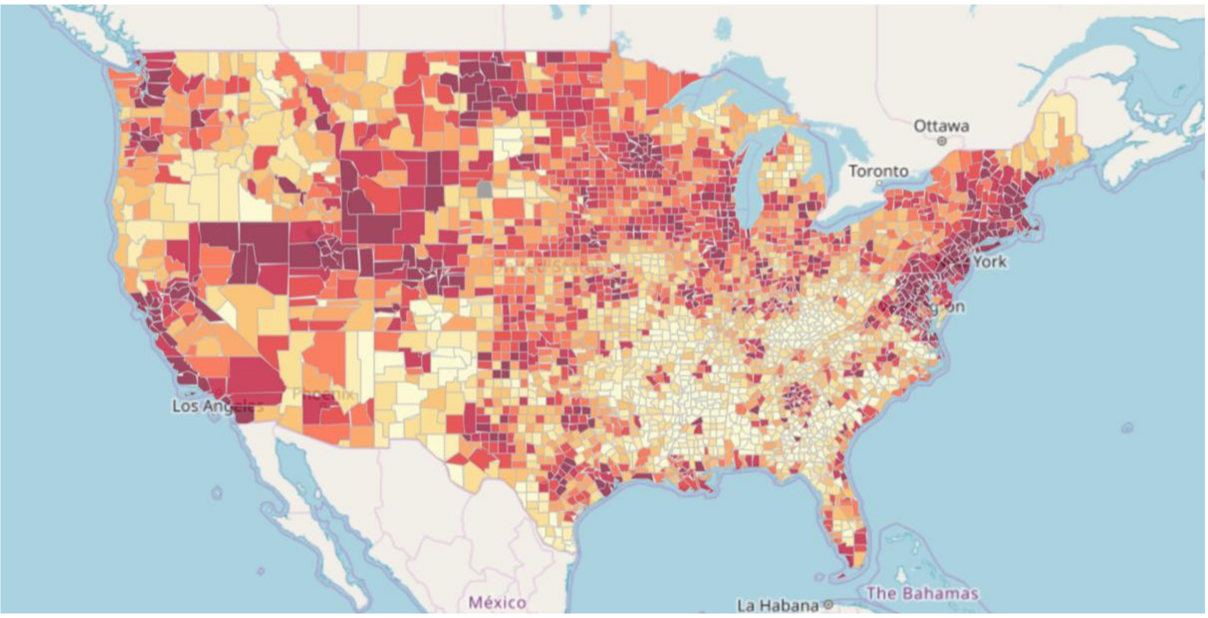
Understanding the Effect of Demographic factors on County Median Incomes using Linear Regression



[4]

Sanchayni Bagade, A20425171
sbagade@hawk.iit.edu

ABSTRACT

The median income of a county is the best indicator of the overall health of the economy at the micro level since it reflects the livelihood, education and state of governance in a county. Hence, predicting median income and understanding the causal factors that affect it can prove to be extremely useful for planning and evaluating the impact of public policy. Linear regression proves to be an effective technique to model linear relationships between dependent and independent variables. In this report, we use linear regression to understand the relationships between the factors that affect the median income at the county level in US. The dataset was generated using publicly available US demographic data for 2016. After removing outliers and feature engineering, the linear regression model found key factors that predict income at a county level to a high degree of accuracy.

1. INTRODUCTION

Predicting the median income at a geographic grain is an extremely useful exercise for city planners, policy makers and demographers. The median income is a great indicator of standard of living of households and hence a great metric to track to evaluate the outcomes of policy measures.

In this report, we chose to model the median income at the county level for US counties in the year 2016. The dataset was created using publicly available US demographic data. Given the nature of the problem, we felt that this problem would be a great candidate to be solved using linear regression as we expect to see a linear relationship between the dependent and independent variables.

However, the abundancy of data available increases the dimensionality of the dataset and we took great care to remove collinear variables in order to ensure that the model generalises. We also accounted for outliers in the dataset which might be due to poor data quality.

This paper is thus organised into sections which detail each step of this effort. We cover the data set collection in Section 2 and the succeeding sections talk about the data pre-processing, outlier removal and model training steps. We dive deep into model results in Section 4.

2. DATA SOURCE

For this analysis we sourced data from the Fact finder census gov, a major source of US demographic data. Based on previous research and our intuition we determined variables that would likely be predictive of median income. We then created the dataset by pulling in these features at a county level for the year 2016.

The variables we chose are:

1. **Earning Ability**: Variables like age and education level which represents a person's capability/earning ability.
2. **Occupation Type**: Percent of people engaged in sectors like Sales and office occupation, management, business, natural resources and so on provides an insight of labor division at overall county level.
3. **Employment Statistics**: Variables like employment rate which are in some sense and indirect or an alternative variable for median income.
4. **External Factors**: Factors like tax rates at the state level, percentage of people married and sex ratio which are known to directly affect the income of a household were included.

The table below provides details on the types, nature and source of each of the variables.

Sr. No	Table	Variable	Source
1	Demographic data	Occupation – 5Levels	American factfinder[1]
		Class of Worker – 4Levels	
		Education level – 3Levels	

		Sex Ratio	
		Employment percentage	
		Median Income	
		ID and ID2	
		State	
		County	
2	Geospatial data	Latitude, Longitude, County and State	2018 U.S. Gazetteer Files
3	Tax data	Property Tax	American factfinder[2]
		Sales and receipt taxes	
4	Marital Status	Percentage of people currently married	American factfinder[3]

3. EXPLORATORY DATA ANALYSIS AND PRE-PROCESSING

3.1. Missing value treatment

There exist missing values in only two columns which are present in the Tax table. Property Tax has 1033 missing values whereas Sales and receipt taxes has 39. We will be imputing this missing data using KNN imputation which predicts the missing value based on the tag of its neighbours. Let’s say two people live in states like California and New York which has most of their demographic values similar then we can say both of them might be paying around the same tax rate. We are being careful with imputing the data by median rather than mean to account for variance in mean due to outliers.

3.2. Scaling

Most of our independent variables are either a percent value (range: [0-100]) or a ratio in 0-1 bounded range thus don’t require any major scaling. However, the dependent variable Median income histogram looks like as below which has a tail towards the right (left skewed) and thus require scaling:

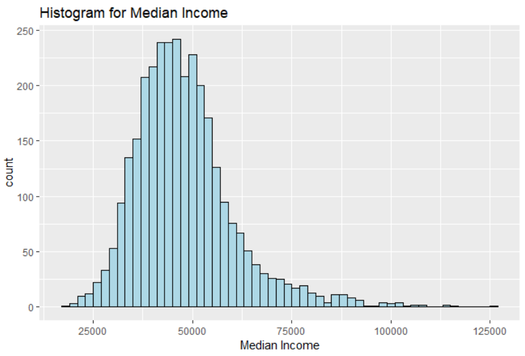


Fig 3.1 Unscaled Median Income

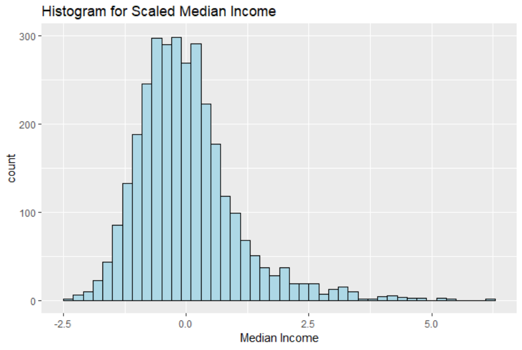


Fig 3.2 Scaled Median Income

3.3. Feature Engineering

We included two derived variables namely, “Working Class” which was defined as the percent of population within the age group of 18 to 74 and “Dependency ratio” is the ratio of dependent population (0-18 age group plus 75+ age group) divided by working population.

Based on our understanding of the problem statement and the independent variables considered for analysis we chose to exclude variables like Percentage of families below poverty line which is nothing but a direct indicator of median income. We also be removed “Employment rate” from the mix as it was a highly correlated with other variables.

3.4. Multicollinearity

(please find below in appendix variable definition table)

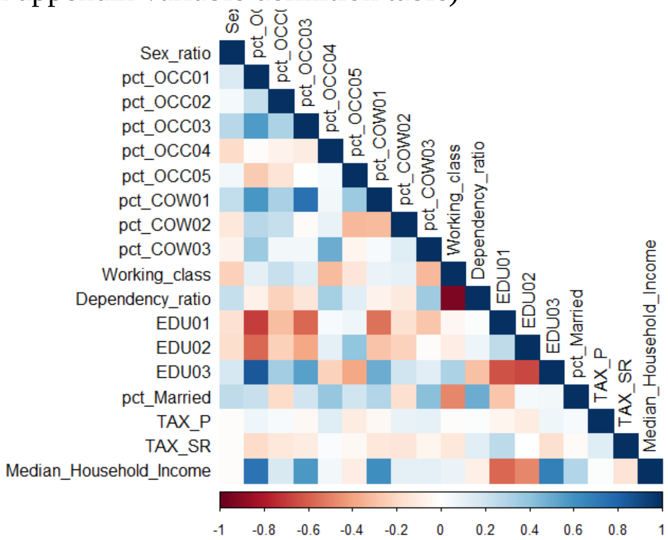


Fig 3.3 Correlation plot

As we can see from the correlation plot above there exists very high collinearity between variables as expected. Multicollinearity in independent variable leads to high variance in the beta values predicted in linear regression. Variance of coefficients in linear regression is defined by $var(\beta) = \sigma^2(X^tX)^{-1}$ and multicollinearity lead to singularity of the inverse term and thus high variance. For removing multicollinearity, we performed Lasso and Ridge regression, Principal Component Analysis (PCA) as well as stepwise subset selection. We will be comparing results of all the mentioned method for removing multicollinearity and variable reduction.

Lasso and Ridge regression both add an extra bias to the regression estimate and thus control the beta variance. Ridge uses L2 norm whereas Lasso uses L1 to achieve this goal.

Sr No	**Variables**	**Linear regression coefficients**	**Ridge regression coefficients**	**Lasso coefficients**
1	β0	2.559026	0	0
2	Sex_ratio	-3.3578	-0.239968723	-0.182743020
3	pct_OCC01	-0.1175	0.264989191	0.360813362

4	pct_OCC02	-0.2384	-0.072255200	0
5	pct_OCC03	-0.1459	0.098256454	0.066348220
6	pct_OCC04	-0.1503	0.060623247	0
7	pct_OCC05	-0.1944	-0.010304716	0
8	pct_COW01	0.23121	0.241769993	0.195010533
9	pct_COW02	0.22405	0.072655490	0
10	pct_COW03	0.07563	-0.174093413	-0.086521546
11	Working_class	-0.0469	-0.076566894	0
12	Dependency_ratio	-1.2678	-0.025099712	0
13	EDU01	-0.0067	-0.066795180	-0.006325159
14	EDU02	-0.0122	-0.096642597	-0.058225635
15	EDU03	0.04452	0.244544862	0.185478273
16	pct_Married	3.03895	0.186946934	0.168179065
17	TAX_P	-0.4736	-0.021029117	0
18	TAX_SR	-0.059	-0.008199343	0

Lasso shrinks few variables to zero because of this underlying use of L1 norm rather than L2. We also choose a value of lambda such that significant number of variables shrink to zero. Ridge doesn’t really shrink coefficients of any variable to zero but if we look at the absolute value of coefficients, we can say that both the methods select almost the same variable which are as follows*:
Sex_ratio,pct_OCC01,pct_OCC03, pct_COW01, pct_COW03,EDU01, EDU02, EDU03 and pct_Married

Principal component analysis is a method of creating orthogonal variables from the set of input variables. It transforms the variables in such a way that the first PCA component would explain the maximum variance in the input data. When we apply PCA on our 17 variables we get 17 new PCA components with variance explained by them as follows:

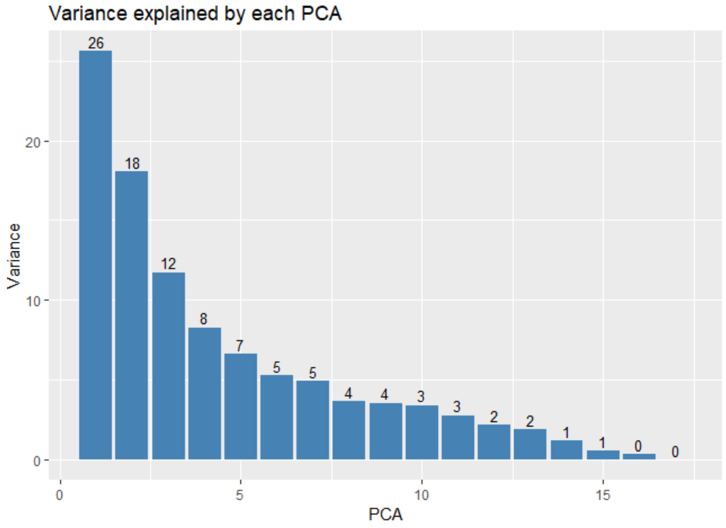


Fig 3.4 Variance explained by PCA components

[*] The other perspective at deciding which variables to pick based on Ridge would be to check percent change in the coefficients. More the shrinkage in the coefficients more it has been penalized for variance. Thus, variables with minimum percent change can be included in the model.

As we can see the variance explained by the PCA decreases gradually (it is of an advantage if most of the variance in the data is explained by first 1-2 components) and it takes up to 7 Principal Components to explain 80% of the variance in the data. To be able to capture and accurately measure the median income we will have to at least consider 5-6 PCA components which would make the result difficult to interpret as these PCA are linear combination of 18 input variables. Thus, to keep our model simple and easy to interpret we will be not using PCA components.

Stepwise subset selection method can also be viewed as a method for understanding the effect of multicollinearity on model predictions. We performed an exhaustive subset selection as our number of predictor variables are relatively less (less than 30). We used Bayesian Information criterion (BIC) for selecting the best subset as it penalizes the model significantly for considering more variables. Following is a scatterplot of BIC versus number of variables included in the model:

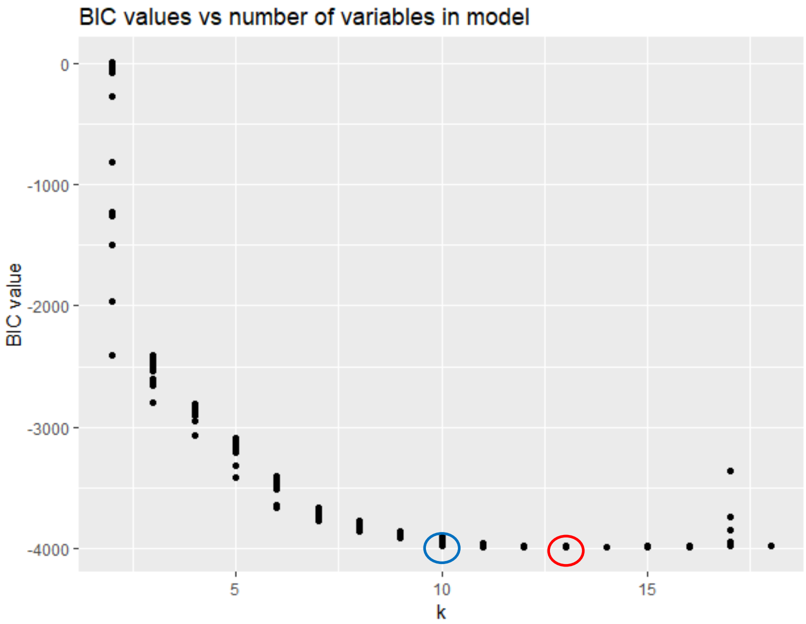


Fig 3.5 BIC plot for subset selection, the red circle is the point of minimum bic value whereas the blue circled point is the k value chosen

The model with minimum BIC has following 12(excluding intercept) variables considered in the modelling: Sex_ratio, pct_OCC01, pct_OCC02, pct_OCC03, pct_OCC04, pct_OCC05, pct_COW01, pct_COW02, Working_class, EDU02, EDU03 and pct_Married. We can also see that there isn't a drastic different in the minimum BIC value at k = 10 and at k = 13 and as our goal here was to reduce multicollinearity and number of variables to we chose the model with minimum BIC at k = 10. Thus, the variables included in the model using BIC for subset selection were:
Sex_ratio, pct_OCC01, pct_OCC03, pct_OCC04, pct_OCC05, pct_COW03, Working_class, EDU03 and pct_Married.

We can relook at the corrplot of the variables from all the above-mentioned methodologies:

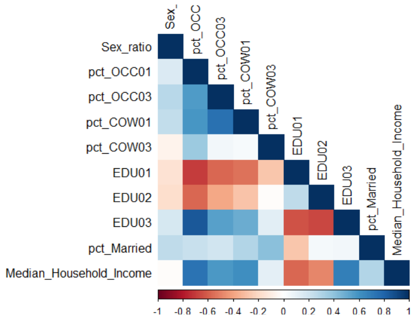


Fig 3.6 Correlation plot of variables selected using Lasso/Ridge

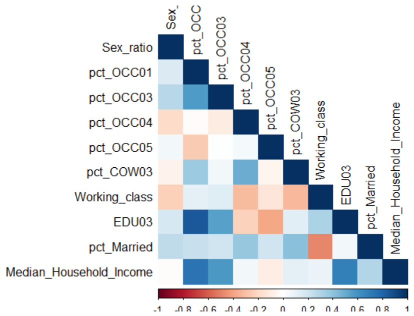


Fig 3.7 Correlation plot of variables using subset selection (BIC)

There still existed significant correlation within dependent variables in case of Lasso regression, whereas in the case of subset selection we had relatively less multicollinearity. The final approach we tried was to select variables based on our understanding of the problem statement and overall correlation plot. The corresponding correlation plot is as follows:

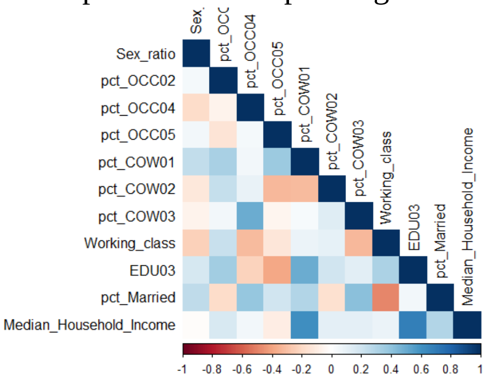


Fig 3.7 Correlation plot of variables selected based on intuition

	Lasso/ Ridge model (model_1)	Subset selection model (model_2)	Intuitive model (model_3)
AIC	5020.030	4873.016	4862.652
BIC	5080.556	4939.595	4935.284
Adjusted R^2	0.7116	0.7249	0.7259

Both subset selection and model 3 performed better than Lasso regression (at lambda= 0.05) in all the three aspects. We proceeded with model 3 as it had less correlated variables with the same performance.

3.5. Outliers and Influential points

For outlier and influential point removal we used a model for tests like studentized deleted residual test and calculating DFFITS measure.

Identifying outliers with respect to X:

Using Hat matrix

(outlier considered for the given set of X variables considered in model 2)

For identifying outliers with respect to X we calculated the leverage of each data point. There exist around 282 outliers which account for around 9% of the data. As it was difficult to verify if these really were outliers or mere exceptions due to the high dimensionality of the data and limited domain knowledge, we chose to remove those outliers.

Identifying outliers with respect to Y:

Studentized deleted residuals [6]

We calculated studentized deleted residual(t_i) values for every n values of Y using the following formula:

$$t_i = \frac{e_i}{\sqrt{MSE_i(1-h_i)}}$$

And if the t_i value exceeded the corresponding t statistics, $t(1\text{-}\alpha/2n\ ;\ n\text{-}p\text{-}1)$ value then it was considered as an outlier. The t value was calculated to 4.322 and based on the above stated criterion we found 6 outliers with respect to Y variable and which were removed from modelling dataset.

Influential points:

DFFITS and Cook's distance

The DFFITS value for the ith case represents the number of estimated standard deviations of Yhat that the fitted value Yhat increases or decreases with inclusion of the ith case in fitting the regression model[7]. We calculated DFFITS values for all n entries and as we had a reasonably sized dataset our threshold value would be 1. Under the mentioned criterion we found no influential points.

Similarly, we calculated Cooks distance using the following formula and it had a threshold of 0.5. As seen in the plot below, we didn't find any influential points.

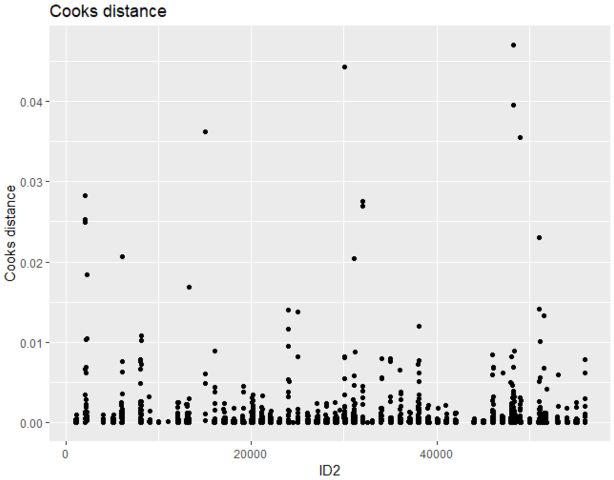


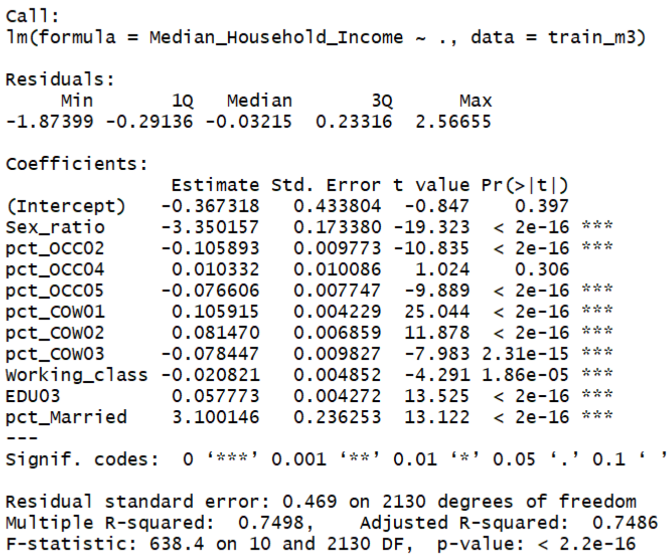
Fig 3.8 Influential points using Cooks distance

$$(DFFITS)_i = \frac{\hat{Y}_\iota - \widehat{Y_{\iota(\iota)}}}{\sqrt{MSE_{(i)} * h_{ii}}} \text{ and Cooks distance } = \frac{\sum(\hat{Y}_\iota - \widehat{Y_{\iota(\iota)}})^2}{p * MSE}$$

4. MODEL DIAGNOSTICS

The dataset was randomly divided in 1:3 ratio of testing and training sets. The model was trained on and refined using the training dataset and further validation and testing was done on the smaller testing dataset.

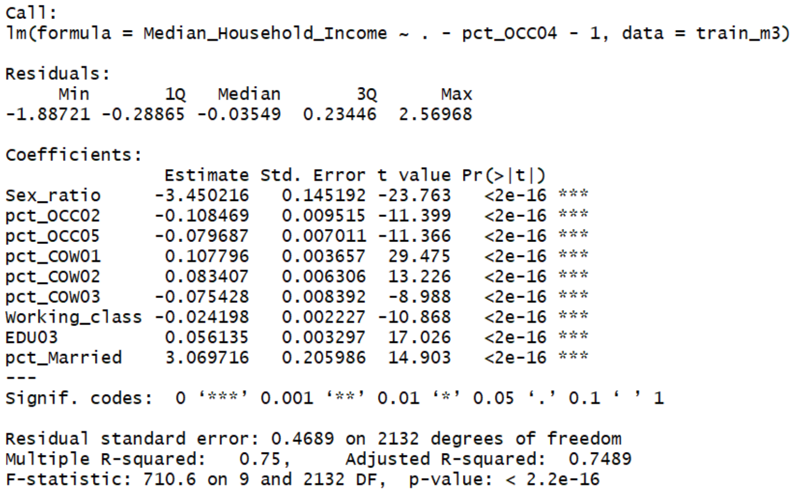
Summary of the model built on the variables selected by model 3:



```
Call:
lm(formula = Median_Household_Income ~ ., data = train_m3)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.87399 -0.29136 -0.03215  0.23316  2.56655 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.367318   0.433804  -0.847    0.397    
Sex_ratio    -3.350157   0.173380 -19.323  < 2e-16 ***
pct_OCC02    -0.105893   0.009773 -10.835  < 2e-16 ***
pct_OCC04     0.010332   0.010086   1.024    0.306    
pct_OCC05    -0.076606   0.007747  -9.889  < 2e-16 ***
pct_COW01     0.105915   0.004229  25.044  < 2e-16 ***
pct_COW02     0.081470   0.006859  11.878  < 2e-16 ***
pct_COW03    -0.078447   0.009827  -7.983 2.31e-15 ***
Working_class -0.020821   0.004852  -4.291 1.86e-05 ***
EDU03         0.057773   0.004272  13.525  < 2e-16 ***
pct_Married   3.100146   0.236253  13.122  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.469 on 2130 degrees of freedom
Multiple R-squared:  0.7498,	Adjusted R-squared:  0.7486 
F-statistic: 638.4 on 10 and 2130 DF,  p-value: < 2.2e-16
```

Fig 4.1 Model summary

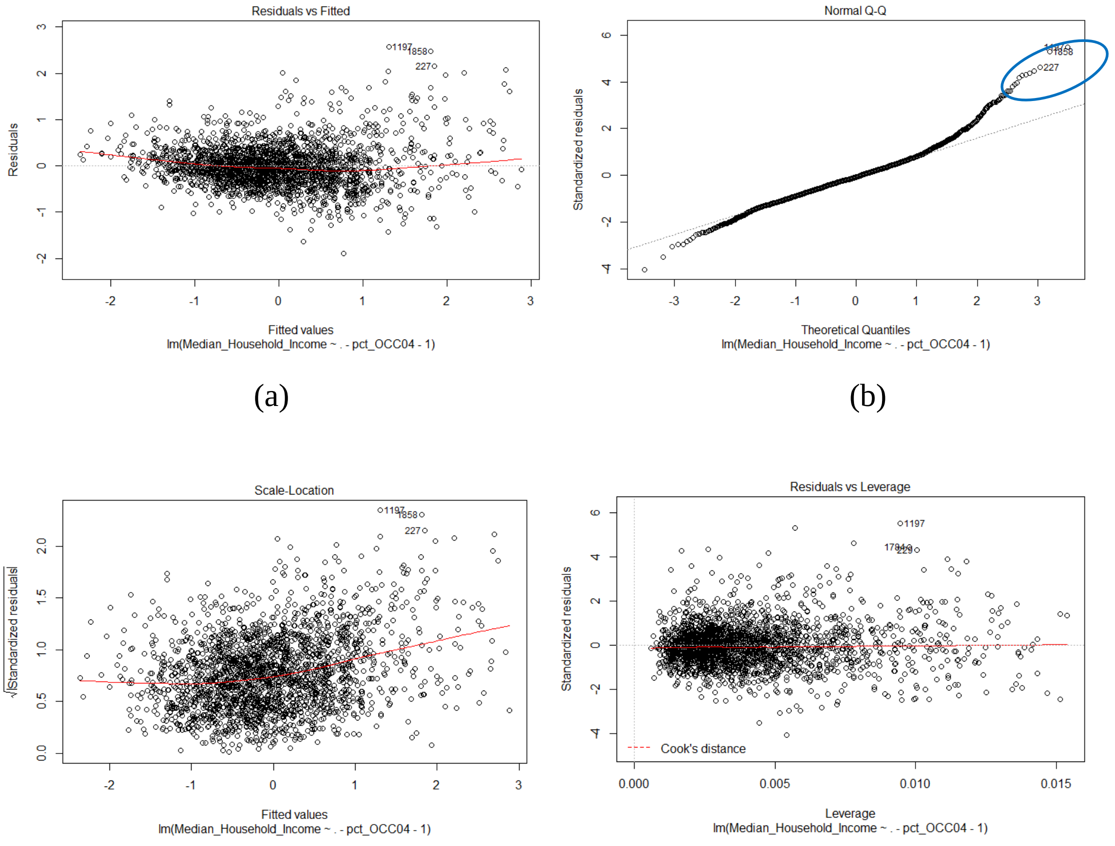
As we can see variable OCC04 and the intercept had a relatively high p value and thus were insignificant. We dropped these two variables and the final model summary is as below:



```
Call:
lm(formula = Median_Household_Income ~ . - pct_OCC04 - 1, data = train_m3)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88721 -0.28865 -0.03549  0.23446  2.56968 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
Sex_ratio    -3.450216   0.145192 -23.763   <2e-16 ***
pct_OCC02    -0.108469   0.009515 -11.399   <2e-16 ***
pct_OCC05    -0.079687   0.007011 -11.366   <2e-16 ***
pct_COW01     0.107796   0.003657  29.475   <2e-16 ***
pct_COW02     0.083407   0.006306  13.226   <2e-16 ***
pct_COW03    -0.075428   0.008392  -8.988   <2e-16 ***
Working_class -0.024198   0.002227 -10.868   <2e-16 ***
EDU03         0.056135   0.003297  17.026   <2e-16 ***
pct_Married   3.069716   0.205986  14.903   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4689 on 2132 degrees of freedom
Multiple R-squared:   0.75,	Adjusted R-squared:  0.7489 
F-statistic: 710.6 on 9 and 2132 DF,  p-value: < 2.2e-16
```

Fig 4.2 Final model summary

The AIC and BIC of the above model is 2843.909 and 2900.599 respectively and has all the predictive variables significant. Following are the diagnostics plots for the model:



(a) (b)

(c) (d)

Fig 4.3 Diagnostic plots

It can be observed from the fitted value versus residual plot that there doesn't exist a significant trend between Yhat and the residuals. We also performed the Breusch-Pagan test to justify the same. From the QQplot, Fig 4.3(b) of the standardized residual we can conclude that the error distribution has a heavy tail towards the right. The same can be verified from the histogram plot of our output variable Median income (Fig. 3.2). Scatter plot 4.3(d) validates our previous claim of zero influential points in the data given the model.

Constancy of Error variance:

Breusch-Pagan Test

This test is used to check if there exists a linear relationship between the error terms and our input variables and a positive result concludes that there exists further trend in the output variable that has not been explained by the model. This leads to the requirement of higher order/ transformed terms in the modelling. We fit a linear model between the log of error term and independent variables and test statistics X^2_{BP} is calculated as:

$$X_{BP}^2 = \frac{SSR^*}{2} \% \left(\frac{\text{SSE}}{n}\right)^2$$

Which gave us a value of 4.5974 which is less than the X^2 (1- α, p-1) which is 18.307 which concludes H_0 : $\log(e^2) \neq \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1}$

F test for the model:

$H_0 : \beta_0 = \beta_1 = \beta_2 = \ldots \ \beta_{p-1} = 0$

H_a : One of the β is not equal to zero

$F^* = MSR/MSE$

F^* can be found in the summary of the model which is equal to 710.6

The corresponding *F* statistics value is *F(1- α, p-1 , n-p)* = 1.601. Thus $F^* > F$ statistics thus we reject the Null hypothesis.

***Performance on Test dataset*:**

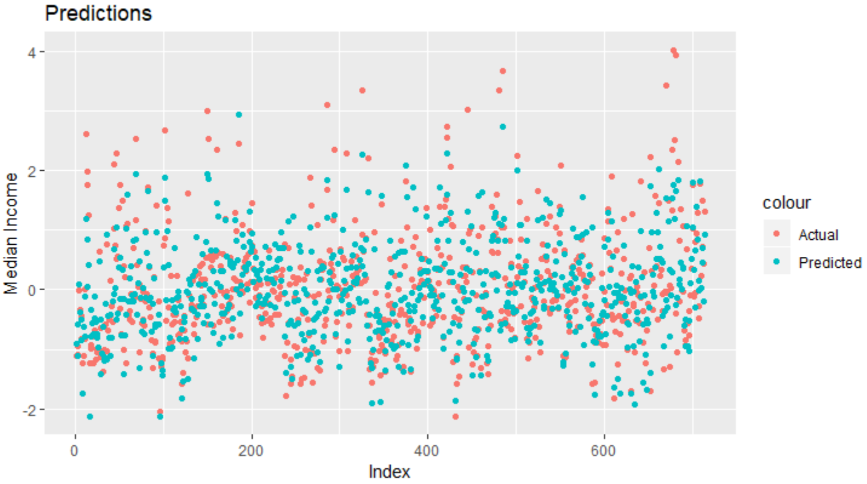


Fig 4.4 Performance on test dataset

The MSPE value for the test dataset is 0.2494, this value is very low as we have standardized our output variable. We can compare this value with the baseline model of $\widehat{Y_\iota} = \bar{Y}$ (0.01898) which is 0.8956. The final model performs 72% better than the baseline model on the training dataset.

As we have scaled the Median income variable our final unscaled prediction expression would be stated as follows:

$$\frac{Y_i - \bar{Y}}{sd(Y)} = -3.45 * Sex\ ratio - 0.108 * pct\ OCC02 - 0.0796 * pct\ OCC05 + 0.1077 * pct\ COW01 + 0.083 * pct\ COW02 - \ 0.0754 * pct\ COW03 \quad - 0.024 * Working\ class + 0.0561 * EDU03 \ + 3.0697 * pct\ Married$$

$\bar{Y} = \ 47973.23$

$sd(Y) = 12605.98$

Final Model:

$$Y_i = 47973.23 - 12605.98 * (3.45 * Sex\ ratio - 0.108 * pct\ OCC02 - 0.0796 * pct\ OCC05 + 0.1077 \quad * pct\ COW01 + 0.083 * pct\ COW02 - \ 0.0754 * pct\ COW03 \ - 0.024 * Working\ class + 0.0561 * EDU03 \ + 3.0697 * pct\ Married$$

5. CONCLUSION

In this project we successfully created a dataset from publicly available US Demographic data and trained a robust linear regression model that can predict the median income at the county level. A number of derived variables were created and we performed a thorough analysis of the multicollinearity between the variables and in removing outliers. During exploratory data analysis step, we came across various key insights in our data like occupation type *Management, business, science and arts occupations* has a high correlation with median income which is intuitive as these are the high paying occupation types. In our final model we have occupation type variables as *Service occupations* (housekeeping, tours, nursing and teaching) and *Natural resources, construction, and maintenance occupations* which are very significant and have negative coefficients which can be comprehended as having negative effect on median income as they are high efforts but less paying occupation types. The same can be said about class of workers, *Private wage and salary workers* are paid relatively more than the counter worker type. As for the *Self- employed class of worker* the source of income can have a high variance as this category spans large horizons from Manufacturing to entertainment. The negative coefficient associated with it can be interpreted as not having a complete track of all the source of income, a smaller number of people opting for the class type or merely due to huge variance in the income. As anticipated, Education type- *percentage of people with Bachelors or higher* has a positive and significant effect on the predicted variable.

Variables like *Sex ratio, Working class and pct_Married* are difficult to interpret in our model. Working class variable is expected to have a positive effect on median income as population within the working-class range is a measure of a county's potential to make higher impact on the GDP but in our model it predicts otherwise. Negative coefficient of Sex ratio might be pointing towards the wage discrimination by gender but we can't be certain about this correlation to be a causality.

The final model that was created had MSPE of 0.2494 on the test dataset which was a 72% improvement over the baseline model. Overall, we have a robust model with relatively less variance in predicted coefficients with R^2_{adj} value as 74.89. Our dependent variables reflect on the health and performance of the county economy.

Future work:

(a) Geo plotting/ analysis in python for understanding distribution of various factors throughout the country. It wasn't possible in R as the necessary feature is paid.

(b) In model diagnostics section we found that the variance of the error terms is not significantly related with the fitted output but the corresponding residual vs fitted values does seem to have a trend. This change in variance can be further reduced by transformation of output variable, preferable log.

(c) Analysing difference between lasso and ridge for feature selection in-depth.

6. APPENDIX

Variable definition table

Acronym	Definition
OCC01	Management, business, science, and arts occupations
OCC02	Service occupations
OCC03	Sales and office occupations
OCC04	Natural resources, construction, and maintenance occupations
OCC05	Production, transportation, and material moving occupations
COW01	CLASS_OF_WORKER-Private wage and salary workers
COW02	CLASS_OF_WORKER - Government workers
COW03	CLASS_OF_WORKER -Self-employed in own not incorporated business workers
COW04	CLASS_OF_WORKER - Unpaid family workers
EDU01	pct_Less than high school graduate
EDU02	pct_High_school_graduate
EDU03	pct_Bachelors_degree_or_higher
BPL	pct_of_families_below_poverty_line_past1yr
Dependency ratio	(Age_group_5to17+ Age_group_75andover+ Age_group_50to74)/ (Age_group_18to34 + Age_group_35to49)
Working class	(Age_group_18to34 + Age_group_35to49)
Sex Ratio	Total Female population/ Total male population
TAX_P	Property Tax
TAX_SR	Sales and receipt taxes
Employed_pct	Number of people employed/ Total population
Pct_Married	Number of people currently married/ Total population

7. References and Citations

1. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_5YR_DP03&src=pt
2. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=COG_2012_COG001&prodType=table
3. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_5YR_B12002&prodType=table
4. Geospatial image: https://rpubs.com/yuorme/267270
5. Applied Linear Statistics Models, fifth edition. Authors: Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li.
6. https://www.whitman.edu/Documents/Academics/Mathematics/2017/Perez.pdf