opportunity, health, and even environmental sustainability. One study noted that "it is unclear whether the poor health status of illiterate individuals in the U.S. is related to illiteracy itself, or to other associated sociodemographic factors," but "literate women are more likely to participate in family planning services... [and] improved knowledge about and utilization of family planning information can, in turn, decrease birth rates and family size, both of which are important factors in improving the health status of women and children in nonindustrialized countries" (Weiss 1991). A later research journal found bi-causality between income per capita and literacy rates with results that conclude that "higher literacy rates lead to increased health expenditures due to increased demand of health services" (Mehmood 2014).

In our analysis, we wondered if we could predict literacy rates based on factors such as poverty levels and unemployment. Thus, we conducted linear regressions to model the relationships between our primary response variable, literacy rates in percentages, and explanatory variables including education less than high school, population below poverty level, unemployment, and state. To acquire data, we originally created a dataset from the National Center for Education Statistics and World Population Review on state-by-state expenditures and literacy rates. Since it focused on states, we only had 50 observations. After we did a few initial tests, we decided to dive deeper into the statistics by expanding our observations by choosing a dataset that focused on literacy rates per county from the Programme for the International Assessment of Adult Competencies (PIAAC). This program assesses and analyzes adult skills including literacy, numeracy, and problem solving in over 40 countries. With US counties as the new observational unit, there are 3,412 observations. From this dataset, we decided to use states as a categorical variable. Data about our other categorical variable, geographic region (north, south, northeast, west or midwest), was found on the US Department of Commerce's Bureau of Economic Analysis website, but we assigned counties a region manually in our dataset during the data cleaning stage. Since we collected data from two credible sources, we had to compile our own spreadsheet of all our variables with a codebook that contains descriptions and sources.

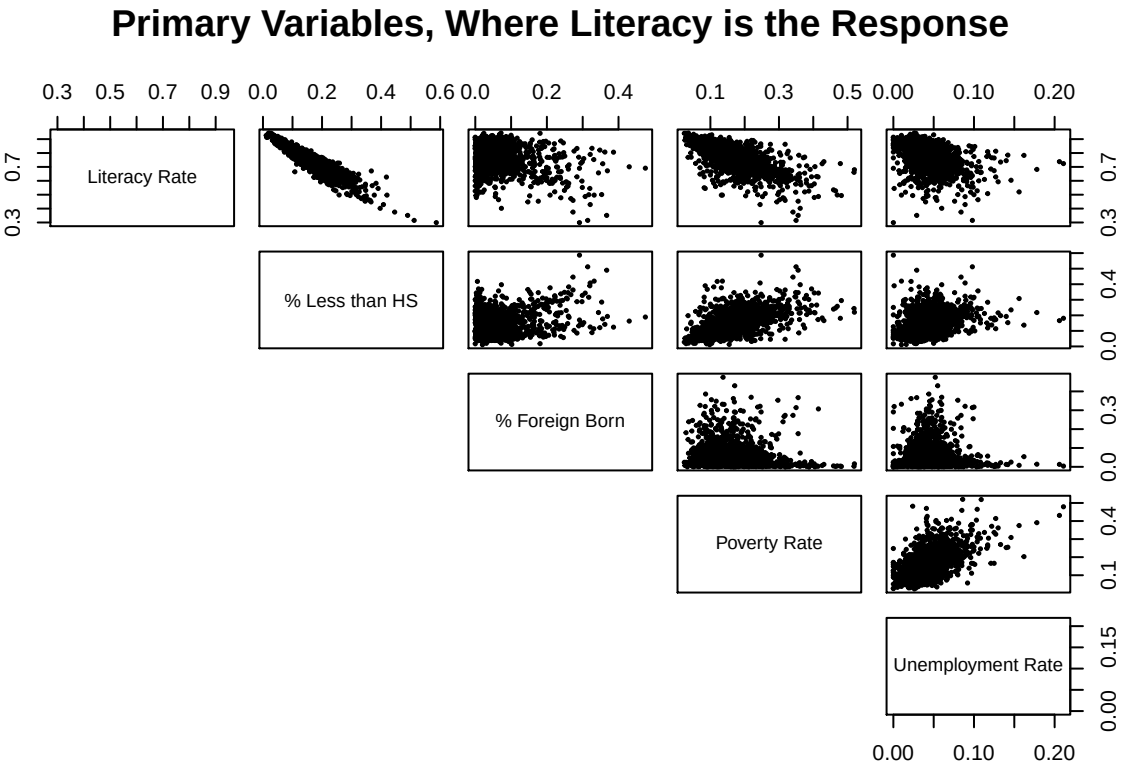
Split the Data

The size of `train.data` will be 80% of the 3142 observations, which is 2513. The variable `train_ind` stores a randomly-selected list of 2513 index numbers, which correspond to observations (US counties) in the data set. We used '123' as our seed number for the set.seed() function when randomly choosing the indexes of the observations for `train.data` in order to ensure the same "random" numbers generate each time. `train.data` is filled with the observations at these 2513 rows in the data, and `test.data` is populated with the remaining 629 counties.

The point of splitting the original data into these two groups is that we will use the training data set to "train" our model and create visualizations, then use the testing data set to test the model. This is a good way to check model validity, because the 629 counties in `test.data` will not be used to create our model.

Data Visualization

Matrix Plot



Based on the matrix plot, it is very obvious that there is an extremely strong, negative, linear relationship between literacy rate and the proportion of the population with less than a high school education, with a correlation coefficient of -0.947. There are also negative relationships between literacy rate and the proportion of the population that is foreign born (r=-0.264), the poverty rate (r=-0.720), and the unemployment rate (r=-0.397) respectively that all appear significant, though less strong than the first relationship.

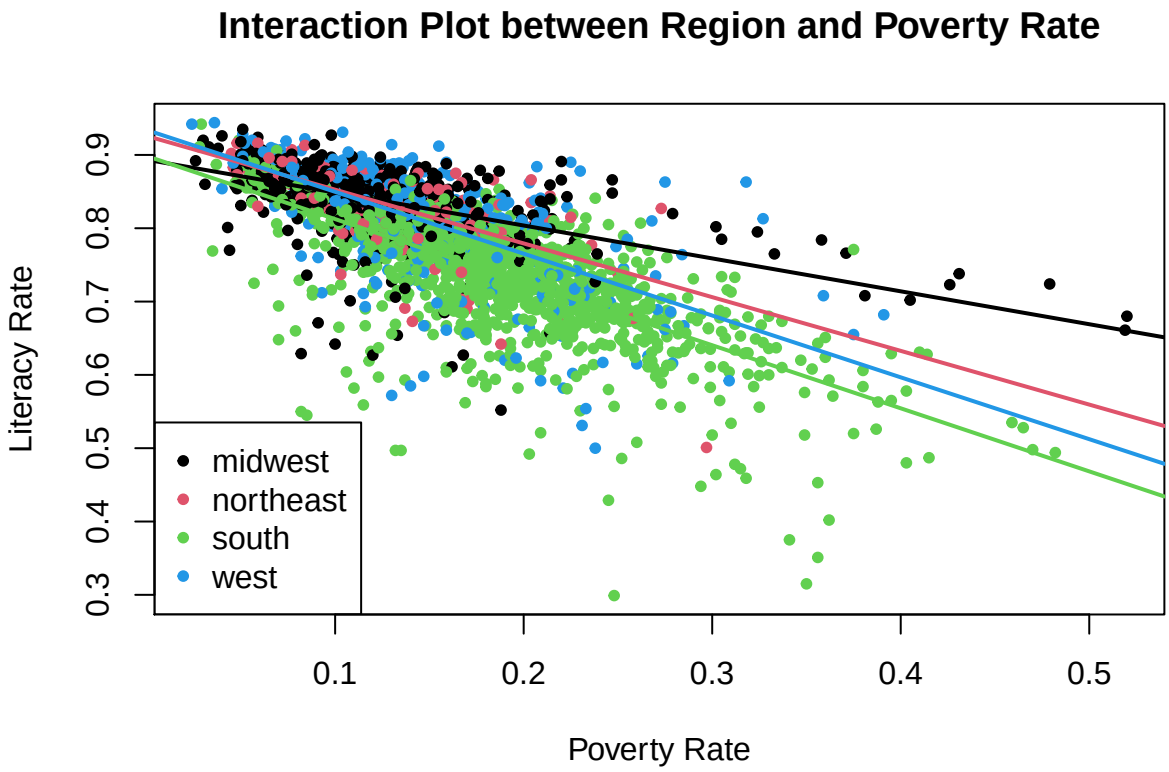
There are definitely positive linear relationships between a few predictors, such as percentage with less than HS education and poverty rate (r=0.633); Less_HS and the unemployment rate (r=0.328); and the poverty rate and unemployment rate (r=0.536), all three of which relationships make logical sense.

The relationship between literacy rate and percentage foreign born could possibly use a transformation to improve linearity and equal variance. The relationships where percentage with less than HS education and poverty rate are the predictors appear to have equal variance issues, so a transformation of literacy rate could help both of those.

Correlation Matrix

	Literate	Less_HS	FB	Poverty_100	Unemployed
Literate	1.0000000	-0.9471220	-0.2644974	-0.7197354	-0.3968121
Less_HS	-0.9471220	1.0000000	0.2251185	0.6330536	0.3284220
FB	-0.2644974	0.2251185	1.0000000	-0.0694472	0.0184059
Poverty_100	-0.7197354	0.6330536	-0.0694472	1.0000000	0.5358739
Unemployed	-0.3968121	0.3284220	0.0184059	0.5358739	1.0000000

Interaction Plot

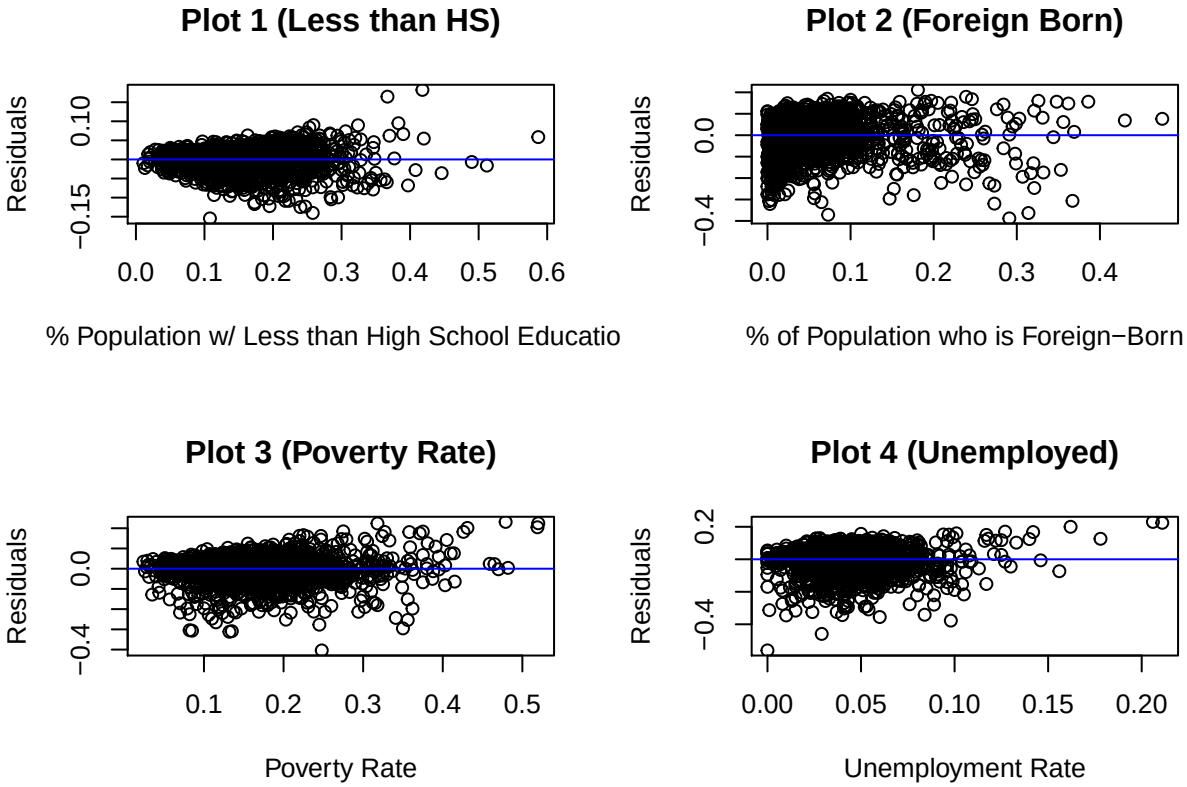


In context, an interaction between the poverty rate and region would mean that the region of the United States that a given county is located in has an impact on how the poverty rate affects the literacy rate of that county, i.e., the poverty rate affects the literacy rates differently in different regions of the US.

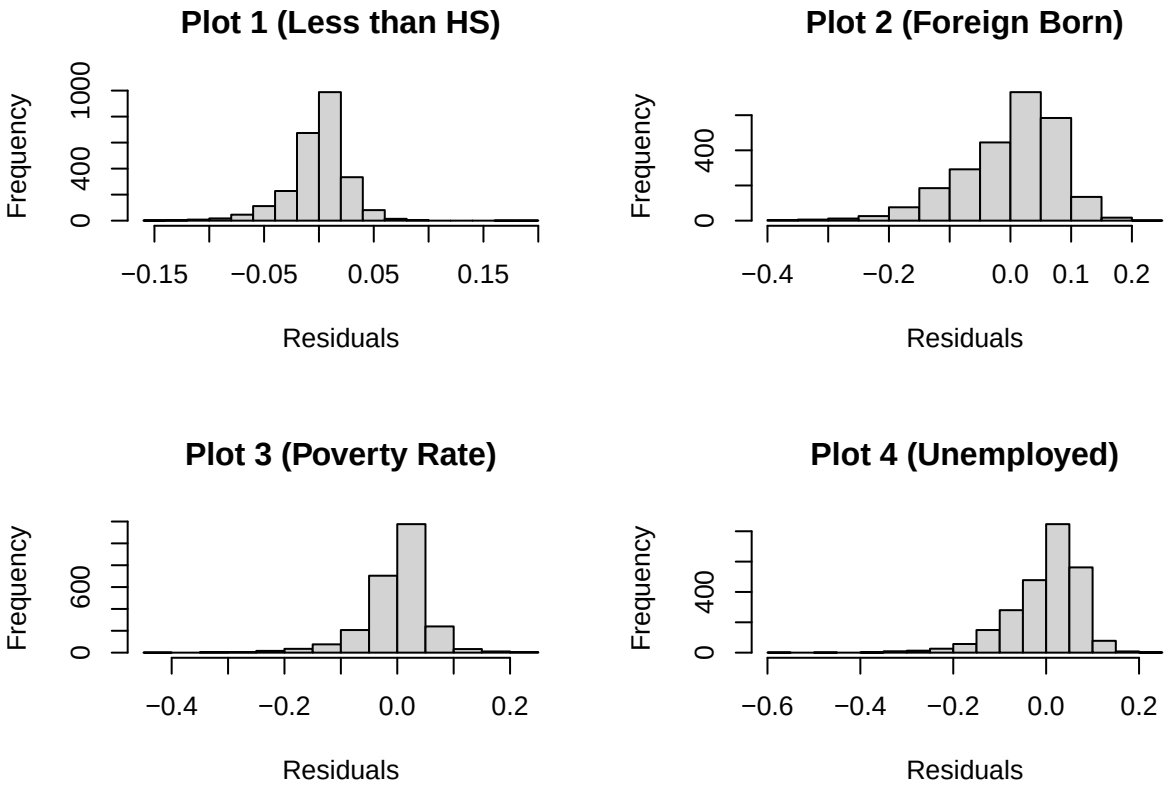
Based on the interaction plot, it appears that a parallel lines plot might be appropriate for the northeast, south, and west, but the midwest line could be making the interaction significant, although it is impossible to tell without running the formal hypothesis test.

Variable Pre-Processing

Initial Residual Plots



The initial residual plots of Literacy Rate with each quantitative predictor are shown above. Plot 1 shows no obvious violations of linearity but a clear equal variance problem. Plot 2 doesn't seem show to show a linearity issue, but there might be an issue with equal variance. Plots 3 and 4 don't have any huge linearity issues but there does seem to be a bit of an equal variance violation in both plots.



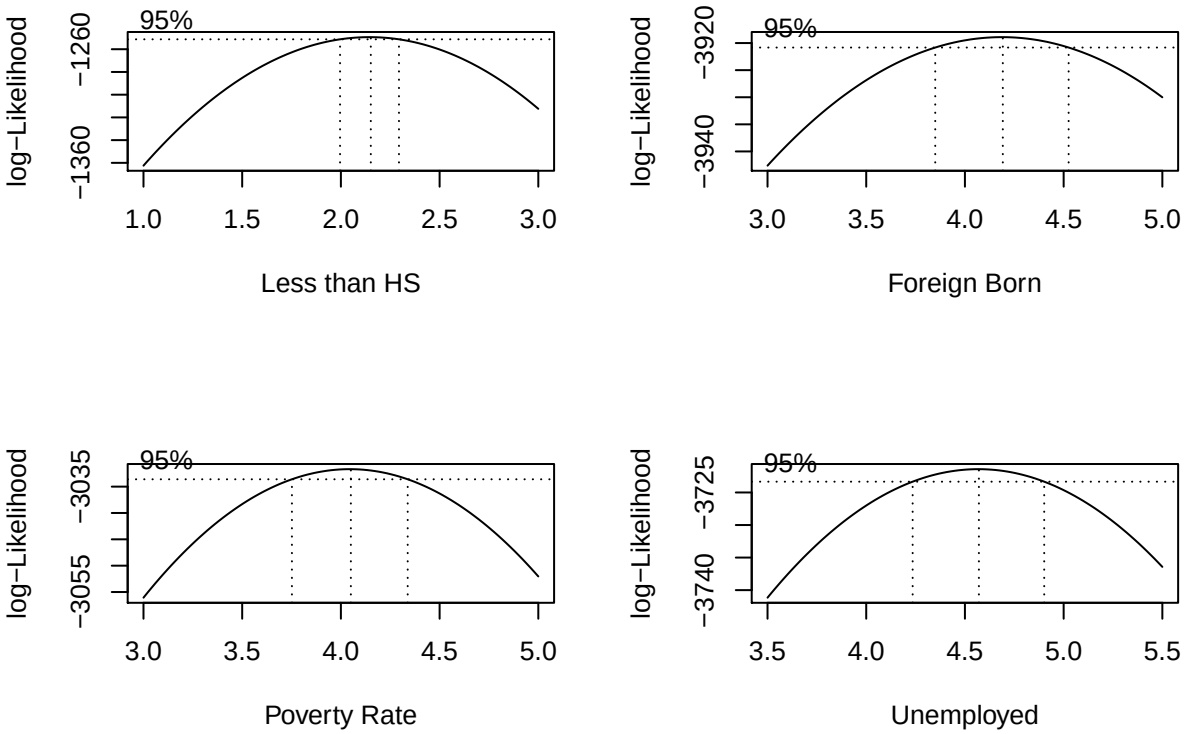
The initial histograms of the residuals from literacy rate and each quantitative predictor are shown above. There are no egregious deviations from normality in any of the plots, although plots 2 and 4 seem somewhat right-skewed.

Preliminary Transformations

Because the main issue identified in the residual plots of literacy rate and each predictor was equal variance, along with some minor issues with linearity, we decided that a transformation of y would be the most likely to fix the problems. We also tried multiple linearizing transformations on the predictors `Less_HS` and `FB`, but nothing seemed to help linearity, so these plots are not included.

Below are the Box Cox plots for literacy rate and each predictor, which was a good place to start in order to determine the correct transformation.

Box Cox



The Box Cox plots suggest raising the response variable to a higher power, but unfortunately they do not all include the same power within the 95% confidence intervals. The power transformation that appears closest to all four confidence intervals is 4. We attempted to include those residual plots below but unfortunately the LaTeX processor would not allow us to do so. Prior to $\texttt{Literate}^4$, we attempted $\texttt{Literate}^2$ and the natural log of Literate, but neither were fruitful for transforming the individual relationships between the response and each predictor.

Unusual Observations

We believe that none of the relationships between literacy rate and an individual relationship have many extreme outliers. One thing we noted is that the somewhat-unusual observations in all four of the relationships plotted above are to the right of the point clouds, and none are to the left. This means that these relationships between the literacy rates and the four predictors are right skewed, which is apparent in the histograms included previously.

Proposed Model

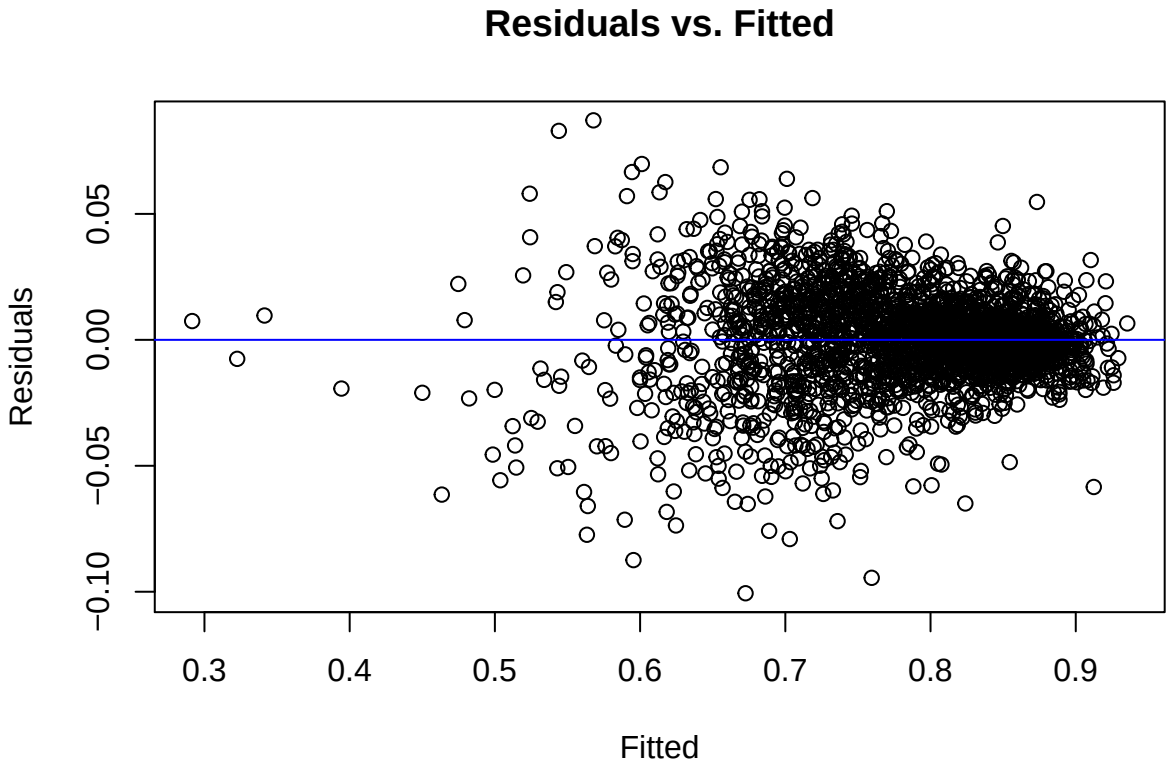
The first model we propose is a linear regression on the literacy rate of the county with the predictors: proportion of the population over age 25 who have less than a high school education, the proportion of the population who is foreign born, the poverty rate, the unemployment rate, and the US state that the county is located in. The abbreviated summary table with overall characteristics of the model is displayed below.

Table 1: Overall Model Characteristics

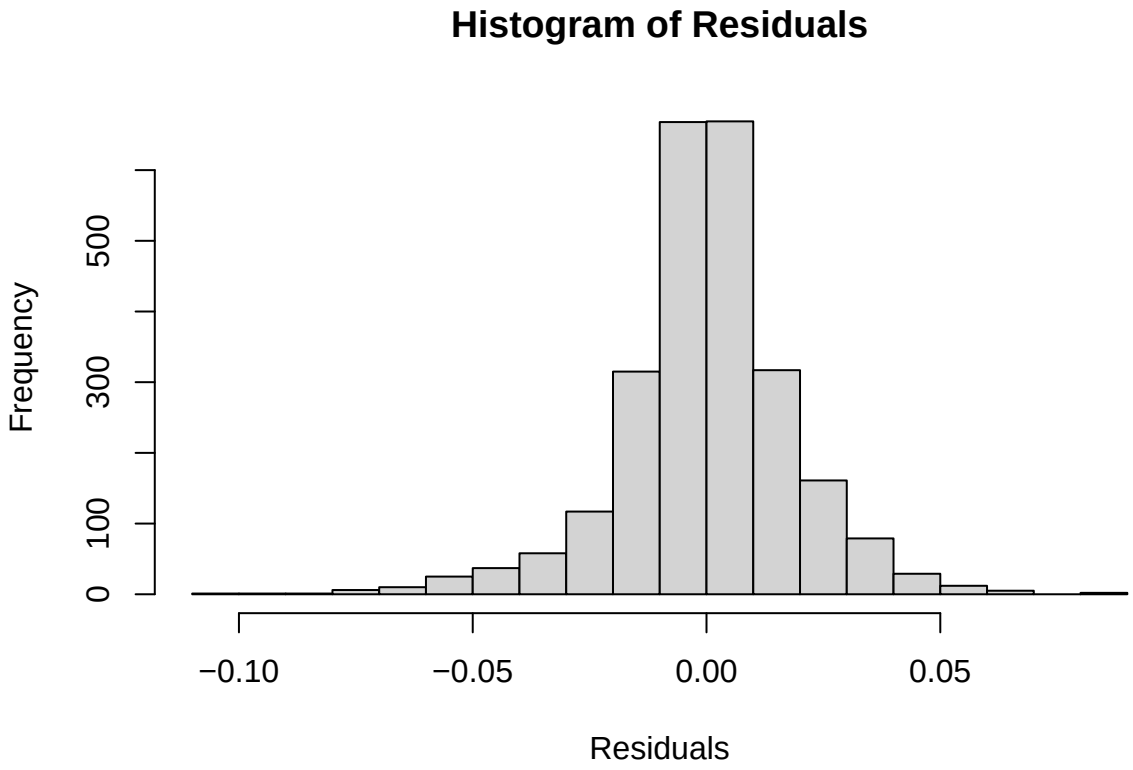
Rsq	Rsq_adj	s	F.statistic	p.value
0.9494	0.9483	0.01893	871.2	<0.001

Residual Analysis

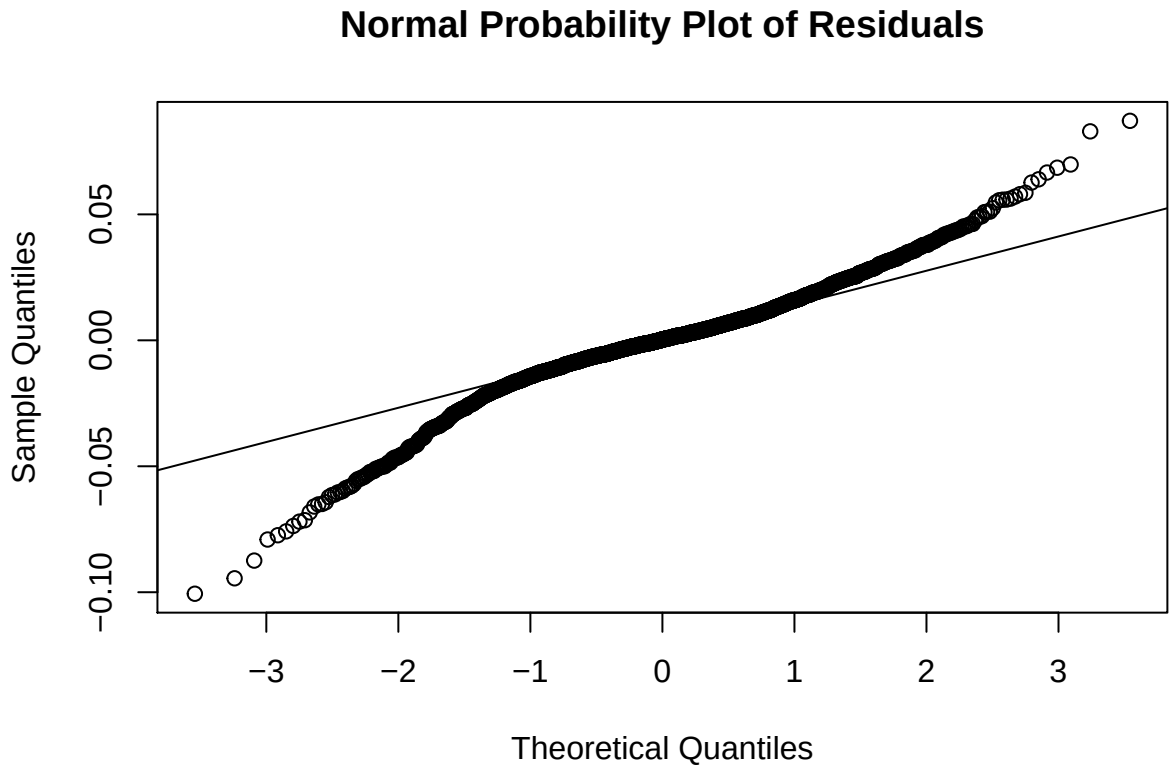
Model Conditions



The Residuals vs. Fitted plot for this model displays very clear violations of equal variance due to the pronounced fanning of the residuals. The linearity condition is fine because while there is fanning, there isn't obvious curvature or other patterns in the Residuals vs. Fitted plot.



The histogram of residuals looks very normal and bell-shaped, which makes sense given the Central Limit Theorem and the large size of our data set.

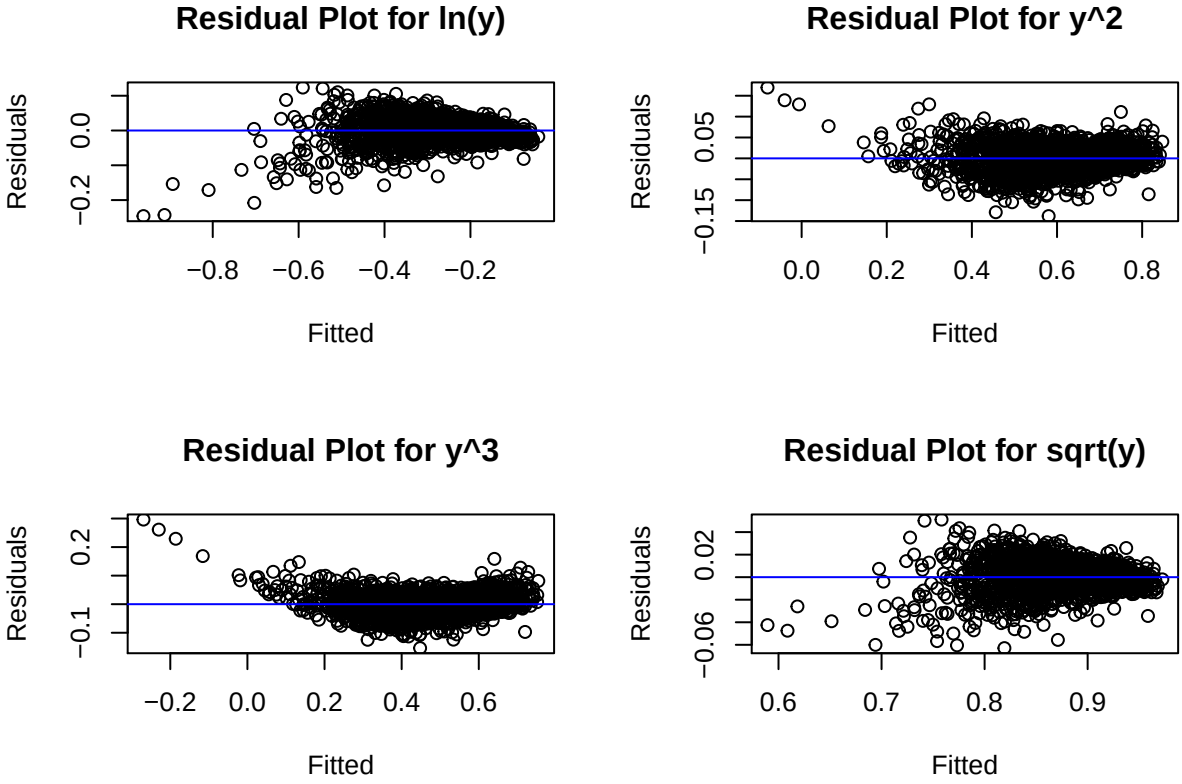


The normal probability plot of residuals definitely shows some issues with normality, but overall this condition is satisfied because of how good the histogram of residuals looks.

The independence condition cannot be checked with residual plots, but we assume that it is satisfied because each observation is an individual US county, and data provided by the National Center for Education Statistics (NCES) are usually reliable.

Just like in the Residuals vs. Predictor plots that were displayed earlier in this report, the residual analysis revealed that the main model condition that is not satisfied is equal variance. Usually when there are issues with equal variance, a transformation of the response variable is the most appropriate, so that is what we tried.

Below are four Residual vs. Fitted plots for this model that we tried, where the response variable literacy rate is transformed.

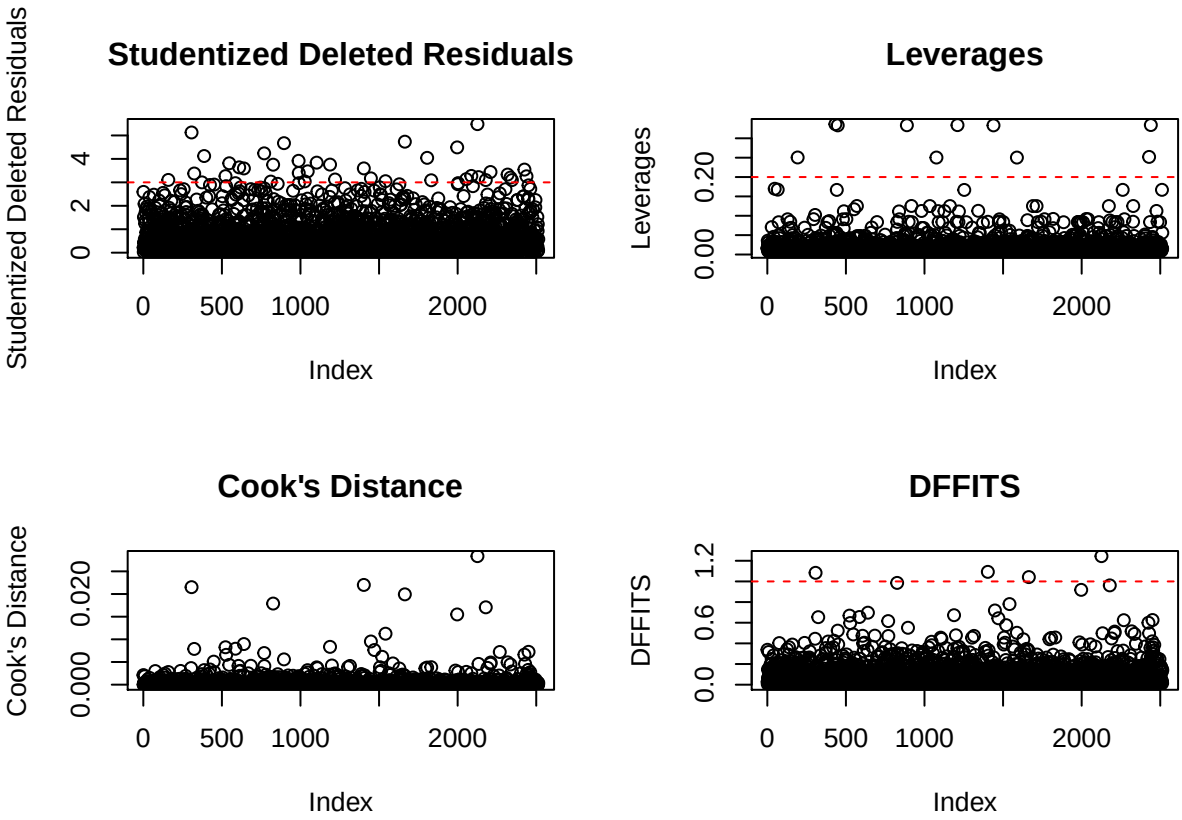


None of the transformations of literacy rate that we attempted (including the four pictures and a few others) appeared to help the equal variance condition without messing up linearity. The Residual vs. Fitted plot that looks the best was for the model with the literacy rate squared, in which the equal variance condition seems to be somewhat better. We ran the Box Cox procedure, and $\lambda = 2$ was right in the middle of the confidence interval, so it confirmed what we found in the residual plot. However, the linearity condition is worse for this relationship, and a couple of outliers that didn't previously exist were created in the process.

Bearing in mind our cycles of residual analysis and re-fitting, we decide to keep the original model and move forward with it for the rest of our analysis. Since the model with literacy rate squared is not too much worse than our first model, we will compare these two models again during the model validation.

Case Influence Diagnostics

After confirming our choice of model, we found the main case influence statistics and plotted them below.



The case influence diagnostic plots show that there are a moderate amount of observations with a studentized deleted residual that have an absolute value greater than 3 (meaning they are in the tail of the t-distribution); 10 observations with high leverage; and a handful with a large influence on the predicted values from the DFFITS plot.

We found that there are four counties above the cutoff points (plotted as the red dashed lines) that have both a large studentized deleted residual and a large DFFITS score. Three of those counties–Mora County, Mckinley County, and Guadalupe County–are located in New Mexico, while the remaining county–Greenlee County–is located in Arizona. It is likely significant that three of the four counties that our analysis flagged as potential influential points are in the same state, although further study would be required to figure out what is going on there.

Fit a Linear Model

Linear Model Equation

The equation for our proposed model is included below. (Note: states are abbreviated to their two digit postal code).

$$\hat{Literate} = 0.959 - .95LessHS - 0.13FB - 0.27Poverty100 - 0.13Unemployed + 0.04AK + 0.04AZ + 0.008AR$$
$$+0.008CA + 0.004CO + 0.02CT - 0.008DE - 0.02DC + 0.006FL + 0.006GA + 0.01HI + 0.02ID + 0.02IL + 0.02IN$$
$$+0.01IA + 0.01KA + 0.03KY - 0.01LA + 0.03ME - 0.004MD + 0.02MA + 0.02MI + 0.02MN - 0.01MS$$
$$+0.01MO + 0.02MT + 0.008NE + 0.008NV + 0.03NH + 0.001NW - 0.04NM + 0.01NY + 0.01NC + 0.02ND$$
$$+0.02OH + 0.02OK + 0.02OR + 0.009PA + 0.03RI - 0.007SC + 0.02SD + 0.02TN - 0.003TX + 0.02UT + 0.02VT$$
$$+0.005VA + 0.02WA + 0.02WV + 0.02WI + 0.02WY$$

Contextualized Model

Based on logic alone, there is an apparent relationship between level of education and literacy rates, which our model supports. Our model implies that the more people who graduate high school, the higher literacy rates will be. Although we cannot make any definite statements about causality, it makes sense that more education will lead to improved literacy rates because one must be able to have a proficient level of reading and writing skills to earn a diploma. Furthermore, it is fair to assume that counties that have large populations of foreign born residents would have lower English literacy rates because English might not have been their first language, so the negative regression coefficient for the Foreign Born variable makes sense as well. However, it is unclear which way causality flows between literacy rates, poverty and unemployment, so it would not be appropriate to contextualize the negative relationships that appear in our model. Nevertheless, it is logical that counties with high rates of poverty will have lower rates of not only literacy but also completion of secondary education.

Model Behavior

The variables in our model fit the data pretty well, based on R^2, adjusted R^2, and s. R^2 is 0.9494, so 94.94% of the variation in literacy rates in US counties can be explained by the regression model with the predictors proportion of population with less than a high school education, proportion of population who is foreign born, poverty rate, unemployment rate, and US state. The adjusted R^2 is 0.9483, which is a good sign that there are not unnecessary predictors since it is so close to the R^2.

The root mean square error (s) is 0.0189, which means the average prediction error is 1.89 percentage points of literacy rate. This is a pretty small s value in context, so we believe that the fit of the model is applicable to our study on literacy rates in US counties.

Regression Coefficient Interpretation

The intercept of our model is 0.9594, so if 0% of the population of a county failed to graduate high school, 0% are foreign born, the poverty rate is 0%, the unemployment rate is 0%, and the state is Alabama (the reference group), the predicted mean literacy rate is 95.94%. It doesn't make very much sense to interpret the intercept since it is virtually impossible for a county to have these characteristics.

The regression coefficient for `Less_HS` is -0.9493, so holding the proportion of foreign born residents in a county, the poverty rate, the unemployment rate, and the state constant, for every one percentage point increase in proportion of population who have less than a high school education, it is predicted that the literacy rate of a county will decrease by 0.9493 percentage points.

The regression coefficient for `FB` is -0.1305, so holding all other predictors constant, for every one percentage point increase in the proportion of county residents who were born outside of the US, we predict literacy rates to decrease by 0.1305 percentage points.

The coefficient for `Poverty` is -0.2745, so holding all other predictors constant, for every one percentage point increase in the poverty rate of a county, we expect literacy rate to decrease by 0.2745 percentage points.

And finally, the coefficient fr `Unemployed` is -0.0939, so holding all other predictors constanr, for every one percentage point increase in the unemployment rate in a county, we predict that the literacy rate will decrease by .0939 percentage points.

For the categorical variable `State`, the reference group is Alabama and the regression coefficient for Alaska is 0.0363. So we predict that holding all four quantitative predictors constant, counties in Alaska will on average have a literacy rate that is 0.0363 percentage points higher than counties in Alabama.

Multicollinearity

In order to examine possible multicollinearity in our data, a table with the variance influence factors (VIFs) for the predictors is included below.

	GVIF	Df	GVIF^(1/(2*Df))
Less_HS	2.566031	1	1.601884
FB	1.701218	1	1.304308
Poverty_100	2.539720	1	1.593650
Unemployed	1.743957	1	1.320590
State	3.640182	49	1.013271

Since the VIFs for all the explanatory variables are between 1 and 5, we can conclude that none of the predictors are highly correlated. Therefore, while there is some moderate multicollinearity in our data, it is not severe enough to warrant further corrective measures.

Statistical Inference

Overall Model Utility Test

For this overall model utility test, the hypotheses are:

H_0: $\beta_{LessHS} = \beta_{FB} = \beta_{Poverty} = \beta_{Unemployed} = \beta_{State} = 0$ H_A: at least one of these $\beta_i's \neq 0$

With a large F-statistic of 871.2 and small p-value < 0.001 from the distribution F(53,2459), we reject the null hypothesis and conclude that the overall model is statistically significant, meaning there is a statistically significant relationship between proportion of population with education less than high school, proportion of population foreign born, poverty rate, and unemployment rate for the model on literacy rates.

Partial F-Test

For this partial F-test, we are testing whether our full model is significantly better than a model that predicts literacy rates from poverty rate alone. The hypotheses for this test are:

H_0: $\beta_{LessHS} = \beta_{FB} = \beta_{Unemployment} = \beta_{State} = 0$ H_A: at least one of these $\beta_i's \neq 0$

The table below displays the output from the ANOVA table for the partial F-test.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2511	8.3977868	NA	NA	NA	NA
2459	0.8809443	52	7.516843	403.4987	0

With a large F-statistic of 403.50 and a small p-value < 0.001 from the distribution F(52,2459), we reject the null hypothesis and conclude that adding the predictors proportion with less than a high school education, proportion foreign born, unemployment rate, and state to the model that only includes poverty rate significantly improved the prediction for literacy rate.

Interaction Model

Next, we will test the significance of the interaction term between region of the United States and poverty rate, which is the relationship depicted in the previous interaction plot. For this test, the hypotheses are:

H_O:$\beta_{region:Poverty} = 0$ H_A:$\beta_{region:Poverty} \neq 0$

The table below displays the output from the ANOVA table for this F-test of the interaction term.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2508	7.104317	NA	NA	NA	NA
2505	6.840437	3	0.2638801	32.21137	0

With a moderately large F stat of 32.211 and small p-value < 0.001, from the distribution F(3,2505), we reject the null hypothesis to conclude that the difference in slopes between the four US regions south, west, midwest, and northeast is statistically significant. Therefore, the impact of poverty on literacy rates is significantly different between regions in the US.

Confidence Intervals

We chose to make a hypothetical county that has the mean values (from the training data) for proportion of population who failed to graduate high school, proportion of population foreign born, poverty rate, and unemployment rate in order to investigate an "average" US county. We placed this county in California because that is where we live.

The table below displays the confidence interval for the mean literacy rate of these "average" counties and the prediction interval for the literacy rate of a future hypothetical "average" county.

Interval.Type	Lower.Bound	Upper.Bound
Confidence	0.77621	0.78771
Prediction	0.74440	0.81952

For these "average" counties, we are 95% confident that the mean literacy rate will be between 77.62% and 78.77%. For a hypothetical "average" county with the same characteristics, we are 95% confident in our prediction that the literacy rate will be between 74.44% and 81.95%.

Model Validation

Model Validation Statistics

In order to validate our model, we will compare some model validations statistics from the model we ultimately chose with our second choice of a model, which was the model with the same predictors, except that the response (literacy rate) is squared. Below is a table that summarizes these statistics between the two models.

Model	MSPR	MSE	Difference	Rsq	Rsq_adj	Rsq_pred	SSE	PRESS	BIC
Literacy Rate	0.01334	0.00035	0.01299	0.9494	0.9483	0.94727	0.8809	0.9187	-12431
Literacy Rate^2	0.04608	0.00076	0.04532	0.9496	0.9486	0.94734	1.9134	2.0008	-10482

We believe that the model validation statistics provides relatively good evidence that our final model has good predictive ability. The MSPR for our model is pretty small at 0.01334 and the MSE is very small at 0.00035. So while our model appears to have overfit the training data slightly in comparison to the predictions for the testing data, it still did a pretty good job, with a difference of only 0.01299. The model with Literacy^2 performed worse for both MSPR, MSE, and the difference between them, so this model definitely overfit the training data.

It is appropriate to compare the R^2 and R^2 adjusted because both models have the same number of predictors, and they are very high for both models. Even though both R^2 and R^2 adjusted are each a tiny bit higher for the Literacy^2 model, it is not a big enough difference to be important. The R^2 prediction is almost also high at 0.9473 both both models. This means that 94.73% of the variability in predicting new observations is explained by both our full model and the model with Literacy^2.

While both models produced PRESS values that are close to their SSE values, our final model has both a smaller difference between SSE and PRESS and smaller values of SSE and PRESS in absolute terms (in comparison to the model with Literacy^2). Our final model also minimized BIC in comparison to the other model.

After computing and comparing these model validation statistics, we are satisfied that we chose the best model we could find, and that this model does a good job both explaining the variability in the data and predicting on new observations.

Fit Model on Full Data

As a final step, we fit our model on our full dataset that includes all 3,142 US counties. Some select summary statistics are displayed in the table below in order to compare the model fit with the training data, the test data, and the full data.

Table 2: Overall Model Characteristics

Data	Rsq	Rsq_adj	s	F.statistic	p.value
Training Data	0.9494	0.9483	0.01890	871.24	<0.001
Testing Data	0.9582	0.9543	0.01740	248.71	<0.001
Full Data	0.9501	0.9492	0.01868	1088.43	<0.001

The summary statistics reveal that our model actually performs better on both the testing data and the full data than it does on the training data. However, the differences are not that large between the model fit on all three data sets, and all three perform well in predicting the literacy rate of a US county from the proportion of residents who did not graduate high school, the proportion of residents born outside the US, the poverty rate, the unemployment rate, and the state.

Conclusion

Our linear regression model proved to be effective in predicting literacy rates of US counties. With a large F-statistic and small p value, an overall F-test determined that the model is significant, meaning there is a statistically significant relationship between education less than high school, foreign birth, poverty, unemployment, states and literacy rates. Furthermore, only twelve out of fifty states did not have slope coefficients that were statistically significant at the 5% level, indicating that state is a useful categorical predictor. Based on the R^2 value, only about 5% of the variation in literacy rates is not explained by our model. We believe that if we were to continue investigating literacy rates, adding race and gender to our model could explain more of the unexplained variation.

As we discussed in the model validation section, when testing our model with the training data, the mean of the squared prediction errors was very small and was relatively similar to the mean squared error. This indicates that the predictive ability of the model is relatively strong. We found that while squaring the Literacy Rate seemed to help equal variance somewhat, in the end it was not a worthwhile transformation because it made the model fit the training data too closely and messed with linearity. In addition, when we transformed literacy, it made interpreting the coefficients much more difficult, so it made sense to keep the simpler model. Critiques of our model include the fact that the MSPE and MSE aren't extremely close and

the equal variance conditions did not greatly improve with transformations. It is possible that more complex statistical techniques would have enabled us to find an appropriate transformation to help with equal variance.

Looking into the future and beyond the United States, further statistical analyses could also focus on adding a cross-cultural literacy comparison by exploring literacy rates around the world with adjusted variables. One study noted that a single statistic about literacy rates cannot tell the whole picture as well. For example, "although Haiti has a higher overall basic literacy rate than Afghanistan (45 vs. 32 percent), in fact, the two countries are almost equivalent in male literacy, but sharply different in female basic literacy rates" (Wallendorf). This journal also noted that literacy can directly influence statistics, as data collection methods like informed consent forms, questionnaires and surveys "may exceed the reading motivation of illiterates, further homogenizing the sample... these implicit exclusions mask a systematic bias" (Wallendorf). Thus, literacy is an important topic to research for not only social science majors, but for others as well, and there are many areas which warrant further investigation.

Appendix

References

Mehmood, Bilal and Syed Hassan Raza, Shabana Mureed. (2014). Health Expenditure, Literacy and Economic Growth: PMG Evidence from Asian Countries. Euro-Asian Journal of Economics and Finance, 2(4):408-417.

Wallendorf, Melanie. "Literally Literacy." Journal of Consumer Research, vol. 27, no. 4, 2001, pp. 505–511. JSTOR, www.jstor.org/stable/10.1086/319625. Accessed 10 Mar. 2021.

Weiss, Barry and Gregory Hart, Ronald E. Pust. (1991). The Relationship Between Literacy and Health. Journal of Health Care for the Poor and Underserved, 1(4): 351-363.

Codebook

An Excel spreadsheet with our codebook will be submitted with this report.