Lauryn Davis
Bradford Dykes
STA 631-01
6 August 2024

Statistical Modeling and Regression Final Portfolio

Overview:

STA 631-01 allows students the opportunity to complete a data modeling project from start to finish. Students are given the independence to choose their domain and apply meaningful statistical models to their chosen topic. In this case, the topic chosen for this portfolio is *historical redlining implications on present economics*. Within past semesters, I was able to work with Dr. Bradford Dykes and Dr. Tamera Shriener to develop a website that highlights geospatial analysis of historical redlining homeowners loan corporation (HOLC) grades. These multi-polygon shape files were collected from the University of Richmond and plotted within an interactive map with current economic information.

A way to extend the current scope of the website is to include statistical modeling of current median home values and test *if* there is any implication of historical policy such as redlining HOLC grades. An outline of the data project will be covered. Additionally, each of the course objectives will be acknowledged and directed to specific topics within the data project to show evidence that it has been completed.

Data Project:

ETL (Extract, Transform, Load) Process:

Data Availability:

As discussed, the University of Richmond has data on over **215 cities** throughout the United States (*Mapping Inequality*). Each city/state grouping is further broken apart to analyze the HOLC grades given to various communities. Furthermore, the R programming census data platform called "TidyCensus" has data on **90%** of the corresponding cities within the year of 2020. The year of 2020 was chosen in order to pull the most cities corresponding to HOLC grades. This census data can be viewed at a more granular **"block level"** group which is incorporated within this project.

Consolidation of information:

In order to extract this information, there are a few different areas that need to be pulled from. The redlining information can be found within this JSON link (*Mapping Inequality*). Additionally the TidyCensus data can be efficiently pulled using this starter guide document

(*Basic Usage of TidyCensus*). Both pieces of information are in an *sf* format, meaning that they are simple structured data features with multi-polygon shape files corresponding to areas of interest. In order to join the two files together on location, I first verified that they have the same R coordinate system. Then, I joined the two files with an R function called *st_join (Spatial Join)* with a parameter specification of *st_is_within_distance (Topological relations)*.

```
#Loading shape files from richmond:
all_shape<- read_sf('https://dsl.richmond.edu/panorama/redlining/static/mappinginequality.json')

#Transforming shape files to have appropriate coordinates system:
all_shape <- st_transform(all_shape, crs = 4326)

#Removing N/A'/ grades that were not equal to A,B,C or D:
no_na_richmond<- na.omit(all_shape)
updated<- no_na_richmond %>%
  mutate(grade = if_else(grade == 'A ', 'A', grade))%>%
  mutate(grade = if_else(grade == 'C ', 'C', grade))
final_richmond <- subset(updated, grade != "F" & grade != 'E' & grade!= '' & grade!= ' ')


#Making the same CRS Code. Shape_2021_median is the shape files for tidy census. They actually come from 2020, so ignore
#that it says 2021
final_richmond_crs <- st_transform(final_richmond, crs = st_crs(shape_2021_median))

#Validating richmond set:
final_richmond_valid <- st_make_valid(final_richmond_crs)

#Filtering out an area id that was causing an error
final_richmond_valid <- final_richmond_valid %>%
  filter(area_id != "4118")


#Merging together based on distance. It default checks the "geometry" variable in terms of closeness.
#df is the tidy census shape files
merged_data <- st_join(df, final_richmond_valid,
                       join = st_is_within_distance,
                       dist = .0001,
                       left = FALSE)
```

Aggregate Idea:

With the idea of regression in mind, which will be further explained in the next section, I performed transformations on various historical variables of concern to align with the model idea. Two of the predictors that will be tested within the model include *category type* and *HOLC Grade*. Category type is a categorical variable with an indicator assigned to a location describing the desirability of the location within the 1930's. Furthermore, the HOLC grade is the grade in which the specific area of the city/state was assigned by the Home Owner's Loan Corporation. This grade impacted the ability of obtaining a mortgage within this time. The grade took various factors into account such as race, poverty status, population, etc. The grading system was often discriminatory to various communities.

In order to align with the idea of predicting current median home values at the block group level (the deepest granularity I can achieve from TidyCensus), I must aggregate category and grade to be at the block group level. Mode can be calculated using the *dplyr* R package. Documentation from *SQLPad* helped in understanding some of the steps that are needed to adequately calculate the mode (*Mode Calculation in R*). This is achieved by taking the mode of

the variable that falls within a specific block group. If there were no duplicate values of either variable within the group, for example if the block group had grades A/B/C one time each, the indicator assigned is: *Tied Groups.*

```
df_agg_1 <- selected_cols %>%
  group_by(across(c(NAME, city, state, all_of(variables_to_include)))) %>%
  summarize(
    grades_list = list(grade),
    mode_grade = {
      freq_table <- table(grade)
      max_freq <- max(freq_table)
      #Adding in tied grades because if the grades are only one of each in a group, we just want it to say Tie:
      modes <- names(freq_table[freq_table == max_freq])
      if (length(modes) > 1) {
        "Tied Grades"
      } else {
        modes
      }
    },
    mode_category = {
      freq_table <- table(category)
      max_freq <- max(freq_table)
      modes <- names(freq_table[freq_table == max_freq])
      if (length(modes) > 1) {
        "Tied Modes"
      } else {
        modes
      }
    }
  )
```

Standardization:

Standardization is a simple process that can be done in R. The primary purpose of standardizing continuous variables is to have them all on the same *scale.* This can allow for easier interpretation of what predictors are influencing your model the most. There are two parameters within this function, *center* and *scale* (each set to the default of true). Centering subtracts the column mean from the actual value. Scaling divides the centered values by their columns standard deviation. In general, the scaled value calculation is: $x_i' = \frac{x_i - \bar{x}}{\sigma}$ (*Scale: Scaling and Centering of Matrix-like Objects)*.

```
df_scaled <- df_filtered %>%
  mutate(across(all_of(variables_to_include), scale))
```

Within this project, all continuous TidyCensus variables within the initial model were standardized to understand what was influencing current median home values the most. I decided to drop ~300 observations that had no TidyCensus data to help the models be less skewed. It could become skewed if I kept them in because they were being labeled “0” for all economic variables. (** Note: due to time constraints, I only standardized my initial model. I did not standardize the final models and feel this would be beneficial to integrate at a later time).

Categorical Predictors:

Within multiple linear regression, it's important to factor categorical variables for analysis. This is particularly useful for categorical variables that have multiple values such as the HOLC grades or category descriptions for various communities. As discussed, I took the mode of each categorical variable within the specified block group and applied a *factor definition.* Factoring can be done with base R and utilizes "dummy coding" methodology. According to documentation, "It creates dichotomous variables where each level of the categorical variable is contrasted to a specified reference level" (*Coding for Categorical Variables*). In this case, each HOLC grade will be contrasted with HOLC grade A as the reference level. I also consolidate category information to be labeled "desirable", "declining", or "other". Finally, I factor the state variable for use in a regression model, setting the reference to Michigan to see how much median home values differ compared to this state. I felt this could be interesting based on my current location and knowledge of Michigan. Interpretation of significant factored categorical variables within the regression model will be stated below along with Michigan comparisons.

```
df_filtered$mode_grade_idx <- as.factor(df_filtered$mode_grade)
df_filtered$mode_category_idx <- as.factor(df_filtered$broad_category)
df_filtered$city <- as.factor(df_filtered$city)
df_filtered$state <- as.factor(df_filtered$state)
```

Multiple Linear Regression Modeling:

Objective goal: I wish to predict 2020 median home values for ~35,000 aggregated block group/census tract/city/states within the United States using historical and current economic and demographic information. I aim for 80% accuracy as defined by MAPE (discussed below).

Test/Train Split:

Within machine learning, there are many steps that must be taken in order for a statistical model to be deployed and **maintained.** In many cases, the data that is worked with in the real world is large (> 100 million observations). In order to develop accurate models, one must train their model and test it frequently. Although the data that is worked with within this project is somewhat small (< 60,000 observations), I can still implement common practice modeling steps that would be crucial in the real world for many situations.

In this case, a simple 70/30 split of the dataset was taken using documentation from GeeksForGeeks. It's also important to set a random seed so you can focus on one baseline training and testing set at a time (*GeeksForGeeks*). If you did not specify a random seed, the datasets would change each time you run the code and would cause additional accuracy testing. (For the sake of this project, I simply set the random seed). Also, I stratify the variable "state" to make sure each state is found at least once within the training and testing set. The sample size for the training set is 24,865 and the testing set is 10,651 observations.

```r
set.seed(123)
library(caret)

#Stratified variable for "state" because we want to make sure the same state appears within the test and train sets:
stratify_variable <- df_filtered$state

train_index <- createDataPartition(stratify_variable, p = 0.7, list = FALSE)

train_data <- df_filtered[train_index, ]
test_data <- df_filtered[-train_index, ]
```

Initial Model:

A baseline model that can be tested is using only numerical predictors. This will probably cause over-fitting, but it's a good way to get an initial gauge of what predictors are being labeled as significant. A way to consolidate predictors and measure accuracy is to use cross validation. That will be discussed in the next section.

Continuous predictors only:

Significant variables:

	Estimate	Std. Error	t value	Pr(>|t|)
Population	7.670847e-02	1.390114e-02	5.518141e+00	3.460222e-08
Median_Year_Structure	-9.743293e-02	5.162305e-03	-1.887392e+01	6.668069e-79
Total_Bedrooms	-8.106732e-02	1.767942e-02	-4.585407e+00	4.553114e-06
Median_Number_Rooms	-5.486105e-02	5.477293e-03	-1.001609e+01	1.435825e-23
Median_Age	8.411099e-02	6.262917e-03	1.343000e+01	5.607564e-41
Median_HouseHold_Income	2.767230e-01	7.681751e-03	3.602342e+01	4.777108e-277
Average_Household_Size	2.034979e-02	5.283694e-03	3.851432e+00	1.177235e-04
Total_200000_or_more	1.813337e-01	1.265333e-02	1.433091e+01	2.148076e-46
Gross_Rent_Percent_of_Income	1.015847e-01	1.001021e-02	1.014810e+01	3.765842e-24
Travel_Time_to_Work	-1.387880e-01	1.163022e-02	-1.193339e+01	9.742624e-33
Total_Food_Stamp_Recievers	-8.515570e-02	7.630829e-03	-1.115943e+01	7.542846e-29
Total_With_Internet_Sub	5.175845e-02	2.198526e-02	2.354234e+00	1.856866e-02
Total_Mobile_Homes	-1.765399e-02	4.720351e-03	-3.739975e+00	1.844518e-04
Total_7_or_More_Person_Households	6.137104e-02	5.424297e-03	1.131410e+01	1.320360e-29
Agg_Earnings_Past_12_Months	2.220818e-01	1.585729e-02	1.400502e+01	2.142860e-44
Total_Black_Alone	-1.746599e-01	9.203427e-03	-1.897770e+01	9.566219e-80
Total_White_Alone	-2.702643e-01	1.225612e-02	-2.205138e+01	9.796631e-107
Total_Bachelors_Degree	5.521234e-02	9.135931e-03	6.043428e+00	1.530211e-09
Agg_Real_Estate_Taxes	1.720748e-01	1.203735e-02	1.429507e+01	3.581812e-46

Although these are all considered significant based on their p-values, the goal would be to have as few predictors as possible to allow for **interpretability** and to **prevent overfitting.** Overfitting has to do with training on too many predictors which causes for generalization issues on the testing set.

Metrics on testing set:

Performance Metrics

Metric	Value
Root Mean Squared Error (RMSE)	305181.237
Mean Absolute Error (MAE)	201066.170
R-squared (R^2)	0.273

Model Selection/Cross Validation:

Now that I have an initial gauge of what predictors are deemed to be significant, I created three different models and implemented K-fold cross validation to test which was allowing for a higher accuracy.

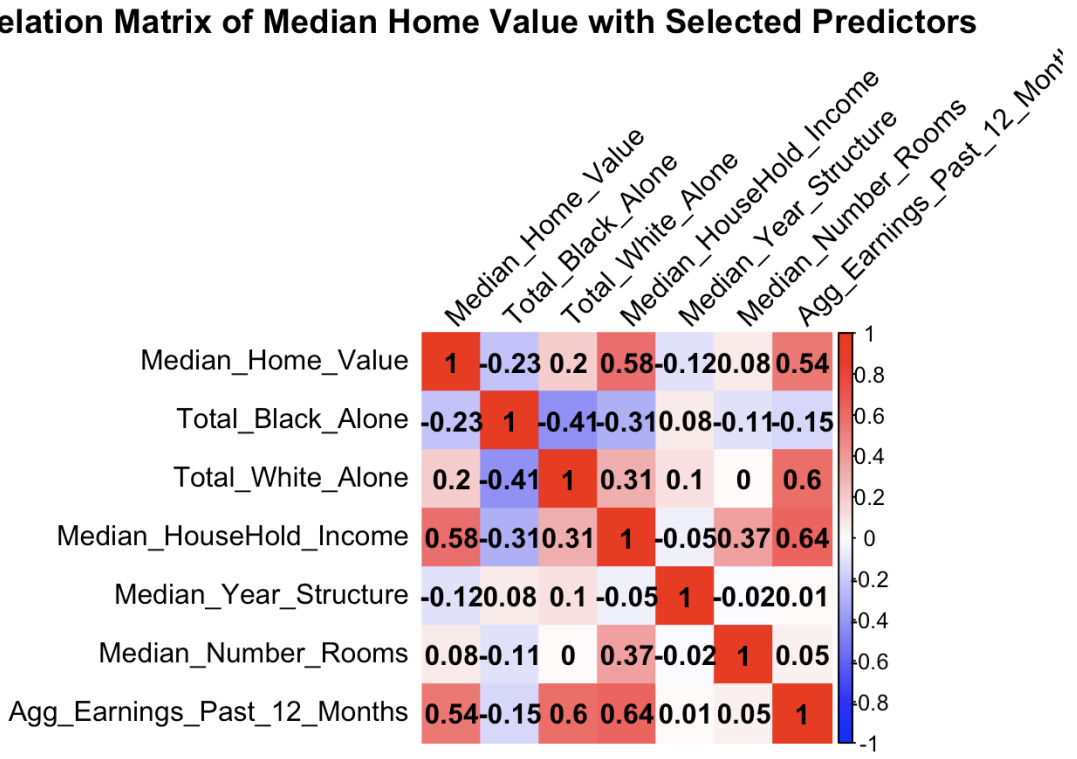
Based on documentation from GeeksforGeeks, K-fold Cross Validation splits the data into various subsets dependent on the parameter that is set for K. For example, if there were 5 folds, the data would be split into 5 subsets, trained on k-1 folds, and tested on the hold-out validation set. This allows you to see if your model **generalizes** well to unseen data that was not included in your training set (*GeeksforGeeks*).

```
cv_control <- trainControl(method = "cv", number = 10,
                           savePredictions = "all",
                           returnResamp = "all")

# Train the model with k-fold cross-validation
cv_model <- train(formula, data = train_data, method = "lm", trControl = cv_control)
```

In the above screenshot, I define the number of folds to be 10 and the “formula” is the fitted model of interest.

The first model included information on the number of people of a certain race in a block group for Caucasian and African Americans, the median year the structure was built, the median number of bedrooms, median household income, and the aggregate block group earnings in the past 12 months. I felt these predictors may provide a better understanding of median home value predictions based on initial model findings and housing marketing research. VIF validation was also conducted. I also provided a correlation matrix to understand the predictors relationships with median home values:



After completing 10 fold cross validation and testing on the hold-out set, the overall average performance metrics across the validation sets included:

Performance Metrics with MAPE Threshold

Metric	Value
Root Mean Squared Error (RMSE)	270434.110
Mean Absolute Error (MAE)	176091.448
R-squared (R^2)	0.429
Proportion of Predictions with MAPE ≤ 0.20 (%)	20.130

The second model included the mode block group HOLC grades and state in addition to those mentioned in the first model:

Performance Metrics for Model with State and Mode Grade

Metric	Value
Root Mean Squared Error (RMSE)	246704.948
Mean Absolute Error (MAE)	157322.106
R-squared (R^2)	0.525
Proportion of Predictions with MAPE <= 0.2	0.264

Finally, the third model included all predictors from model 1, state, and an interaction term between the mode block group HOLC grade and mode block group desirability category:

Performance Metrics

Metric	Value
Root Mean Squared Error (RMSE)	246705.624
Mean Absolute Error (MAE)	157321.318
R-squared (R^2)	0.525
Proportion of Predictions with MAPE <= 0.2	0.264

Each of the metrics utilized in this project provide important information about how your model is performing. For example, root mean squared error tells us how far off we were between predicted and actual values. Additionally, MAE tells us how far off we were, on average, when comparing the absolute value between predicted and actual median home values. R^2 is the percentage variation that the predictors explain for median home values.

Finally, MAPE is the moving average percentage error. I wished for this value to be low to allow for better accuracy in predictions. A data scientist I work with favors this measurement because it's interpretable and generalizes a bit better. My objective goal is to achieve 80% accuracy through MAPE. With that being said, we'd want absolute percentage errors to be less than or equal to 20%. (Mape equation below):

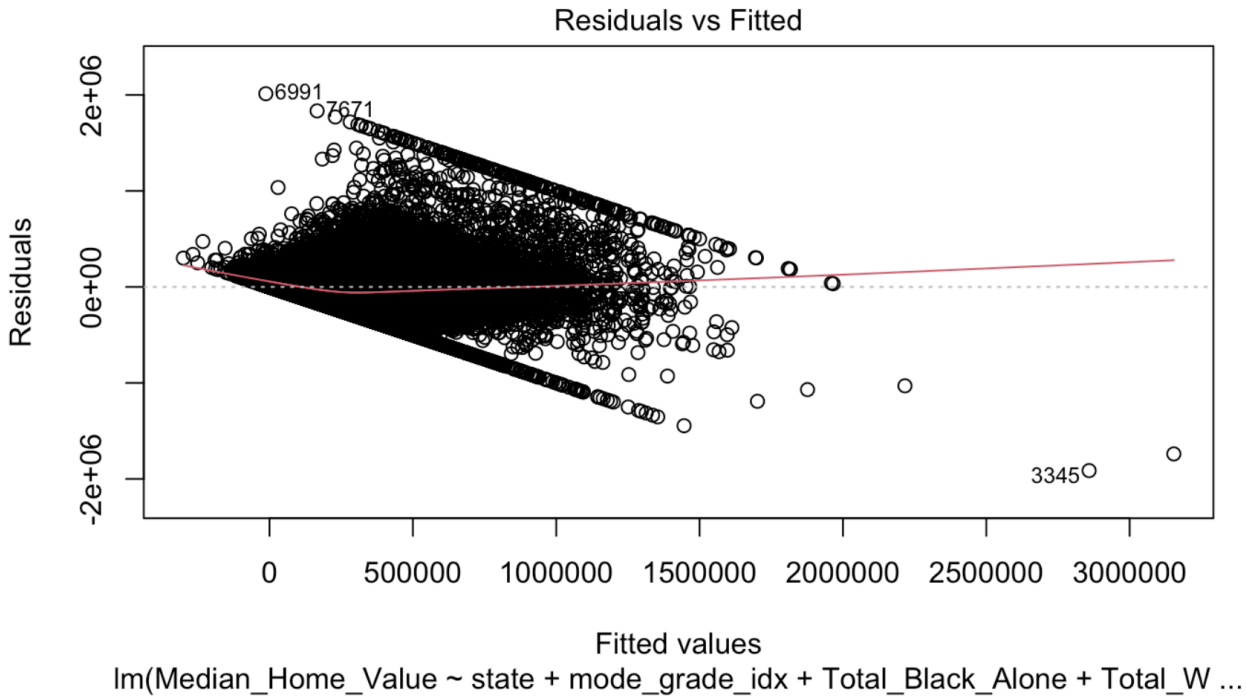
$$\mathrm{M} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

Based on these results, I would probably move forward with the model that does not include the interaction term. I feel that it did not contribute to the model in a significant manner and didn't improve performance metrics. From model 2, I can interpret RMSE by saying we are, on average, off by about $246,000 dollars when predicting median home values. Additionally, the R^2 allows us to say that, on average, 52.5% of the variance in median home values is explained by the predictors. Finally, on average, 26.4% of the observations within the model have 80% MAPE accuracy.

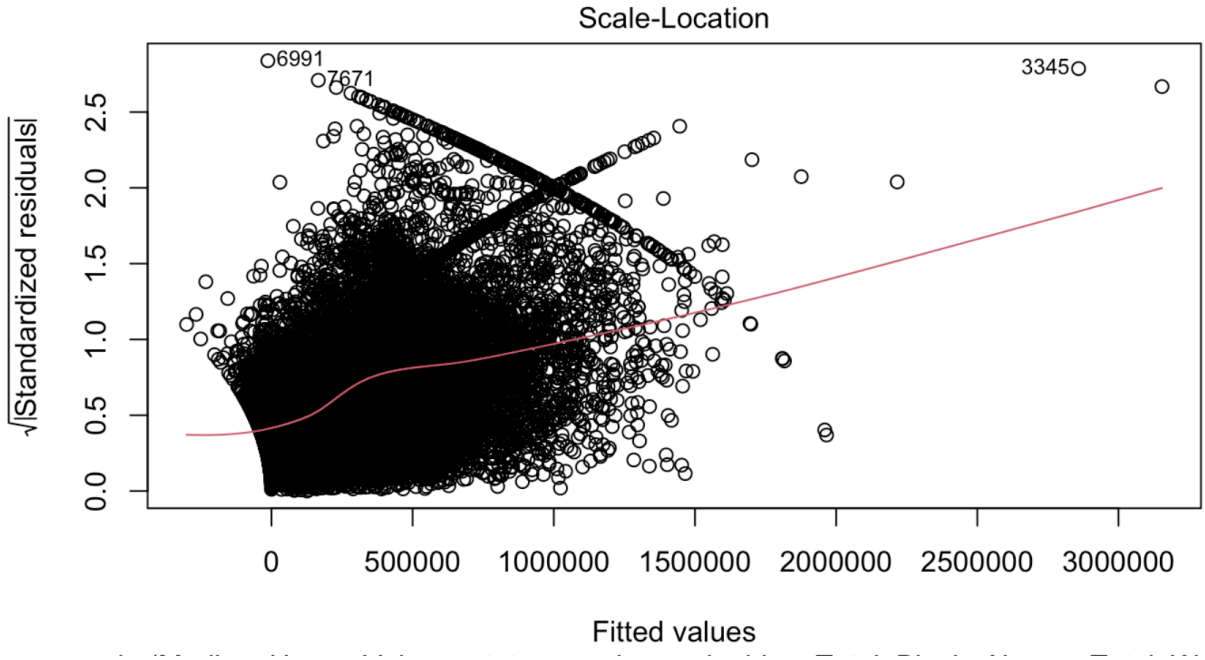
Model Assumption Checks:

It's important to acknowledge assumptions when performing regression analysis. The four primary assumptions of linear regression include linearity, constant variance, normality, and independence (James, 93). If one or more of these assumptions are not met within your model, you may need to consider transforming a variable or choosing a different predictive method.

Linearity:



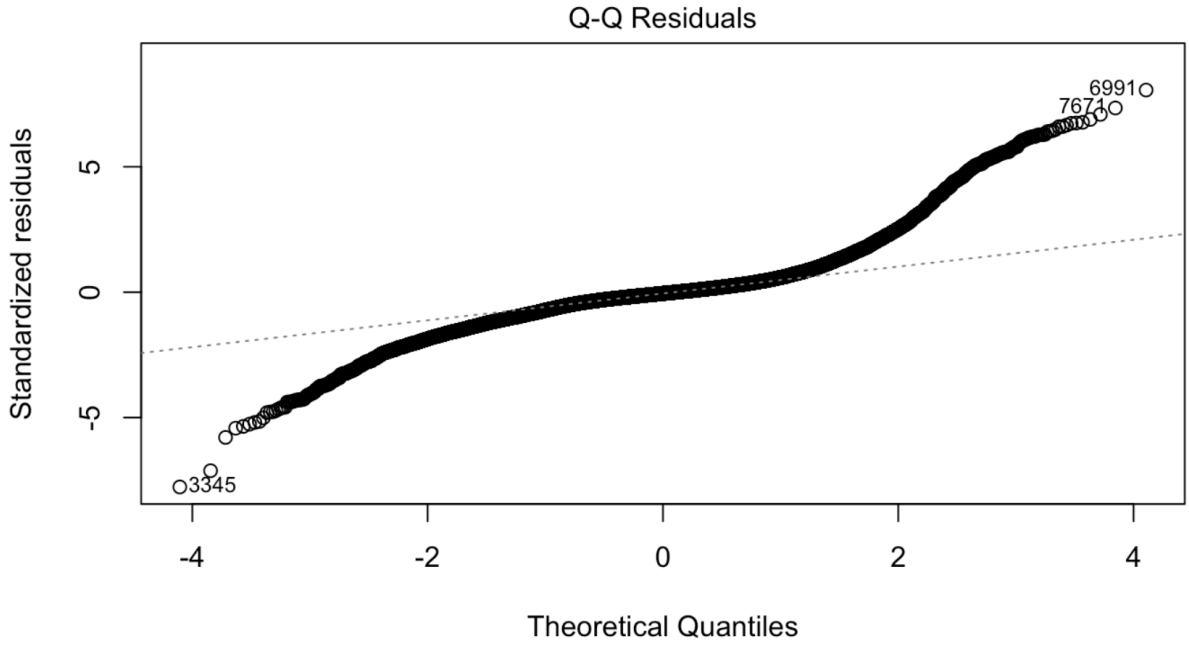
Constant Variance:



Independence:

	GVIF	Df
state	1.930538	36
mode_grade_idx	1.316818	4
Total_Black_Alone	1.530251	1
Total_White_Alone	2.364127	1
Median_HouseHold_Income	2.620937	1
Median_Year_Structure	1.116423	1
Median_Number_Rooms	1.565190	1
Agg_Earnings_Past_12_Months	3.160513	1

Normality:



From these plots, I found that most assumptions were violated. I see that the linearity assumption does not exhibit a random scatter, constant variance looks skewed, VIF multicollinearity does not seem to be an issue, you would, however, need to test with Durbin-Watson to verify if independence is validated. Finally, normality looks slightly violated because the standardized residuals don't follow the dashed line as much as I'd like them to.

This is why it's important to check assumptions before productionalizing or publishing a model to see if it is unbiased and representative of the population. Next steps on potential solutions to these issues will be stated in a subsequent section.

Final Model Interpretation:

As stated above, I decided to move forward with the model that included the six numerical variables mentioned, state, and mode block group HOLC Grade. A summary of this trained model is provided below:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.117e+04	1.495e+04	4.760	1.95e-06	***
stateAL	3.529e+04	1.710e+04	2.064	0.039036	*
stateAR	5.098e+04	3.714e+04	1.373	0.169880	

stateAZ	9.243e+04	3.424e+04	2.700	0.006946 **
stateCA	4.120e+05	1.021e+04	40.352	< 2e-16 ***
stateCO	1.781e+05	1.643e+04	10.841	< 2e-16 ***
stateCT	4.237e+04	1.278e+04	3.315	0.000917 ***
stateFL	1.406e+05	1.201e+04	11.707	< 2e-16 ***
stateGA	7.041e+04	2.063e+04	3.414	0.000642 ***
stateIA	-9.306e+03	1.540e+04	-0.604	0.545594
stateIL	4.717e+04	8.725e+03	5.407	6.47e-08 ***
stateIN	5.924e+02	1.208e+04	0.049	0.960903
stateKS	-1.376e+04	2.271e+04	-0.606	0.544508
stateKY	2.785e+04	1.779e+04	1.565	0.117490
stateMA	2.015e+05	9.791e+03	20.579	< 2e-16 ***
stateMD	2.176e+04	2.637e+04	0.825	0.409377
stateMN	2.106e+04	1.324e+04	1.591	0.111674
stateMO	1.466e+04	1.752e+04	0.837	0.402773
stateMS	7.436e+04	2.500e+05	0.297	0.766171
stateNC	8.046e+04	1.813e+04	4.437	9.16e-06 ***
stateND	-3.130e+03	2.501e+05	-0.013	0.990015
stateNE	-7.935e+03	1.732e+04	-0.458	0.646886
stateNH	4.081e+04	3.231e+04	1.263	0.206473
stateNJ	8.482e+04	9.497e+03	8.931	< 2e-16 ***
stateNY	2.258e+05	8.071e+03	27.972	< 2e-16 ***
stateOH	-8.846e+03	9.445e+03	-0.937	0.348995
stateOK	2.262e+03	2.411e+04	0.094	0.925261
stateOR	2.030e+05	1.683e+04	12.056	< 2e-16 ***
statePA	1.438e+04	9.088e+03	1.583	0.113530
stateRI	6.240e+04	1.727e+04	3.614	0.000302 ***
stateTN	8.298e+04	1.467e+04	5.657	1.55e-08 ***
stateTX	4.861e+04	1.245e+04	3.903	9.51e-05 ***
stateUT	9.845e+04	2.186e+04	4.504	6.69e-06 ***
stateVA	6.579e+04	4.861e+04	1.354	0.175901
stateWA	1.957e+05	1.337e+04	14.632	< 2e-16 ***
stateWI	1.372e+04	1.184e+04	1.159	0.246383
stateWV	3.046e+04	3.114e+04	0.978	0.328034
mode_grade_idxB	-4.469e+04	9.214e+03	-4.850	1.24e-06 ***
mode_grade_idxC	-4.976e+04	8.891e+03	-5.597	2.21e-08 ***
mode_grade_idxD	-5.460e+04	9.313e+03	-5.863	4.60e-09 ***
mode_grade_idxTied Grades	-4.425e+04	8.823e+03	-5.016	5.32e-07 ***
Total_Black_Alone	-2.415e+02	1.308e+01	-18.455	< 2e-16 ***
Total_White_Alone	-2.284e+02	1.156e+01	-19.759	< 2e-16 ***

```
Median_HouseHold_Income      2.082e+00  5.463e-02  38.105  < 2e-16 ***
Median_Year_Structure       -2.427e+01  1.801e+00 -13.474  < 2e-16 ***
Median_Number_Rooms          5.622e+03  1.447e+03   3.885 0.000102 ***
Agg_Earnings_Past_12_Months  3.813e-03  8.484e-05  44.939  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 249900 on 24829 degrees of freedom
Multiple R-squared:  0.5124,        Adjusted R-squared:  0.5115
F-statistic: 567.2 on 46 and 24829 DF,  p-value: < 2.2e-16
```

I completed an overall F test by answering:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1 : \text{ at least one of the regression coefficients is different from } 0.$$

From the p-value at the bottom of the model summary, we have evidence at the .05 alpha level that at least one of the regression coefficients is different from 0 and is therefore influencing the prediction of block group median home values in 2020.

The final model equation will be quite long, but I will provide a few examples of what various model coefficients mean within the trained model estimates summary:

- For each dollar increase in median household income, there is a $2.082 dollar increase in median home value while all other predictors are held constant, or in other words, controlling the effect of the other predictors to isolate median household income to look at its impact on median home value.
- When the median home value of interest comes from California, for example, we can say that the median home values are, on average, $412,000 more than Michigan values.
- Finally, we find that mode block group HOLC Grades B, C, and D are significant at the .05 alpha level. Therefore, we find that a median home value of interest that falls within the mode HOLC Grade D will have, on average, a $54,600 lower median home value compared to mode HOLC Grade A.

Next Steps:

Overall, this data project was a great first iteration for predicting a recent economic indicator in terms of other information about the community and further connecting that to historical information. One of the key next steps that needs to be addressed is how assumptions can be improved. I found that many of the key regression assumptions were violated and can

therefore provide inaccurate predictions of the target variable at hand. Some ways I can address this is through transforming various predictors, pulling in additional historical or economic data, or combining information together to allow a more broad model.

Another next step I would love to take is incorporating this within my redlining website. Currently, I have a geospatial representation of the HOLC grades layered over current economic information about the communities. It would be great to add a tab to this site that includes this data project and the key findings that were made. This would allow people to see the potential significant impact that historical policies have on the current communities today.

Objective 1: Describe probability as a foundation of statistical modeling, including inference and maximum likelihood estimation

Within this course, there are a few different ways that students can understand probability as a foundation of statistical learning. Firstly, P-values within statistical models are a great way to analyze significance and test various hypotheses relating to the larger topic at hand. According to documentation, "A P-value measures the probability of obtaining the observed results, assuming the null hypothesis is true" (Beers). I was able to effectively interpret the F-test within my statistical model and point to various predictors that were considered significant based on the P-value they were receiving.

Additionally, the use of cross validation and various modeling metrics such RMSE, MSE, MAE, and MAPE, provide meaningful interpretations of how the probabilistic predicted values would look based on analyzing various relationships between the observed and expected values. For example, if the mean squared error is relatively low, that means that the difference between the expected and observed values is low and therefore allows for more accurate probabilistic predictions of median home values.

Objective 2: Determine and apply the appropriate generalized linear model for a specific data context

Within the textbook titled, "An Introduction to Statistical Learning" by Gareth James, Sections 3.1 and 3.2 were quite helpful in understanding when and how to use the appropriate linear model for a data situation. For example, the most common approach for simple linear regression is said to occur when you'd like to predict a quantitative response based on a predictor variable. The notation for simple linear regression is stated as follows: $Y \approx \beta 0 + \beta 1 X$ (James, 61).

Based on the context from the textbook, I was able to apply a multiple linear regression model to a specific data context. I am able to incorporate tidy census and historical data to

predict 2020 block group median home values:

```
#Fit the generalized linear model:
fit <- lm(Median_Home_Value ~ state + mode_grade_idx + Total_Black_Alone + Total_White_Alone +
          Median_HouseHold_Income + Median_Year_Structure + Median_Number_Rooms +
          Agg_Earnings_Past_12_Months, data = train_data)

#Viewing Model Summary:
summary(fit)
```

Furthermore, I am able to acknowledge the assumptions that are needed for this model and provide recommendations for how to improve the model. The ETL processes leading up to performing the regression are crucial to understand in order to adequately predict median home values. Therefore, I feel I have completed this objective by completing a first iteration from start to finish.

Objective 3: Conduct model selection for a set of candidate models

Within the Initial Model section, I begin with a naive model that includes all numerical predictor variables in the regression model. I further refined my model by taking a subset of the numerical predictors and testing a few other models that included historical information about the communities. I was able to come up with the subset of numerical predictors by thinking about what predictors would be meaningful to the population at large and seeing how significant they were in the initial model.

The use of K-fold cross validation allowed me to understand what models were generalizing well to unseen data and also had a meaningful/interpretable prediction of median home values. From this, I was able to conclude that I did not need the "category" variable from the redlining data within my model because it was not influencing median home values according to metric interpretation and P-values.

Objective 4: Communicate the results of statistical models to a general audience

Interpreting the coefficients within the multiple linear regression model is important to understand. I feel I was able to explain numerical and factored categorical predictors in my model to a general audience. Also, I clearly understood what the reference variable acts as within qualitative predictors through my explanation of the state predictor being compared to Michigan and the mode HOLC grades being compared to the mode HOLC grade A. Additionally, I was able to explain what predictors were significant to the model, what the conclusion of the F-test was, and how model performance metrics help in understanding the accuracy of your model.

I was also able to follow-up on next steps that could be taken based on these results. This allows the audience to understand how the scope could be expanded, what improvements could be made, and if there were any challenges that occurred. Communicating the results of a statistical model to a general audience in a way that they will understand is key in getting your

points across. With that being said, I feel I have done a thorough job in explaining my findings and have met this objective.

Objective 5: Use programming software (i.e., R) to fit and assess statistical models

Throughout the entirety of this data project, I have been able to utilize the R programming language to provide informative visualizations, data transformations and preparations, and statistical modeling techniques. Various R packages I utilized were Ggplot2, Tidyverse, Dplyr, Modeest, Caret, and many more! I assessed the statistical models within the initial models and model selection sections of the data project.

Additionally, I was able to successfully complete the extract, transform, and load processes of a data project through joining various datasets together, transforming the coordinate systems, aggregating categorical information, and adequately loading this information for my various regression models. Overall, I feel I have gained a lot of valuable R knowledge within STA 631 and STA 518 by completing various activities relating to statistical model creation and assessment. I feel this data project has showcased my R skills in many ways and I look forward to implementing more projects in the future!

Participation in the Course Community:

There are many ways that I was able to contribute to our course community. Firstly, one of my favorite assignments that we did in this course was the exploratory data science job ads analysis. I was able to present my findings to the group and share GVSU courses that relate to these fields. The primary job ads I discussed were a managerial, AI, General Motors, and economic data scientist. I also was able to provide salary information and job requirements for each posting to allow students to get a gauge as to what is expected within these roles.

Additionally, I was an active participant within our course meetings, asking valuable questions about activities we had the opportunity to work on such as logistic or linear regression topics. I connected these topics to our Teams chat where I discussed important findings from *Data Feminism* and our personal skills within data science. By sharing professional experiences I had within data science, it allowed students to compare their work with mine and prompted meaningful conversations about education and the real world. Overall, I feel I have gained valuable knowledge from this course and put in the effort to learn new ideas relating to model selection, interpretation, and data transformation.

Works Cited

Beers, Brian. *P-Value: What It Is, How to Calculate It, and Why It Matters*, 1 Aug. 2024, www.investopedia.com/terms/p/p-value.asp#:~:text=Key%20Takeaways,significance%20of%20the%20observed%20difference.

CODING FOR CATEGORICAL VARIABLES IN REGRESSION MODELS | R LEARNING MODULES, UCLA, stats.oarc.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/.

James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*. 2nd ed., 2021.

K-fold Cross Validation in R Programming, GeeksForGeeks, 28 Dec. 2021, www.geeksforgeeks.org/k-fold-cross-validation-in-r-programming/.

Mapping Inequality, University of Richmond, dsl.richmond.edu/panorama/redlining/.

Mode Calculation in R: A Step-by-Step Guide, SQLPad, 5 May 2024, sqlpad.io/tutorial/mode-calculation-r-step-by-step-guide/#:~:text=A%3A%20You%20can%20calculate%20mode,offer%20functions%20for%20mode%20calculation.

Basic usage of tidycensus, edited by Kyle Walker, Tidycensus, walker-data.com/tidycensus/articles/basic-usage.html.

Scale: Scaling and Centering of Matrix-like Objects, DataCamp, www.rdocumentation.org/packages/base/versions/3.6.2/topics/scale.

Spatial join, spatial filter, edited by Edzer Pebesma, Github.io, r-spatial.github.io/sf/reference/st_join.html.

Split the Dataset into the Training & Test Set in R, GeeksForGeeks, 25 July 2022, www.geeksforgeeks.org/split-the-dataset-into-the-training-test-set-in-r/.

Topological relations, Github.io, tmieno2.github.io/R-as-GIS-for-Economists/topological-relations.html.