MSCA 31007 Statistical Analysis
Assignment 2

Introduction

This report summarizes the statistical modeling and analysis results associated with the Census dataset, which contains the United States census data to the tract-level from 2015 to 2019 and is native to *tidycensus* package in R. The purpose of this report is to document our study of two target variables: the proportion of bachelor's degree attainment (the first part of the report) and the proportion of households living below the poverty line (the second part of the report).

To understand the relationship between different predictors in our dataset to the proportion of bachelor's degree attainment, we focused on: i) identifying and explaining the differences in the performances between single and multiple prediction models; ii) comparing the differences in college attainment rate across geographic locations; iii) proposing changes to predictors to improve college attainment rate; and iv) identifying target households that would likely respond well to a policy used to improve college degree attainment rates by the government.

To understand the relationship between different predictors in our dataset to the proportion of households living below the poverty line, we i) chose 10 relevant variables from the Census data to be our model predictors; ii) conducted different model selection algorithms to identify the best model; and iii) performed transformations on chosen features to further improve our model performance.

Part I. Study of Proportion of College Attainment Rate

Model Selection for College Attainment Rate Prediction

Using the variables in data dictionary I, we create two models to predict college degree attainment (*propbac*). First, we generate a single predictor model (Model A) with only median household income as the predictor. Then, we build a model with all predictors (6 predictor variables) in the data dictionary I (Model B).

Table 1: *Data Dictionary I*

	Variable Name	Census ID	Variable Description
1	*medhhinc*	DP03_0062E	Estimation of median household income (dollars)
2	*totpop*	DP05_0001E	Estimation of total population
3	*medage*	DP05_0018E	Estimation of median age (years)
4	*propbac*	DP02_0065PE	Proportion of population 25 years and over with bachelor's degree attainment
5	*propcov*	DP03_0096PE	Proportion of Civilian noninstitutionalized population with health insurance coverage
6	*proppov*	DP03_0128PE	Proportion of people whose income in the past 12 months is below the poverty line
7	*proprent*	DP04_0047PE	Proportion of people who occupied housing units

Table 2: *Model comparison between Model A (with only medhhinc) and Model B (with all 6 predictors in data dictionary I)*

Model	Adjusted R Square	Residual Difference	Residual Sum of Squares	Degrees of Freedom	Sum of Squares	F	Significance
A	0.5347	1312	101496				
B	0.7136	1307	62241	5	39255	164.86	< 2.2e-16***

Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$

Adding more predictors to the linear model increased adjusted R-squared by 0.1789 compared to the single predictor model. About 71.36% of the variation in average college degree attainment is explained by the average total population, median age, median household income, proportion of insurance coverage, the proportion of families under the poverty line, and proportion of rent over income. Only 53.47% of the variation in mean college degree attainment is explained when only median household income is used. The change in adjusted R-squared is expected because college degree attainment cannot be explained by only one variable. Many reasons influence the college degree attainment rate. In addition, the difference in explanatory power between the two models is significant ($p < .05$) with a difference of 39,255 in the residual

sum of squares, which means that the multiple predictor model is indeed better at explaining college degree attainment. The multiple predictor model not only adds explanatory power but also improves the fit to OLS model assumptions. By plotting the empirical densities of the residuals (*Figure 1*), we can see that the multiple predictor model has more residuals close to 0 and minimizes the sum of squares of the residuals more. Thus, the multiple predictor model provides a better fit for college degree attainment than the single predictor model.

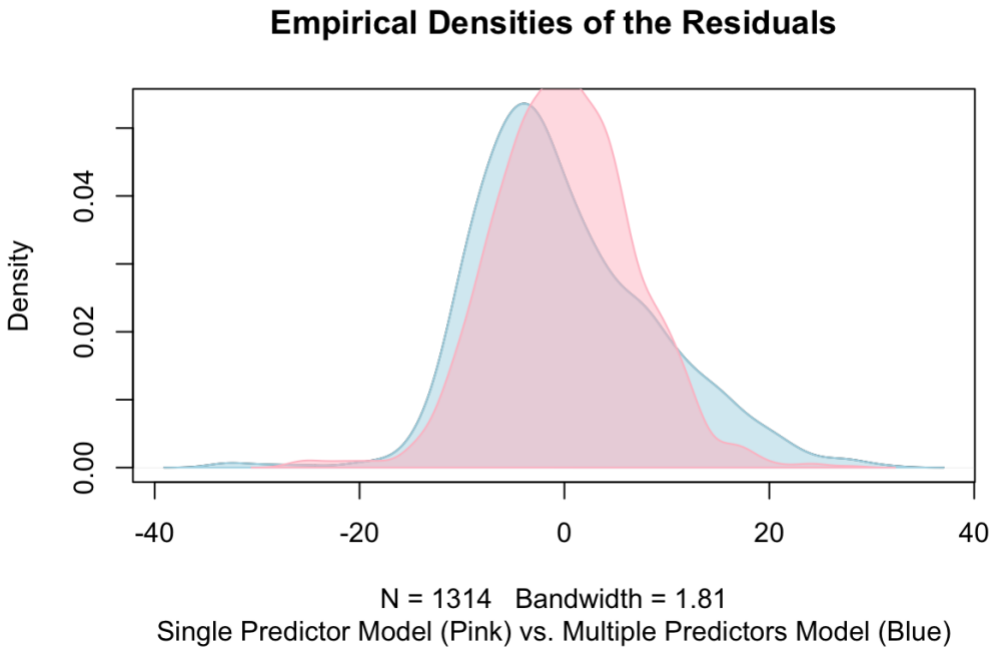


Figure 1: *Empirical densities of the residuals for Model A (in pink) and Model B (in blue)*

Moreover, by performing ANOVA, we note that the results show a degree of freedom of 5 (which corresponds to the 5 additional parameters we have in the complex model) with a very small p-value ($p < .05$), indicating that the multiple predictor model is a significantly better fit than the single predictor model. From the outcomes of adjusted R-squared, empirical densities of the residuals, and ANOVA, it is clear that the multiple predictor model is a better choice than the single variable predictor. We should use the complex model with six explanatory variables to make a recommendation in making future policies.

Policy Proposal for Increasing College Attainment

As total population and median age have no significant power in explaining college degree attainment (ps > 0.05 and rs < 0.1), we mainly focus on four explanatory variables (Model C): median income (*medhhinc*), proportion of insurance coverage (*propcov*), proportion of below poverty line (*proppov*), and proportion of people who are paying rent (*proprent*).

Among these four predictor variables, proporent is the hardest to adjust without a policy lever that decision-makers could use to increase college degree attainment. It is impossible for policymakers to give everyone property and difficult for them to provide everyone with a place to live without charging. *Medhhinc*, *propcov*, and *proppov* are comparatively easier to have a corresponding policy lever.

Table 3: *Four Predictor Model (Model C)*

Predictors	Coefficients	Standard Error	Significance
(intercept)	-50.9575	3.2898	<2e-16 ***
medhhinc	0.0002	< 0.0001	<2e-16 ***
propcov	0.5628	0.0369	<2e-16 ***
proppov	-0.3771	0.0241	<2e-16 ***
proprent	0.2819	0.0110	<2e-16 ***

Significance codes: *** p < .001, ** p < .01, * p < .05

Adjusted R-Squared: 0.7135

From Model C, we know that the college degree attainment rate (*propbac*) increases by 1 unit for every 0.5628 increase in the proportion of insurance coverage (*propcov*) if all other variables are held constant. Since the variable *propcov* affects *propbac* by the largest among all four variables and the government can launch feasible proposals about it, we start with insurance coverage when trying to increase the college degree attainment rate by 5 percent. If we expand the mean of *propcov* to 100, the response variable *propbac* will grow about 5 percent if all other explanatory variables stay the same.

In reality, however, it is impossible for us to raise all proportion of insurance coverage to 100, nor can all other three explanatory variables stay constant while insurance coverage changes. This is because insurance coverage is correlated with other predictors. *Propcov* has a correlation of 0.5147 with *medhhinc*, -0.3309 with *proppov*, and -0.3396 with *proprent*. Due to this reason, we should adjust all predictors although it is difficult for the government to control the proportion of people who rent a place in real life. Moreover, since everything has a non-weak correlation with the proportion of insurance coverage, simply raising the proportion of insurance coverage will not add 5 more percent to college degree attainment. *Proppov* also should not exceed 100 because having a percentage more than 100 does not make sense in the situation. Thus, we would like to adjust all variables to attain a higher *propbac*.

Since proportion of insurance coverage (*propcov*) has a positive relationship with median income (*medhhinc*) and negative correlations with proportion of poverty (*proppov*) and proportion of people who rent a place (*proprent*), we increase *propcov* and *medhhinc* and decrease *proppov* and *proprent*. Interestingly, the proportion of people who rent a place negatively correlates with college degree attainment but maintains a positive relationship in the linear model. Although decreasing *proprent* will negatively affect the increase in *propbac* in the model, we still choose to reduce *proprent* due to its correlation with other explanatory variables and the response variable. If the proposal can raise *medhhinc* by 2,000 dollars and *propcov* by 5.1 percent and reduce *proppov* by 6.9 percent and *proprent* by 0.5, the college degree attainment will increase by 5.09% on average.

National College Attainment Rates across Geographic Locations

A. Cook County vs. National

In this part, we want to dive deep into the average college attainment rate *(propbac*) of Cook County compared to the national average *propbac* excluding Cook County, as well as the *propbac* of one specific tract in Cook County compared to *propbac* of other tracts in Cook County. The dataset of the total population (*totpop*) and *propbac* for every tract in the United States is retrieved with 73,056 rows. After cleaning and filtering tracts with non-missing population, non-missing college degree data, and population of at least 100, we have two datasets, including 70,976 values without Cook County and 1,315 values for Cook County. The national equal-weight average for tract-level college degree attainment (excluding Cook County, IL) is 18.7875, and the average weighted by population is 19.4978.

We conducted a hypothesis test to see whether the tracts from Cook County could share the same equal-weighted average college degree attainment as the national average of the remaining tracts in the United States. The null hypothesis is that the equal-weight average college degree attainment in Cook County ($\mu 1$) is equal to the national average college degree attainment excluding Cook County ($\mu 2$). The alternative hypothesis is that the equal-weight average college degree attainment in Cook County is not equal to the national average college degree attainment.

$$H0: \mu 1 = \mu 2 \qquad Ha: \mu 1 \neq \mu 2$$

The hypothesis test is conducted using one sample t-test by treating the national college attainment rates as a constant. The result shows that the p-value is less than 0.05, so we rejected the null hypothesis. Cook County (M = 21.8634, SD = 12.8899) did not share the same equal-

weight average for tract-level college degree attainment as the national average, $t(1314) = 8.6533$, $p < .05$).

Table 4: *Census data for tract that contains the Gleacher Center and NBC Tower*

geoid	name	totpop	medage	medhhinc	propbac	propcov	proppov	proprent
17031081403	Census Tract 814.03, Cook County, Illinois	8450	34.2	115642	38.8	99.1	11.9	68.9

B. Tract with Gleacher Center/NBC Tower vs. Other Tracts in Cook County

We are interested in whether the tract that contains the Gleacher Center and NBC Tower has the level of college attainment that can be predicted by the same pattern as Cook County. We located it in the dataset by its *geoid* and found that this tract has a population of 8,450 and a college degree attainment rate of 38.8%. Our point estimate for the predicted college degree attainment rate by the regression model in this tract is 45.843 percent, with a 90% confidence interval between 44.5511 and 47.1349 percent. However, the actual college degree attainment rate for this tract is 38.8%, which is not contained in this 90% confidence interval. When we refit the linear regression model weighted by population, the point estimate for the predicted college degree attainment slightly decreased to 45.3691 percent, with a 90% confidence interval between 44.3095 and 46.4288 percent, which is narrower than the 90% confidence interval of the original linear regression model.

Table 5: *Model weighted by the total number of population (Model D)*

Predictors	Coefficients	Standard Error	Significance
(intercept)	-55.7253	3.4522	<2e-16 ***
medhhinc	0.0002	< 0.0001	<2e-16 ***
totpop	< 0.0001	< 0.0001	0.7470
medage	0.0705	0.0373	0.0589
propcov	0.5782	0.0381	<2e-16 ***
proppov	-0.3616	0.0264	<2e-16 ***
proprent	0.2955	0.0118	<2e-16 ***

Significance Codes: *** p < .001, ** p < .01, * p < .05

Adjusted R-Squared: 0.7126

In order to validate the linear regression model used to make predictions of college attainment rates, so it captured relevant features of the data, we simulated the model coefficients of the regression model that are unweighted by population. To achieve the Central Limit Theorem, we simulated 10,000 sets of possible model coefficients based on a multivariable normal distribution assumption with a mean equal to the estimated model coefficients and a standard deviation equal to the model coefficients' standard error. Then we used the simulated model coefficients to predict the college attainment rate for the tract that contains Gleacher Center and NBC Tower by conducting a matrix multiplication. The mean of the 10,000 predictions from simulated coefficients is 45.8486, with a standard deviation of 0.7918. The 90% confidence interval of the predictions' mean is [44.5183, 47.1789]. While the actual college attainment rate for this tract is 38.8%, it is not included in the 90% confidence interval. Comparing this simulated confidence interval with the CI predicted by the linear regression model, we see that the simulated confidence interval is slightly wider than the confidence level predicted by the linear regression model. Therefore, we conclude that the unweighted regression model captures less uncertainty in the data, which can help to explain why it cannot predict the actual college attainment rates for the tract that contains Gleacher Center and NBC Tower.

Memo to County Officers Regarding the New Public School Program Proposal

Based on our model and analysis, we are confident that the new public school program will raise the college attainment level. By looking at the lowest quartile residuals (the residuals that have largest negative values; *Table 6*) from Model B with 6 predictors, we found out these tracts have lower average in median age (medage), median household income (medhhinc), college attainment rate (propbac), proportion of below poverty line (proppov) while higher in average health insurance coverage rate (propcov) and proportion of people who are paying rent (proprent) compared to all other tracts in Cook County. This indicates that people in this quartile have the potential to go to college, according to how we have studied and understood the characteristics of a typical college-educated person, but they decide not to. We believe that targeting this group of people with an effective public school program will encourage more college attendance, because the program will only need to target the willingness of these people, instead of providing other incentives to resolve other concerns that people in other quartiles might have.

Table 6: *Average value for each variable for Cook County and the quartile with lowest residuals*

	totpop	**medage**	**medhhinc**	**propbac**	**propcov**	**proppov**	**proprent**
Cook County	3955	37.49	68114	21.86	91.26	16.24	45.05
Lowest Quantile	3829	36.74	66444	13.84	91.34	16.0	47.64

Part II. Study of Proportion of Households Living under Poverty Line

Choosing Variables to be Model Predictors

At this point of the report, we will temporarily move on from using *propbac* as our response variable and shift our focus to predicting the proportion of households living below the poverty line, tract-wise. In order to model the proportion of households living below the poverty line, we have selected the following ten variables that we think will have a linear relationship with our response variable. All chosen variables need to satisfy the following criteria: i) each variable needs to have at least 90% non-missing value, and ii) each variable needs not directly to measure income. The list of chosen variables is as follows, to the tract-level:

Table 7: *Data Dictionary II*

	Variable Name	**Census ID**	**Variable Description**
1	*medage*	DP05_0018E	Median age
2	*propbac*	DP02_0065PE	Proportion of bachelor's degree attainment
3	*propcov*	DP03_0096PE	Proportion of health insurance coverage
4	*proprent*	DP04_0047PE	Proportion of people who rent a house or apartment
5	*propcomp*	DP02_0151PE	Proportion of households with a computer
6	*medhouseval*	DP04_0089E	Estimated median value of owner-occupied houses
7	*propsch*	DP02_0053PE	Proportion of population 3 years or older enrolled in school
8	*propnotus*	DP02_0096PE	Proportion of foreigners (not US citizen)
9	*propempl*	DP03_0004PE	Proportion of employed population in civilian labor force

10	*propveh*	DP04_0057PE	Proportion of households that have vehicles

Conducting Model Selections

After choosing these ten variables, we conducted descriptive statistics and plotted each variable's histograms to understand the ranges of our variables, as well as to do a sense-check to look for potential outliers or data biases. We then proceeded to do the model selection with minimized Aikake's Information Criterion (AIC), a goodness-of-fit score, as our target metric. Since the number of predictors is reasonable (10), we can conduct an exhaustive search for the best possible model. However, in order to be more cautious, we will also use backward and forward selection algorithms to find good models, and compare the results of the different model selection algorithms.

1. **Exhaustive Search**

The exhaustive search examines every possible combination of all attributes and gives their adjusted R-squared values. A model with a high adjusted R-squared value indicates that it is better at predicting response values with less error. The highest R-squared chosen by the exhaustive search is the model with 9 variables and selects the best model with an adjusted R-squared of 0.6905 (Model E).

2. **Forward Selection**

The forward selection algorithm compares the AIC of each model by adding one variable at a time and returns the best model as the same as the exhaustive search did (please refer to Model E and *Table 8 & 9* below).

3. **Backward Selection**

The backward selection algorithm drops variables from the full model and evaluates resulting model performance by each drop. Once again, it chooses the best model in the same way as the exhaustive search and forward selection did (please refer to Model E and *Table 8 & 9* below).

Table 8: *Results from model selection methods*

Method	AIC	RSS
Forward Selection	250622	2541566
Backward Selection	250622	2541566

All three methods select the same model with 9 explanatory variables with an AIC of 250622, RSS of 250,704, and adjusted R-squared of 0.6905. The best model was found to include the following variables:

Table 9: *Best model selection by three methods (Model E)*

Predictors	Coefficients	Standard Error	Significance
(intercept)	64.0729	0.5058	<2e-16 ***
proprent	0.2026	0.0014	<2e-16 ***
propempl	-0.4628	0.0030	<2e-16 ***
medage	-0.3769	0.0052	<2e-16 ***
propbac	-0.1034	0.0033	<2e-16 ***
medhouseval	0.0000	0.0000	<2e-16 ***
propcomp	-0.0013	0.0000	<2e-16 ***
propcov	-0.1107	0.0041	<2e-16 ***
propsch	0.0005	0.0001	<2e-16 ***
propnotus	0.0096	0.0010	<2e-16 ***

Significance codes: *** p < .001, ** p < .01, * p < .05

Adjusted R-Squared: 0.6905

The only variable that the best model eliminates from the full model with 10 variables is the proportion of households that have vehicles (*propveh)*.

Even though our three methods of choosing the best/good model all point us to the same answer, the best model in this case still failed all our Independent and Identically Distributed Random Variable (IID) tests, which consist of the Durbin-Watson test, Breusch-Pagan test, and Kolmogorov-Smirnov test. They all have p-values close to 0 (*ps* < .05*)*, which indicates that the

model's residuals i) have autocorrelation, ii) lack homoscedasticity, and iii) are not normally distributed.

In order to evaluate the performance of our model, we will use Root Mean Square Error (RMSE), R-squared, and Adjusted R-squared as our target metrics. RMSE is used to measure the differences between values predicted by the model the values observed (the lower, the better), R-squared represents the proportion of the variance for a dependent variable that's explained by an independent variable (the higher, the better), while Adjusted R-squared shows whether adding additional predictors improves a regression model or not (the higher, the better). The RMSE of the model is 6.0396, with an R-squared and adjusted R-squared of 0.6905. We can conclude that the model does a good job explaining the response's variance (69.05%), but we need to be conscious of the effects that a lack of IID does on our results, such as the understated standard error of the coefficients.

Experiments with Transformations of Predictors

Although the explanatory power of our chosen model in part A was good, we will try to improve our model by conducting feature engineering in this section. First, we will determine which variables to transform based on their distribution.

We hypothesize that we would need to do transformation on *proppov, medhouseval, propcov, propcomp,* and *propsch* to transform these skewed variables (*Figure 2*) so that they can approximately conform to normality. Note that having a normal distribution is not a necessary requirement for predictor or response variables in modeling; however, it is a good starting point for us to experiment on. In addition, we created a product function of two variables: *propempl* and *propbac*, which is essentially a feature that describes the proportion of bachelor's degree attainment that is employed, tract-wise. We think this new feature might be useful to predict the proportion of households living below the poverty line because of the hypothesis that people with a college degree will tend to be able to find better-paid jobs, ones that can help them live above the poverty line. We suppose that mere employment rate or bachelor's degree attainment rate might not be as strong indicators of what level of compensation one can get as the new feature.

We then tested the effect of different transformed features on our model to identify the best combinations of predictors. For the options that contain transformation to the predictor *proppov,* we un-transformed the predicted values before calculating RMSE, so that we can

compare their RMSE to our baseline model. More specifically, we tested the following combinations:

Option 1. Log *medhouseval*, keeping all other variables the same

Result: RMSE of 5.8522

Option 2. Square-root *proppov* and log *propcomp*, keeping all other variables the same

Result: RMSE of 5.8286

Option 3. Square-root *proppov*, log *propcomp*, and log *medhouseval*, keeping all other variables the same

Result: RMSE of 5.8378

Option 4. Square-root *proppov* and log *medhouseval*, keeping all other variables the same

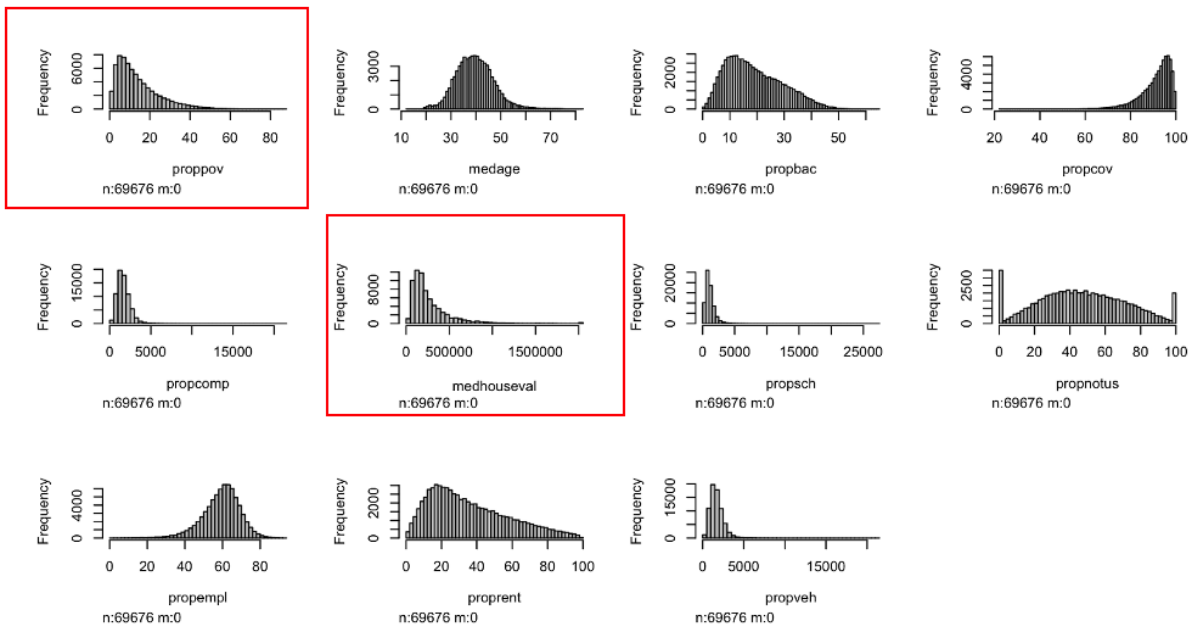
Result: RMSE of 5.6312

Option 5. Square-root *proppov*, log *medhouseval*, and product function between *propempl* and *propbac*, keeping all other variables the same

Result: RMSE of 5.6360

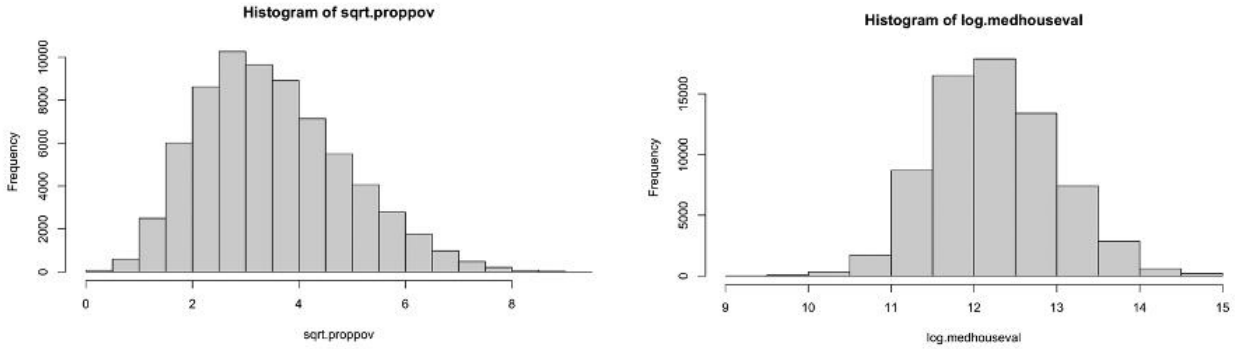
Even though there are multiple non-normal variables, we found that normalizing just *proppov* and *medhouseval* led to the best result in our evaluation metrics (adjusted R-squared and AIC).

***Figure 2**: Distribution histograms of all variables in the best model*



The transformed variables look much more normally distributed, as visualized in *Figure 3*.

Figure 3*: Histograms for transformed proppov and medhouseval*



Recall that our starting model (best fit model before feature transformation) has an RMSE of 6.0396. Out of all our combinations above, even the least good model (**Option 1**) already has better RMSE than our baseline model (6.0396 vs. 5.8522). We can conclude that the non-normality of our features really affected their usefulness as predictors in our regression. However, we also found that we should not normalize every non-normal feature (evidenced by the difference in performances between **Option 3** and **Option 4**) because the distribution of the residuals needs to be normal, not the data distribution. Normalizing variables will make their residuals more normal, but an unnecessary data transformation will reduce the information that such data has and thus weaken our model. Therefore, we chose to normalize only *proppov* and *medhouseval*. We can also see that contradicting our hypothesis, the new feature (*propempl*propbac)* does not have a positive explanatory effect in our model, as **Option 5** has worse RMSE than **Option 4**. In fact, **Option 4** has the best RMSE (the lowest RMSE) out of all our models using transformed features and the baseline model, an RMSE of 5.6312.

Unfortunately, it is worth noting that the identified best fit model (after transformation) still does not pass any of our IID tests (Durbin-Watson test, Breusch-Pagan test, and Kolmogorov-Smirnov test), having all p-values of zero (*ps* < .05*)*. In reality, it is challenging to build a predictive model that has good explanatory power and satisfies IID tests. Therefore, failing to pass those tests does not mean that our model is ineffectual, but rather that we need to be conscious when making conclusions about our p-values and confidence intervals.

Conclusion

To summarize some of the key conclusions from the report, our first aim of the analysis is to select a better model to predict the college degree attainment rate. Using 6 predictors (total population, median age, median income, proportion of health insurance coverage, proportion of people below the poverty line, and proportion of people renting), the model makes good predictions of college degree attainment rates. Since the health insurance coverage ratio has the highest coefficient, policymakers can focus on creating policies that increase this ratio to increase college degree attainment rates. However, they must consider that all variables are interrelated, so adjusting one variable will also affect the others. In addition, college degree attainment is clearly segmented by location, as evidenced by the average college attainment rates difference between the tract containing Gleacher Center and NBC Tower and other tracts within Cook County. After looking at the lowest quartile residuals in the tract-level data of Cook County and identifying the characteristics associated with households in this quartile, we believe that a new public school program targeting households in this quartile can help increase college attendance. Finally, to see how other ACS variables can help explain the tract-level proportion of households living below the poverty line, we conducted multiple model selection processes to identify the best model and experimented with different ways of transformations to improve our model. Although the transformed model failed all Independent and Identically Distributed Random Variable tests, it still possesses high explanatory power. Overall, the report illustrates the complex relationships among variables in the Census dataset. For this reason, we must consider the interrelations among various factors before taking any action and/or developing any policy that draws from the insights of our data.