# ST538 Project 1

# Analysis of Maryland Resident’s Income in 2019

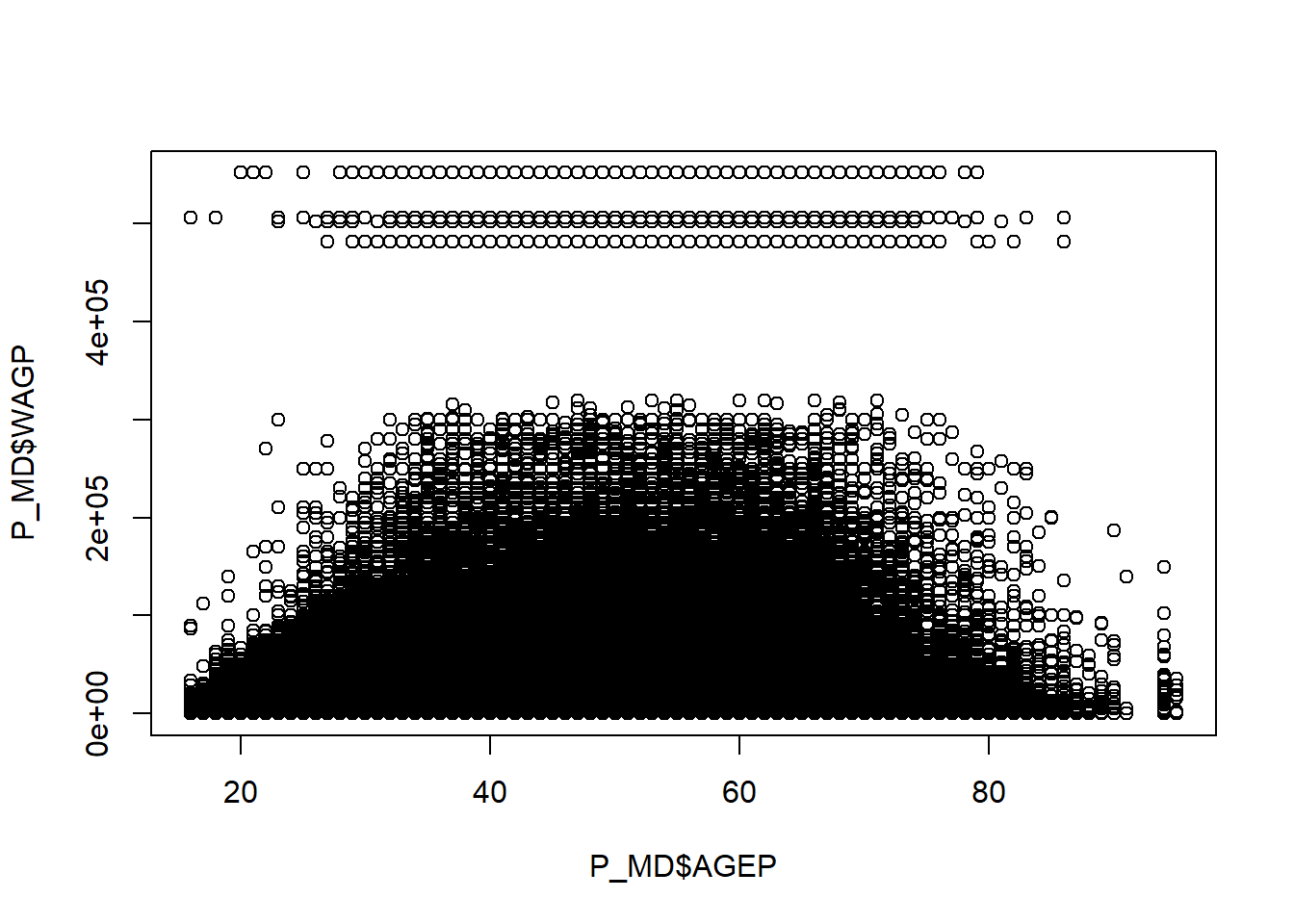
## Introduction

Using data from the American Community Survey conducted by the US Census Bureau, an analysis of Maryland residents’ income was conducted to identify variables which significantly impact an individual’s yearly earnings. This analysis seeks to answer the following two questions: 1.) In Maryland in 2019, what was the effect of being employed in a given sector or industry of the economy on the wages earned over the last year? 2.) Of the people in Maryland who made more than $10,000 USD in the last 12 months, what is the effect their industry has on their income? Additionally, this analysis seeks to predict the income of a person in Maryland based on educational attainment and age. The variables of interest in this paper are individual income, age, sex, level of education, worker classification, and occupational classification.

## **Q1: In Maryland in 2019, what was the effect of being employed in a given sector or industry of the economy on the wages earned over the last year?**

## Exploratory Analysis

To get an idea of how to approach this analysis, we looked at exploratory plots of the explanatory variables against the WAGP variable as the response.



We then fit an initial model with the WAGP variable as the response and the AGEP variable as the explanatory variable. From this initial model fitting, we were able to see that the residuals were not evenly distributed. The log transformation helped the distribution substantially although, we found there to still be a potential need for higher-order transformation because the residuals exhibited some non-linearity.

## We may want to remove the responses from people who said they attended zero years of school. Given that these people have MUCH higher incomes than those with one or two years of school, we have to assume that this was due to misreporting. We will report only for those who have had any school.

## When we log transformed wage, school suddenly appeared to have a decreasing effect on wage, because that is the case for people who didn't graduate high school. Further transformations were implemented via a Box Cox transformation for the WAGP variable.

## Model Fitting/Selection

We used best subsets to look for a good model that considers polynomials of age and school, as well as interactions of age and school. We evaluated the results of the best subsets selection method using the rss, adjusted r squared, cp and bic statistics. Most statistics are optimal with an 8 variable model, so we will use the 8 variable model. We then added our variable of interest, COW, to the model so that we could report the effect of class of worker on income, after adjusting for all of our previous explanatory variables.

## Results

Our findings are relevant for people in Maryland who make more than $100/year and less than $400,000 per year, and have had one year or more of schooling. To interpret the results of our coefficients, we needed to back transform the log of the WAGP variable. Our coefficients indicated that, after adjusting for age, sex, and schooling, the federal government was the only worker class that resulted in a wage increase, compared to that of the private industry (our reference category). The coefficient of 5.3 (after adjusting to account for our log transformation) indicated that a being an employee of the federal government increased wage by about 5.3 percent as compared to being a private sector employee. Each other class of worker showed a wage decrease as compared to the private sector.

## **Q2: Of the people in Maryland who made more than $10,000 USD, how much was their income impacted by the industry they work in and the type of worker they are?**

## Exploratory Analysis

For this investigation, we used the WAGP variable as the response along with the AGEP, COW, SOCP, and SCHL variables as explanatory variables. We were not interested in anyone under the age of 18 nor anyone who had received retirement income since these observations had incomes which varied significantly. We removed individuals who did not attend any school as it appears this factor level of the SCHL variable may be due to a reporting error. The income for individuals who did not attend any school is greater than individuals who attended any amount of school. Additionally, we removed individuals whose income appeared to be an outlier (<$10,000 | >$400,000). Our question of interest seeks to create a floor for the income variable as a way to ensure the individuals had some level of income which could be analyzed. Individuals who made less than $10,000 may have been influenced by several external factors.

The standard occupational classification (SOCP) variable was too specific to answer our question, which resulted in a transformation of the variable. Using the PUMS data dictionary, we applied a more generalized classification for each individual where the first three numbers of the SOCP variable corresponded to their field of employment.

A boxplot was created to compare the income level amongst the different occupational classifications. This boxplot suggested that there were number of high earners which were still considered outliers. Since it appeared the untransformed data contained outliers, a log transformation of the response variable was applied. This transformation helped to drastically reduce the number of outlier data points.

Chart, box and whisker chart

Description automatically generated

As mentioned earlier in this report, it was found whether an individual graduated from high school had a significant impact on their earnings. For this reason, the high school graduation indicator variable was also included in this analysis.

## Model Fitting/Selection

The generalized linear model was created which used the log of individual wages earned (WAGP) variable as the response. The output of this model suggested that the model was significant at a p-value of < 0.001. In addition to the overall model significance, the majority of the individual variable coefficients of the model were reported at a significance level of <0.05. A plot comparing the model residuals to the fitted values was created. This plot showed the points were normally distributed, ensuring that the model was appropriate for the data. This suggests income can be adequately explained by a linear combination of the explanatory variables.

Model selection was preformed using the grouped lasso technique as there were two factor variables in the data that contained high numbers of factors. The COW variable contained 8 different factors and the SOCP variable contained 26 different factors. These columns are grouped together using the groups argument within the grpreg() function. The grouped lasso model we created used the Lasso penalty.

After the construction of the model, the coefficient paths were plotted against the lambda penalty. It should be noted that a number of paths remained centered about zero which suggesting that the grouped lasso model had indicated these variables did not have an effect on the log(WAGP) of an individual. Through cross validation, we were able to calculate the coefficients based on the grouped lasso model.

## Results

The coefficients of the grouped lasso model suggested that there were several factors within the standard occupational classification (SOCP) variable that were significantly different than zero. The lasso regression model identified the ENG, LGL, and MGR occupational classes as having a large positive impact on an individual’s log wages. At a minimum these occupational classes add over $1000 to an individual’s yearly income when compared to the model intercept. The EAT, EDU, HLS, MIL, PRS, and TRN occupational classifications had large negative impacts on the log(WAGP). At a minimum, these occupational classes explain a decrease in income of $1,000 when compared to the model intercept. From the grouped lasso model, we can conclude an individual’s occupational classification has a significant impact on their yearly income. However, other variables such as age and education level have a much greater impact on an individual’s yearly income.

## **Q3: Predict the income of a person in Maryland based on educational attainment and age.**

## Exploratory Analysis

The ggparis() function found in the GGally package allowed us to quickly view several different key pieces of information during our initial analysis of the data.

Throughout all of our variables, there does not appear to be any extreme outliers which will confuse our prediction model. Additionally, the ggpairs function displays the correlation for each numeric variable. While there is some correlation between WAGP, AGE, and SCHL, there does not seem to be any evidence of multicollinearity. If there was evidence of multicollinearity, we would still proceed as our analysis is focused on predicting the income of a person in Maryland.

We plotted each explanatory variable against WAGP to determine if there were any outliers. There appears to be 665 total WAGP outliers. These outliers are top coded with a maximum WAGP value at just over 500,000 USD. However, we are not concerned about these top coded outliers as there are normally high income earners throughout the work force. These top coded response variables make up less than 1% of our data source. Additionally, since these variables are top coded at ~500,000, we have reduced the amount of leverage each outlier data point has on our overall regression model.

When we examined the effect of the education and age variables on the response variable, we saw two distinct splits for both variables. Each variables showed an increasing level of income and a decreasing level of income within the same plot. For education level, we saw a flat or decreasing level of income for education levels up to level 12 (no high school degree). After this point (high school degree) we then saw a rising income level. This may lead us to develop an indicator variable for high school diploma. When we looked at the age variable, we saw a downwards trend in income level after age 50. This is most likely due to the individuals beginning to enter retirement age. This may also lead us to develop an indicator variable for late career or retirement age.

## Model Fitting/Selection

The models selected for the prediction model consisted of using both validation set approach and cross-validation method, which eventually lead to using best subset selection model. The reason for using the validation approach and cross-validation method was to evaluate the prediction accuracy of our model and dataset. The test MSE error was also computed for the predictor variables. The output from the code helps to evaluate the model because it gives us the best model that contains a certain number of variables. We’re able to see the consistency between the error rates for our variables. In addition, we used the resampling procedure to estimate the model’s performance for making a prediction for a portion of the sample.

This is done so by splitting the observations into a training and test model. The validation set error is also computed to help us understand the best model of each model size in our model. We get an idea of how the prediction variables may attribute to our intended model estimates. The effort for using the validation approach and cross-validation method eventually lead us to use the best subset selection model. The best subset selection model helps to indicate the strongest variables that can be used for any predictions that may result from our model. In a sense, our best subset selection model accounts for all our predictor variables. This model was computed by using the regsubsets() function, which helped to identify the best model that contains a given number of predictor variables. In a sense, we can compare between models to see if our predictor variables may share similar strengths and weaknesses to make predictions in our final model with our final selection of predictor variables.

## Model Diagnostics

After arriving out our model, we next checked the fit of the model. To check the model fit, we computed the residual diagnostics of the **pred\_best** model by utilizing the augment() function. This allowed us to examine the residuals against each of the predicator variables, as well as, the fitted values. The AGEP, COW, MAR, and SEX variables plotted against the residuals did not show any obvious patterns however, the SCHL variable and the fitted-values displayed linear relationships in their respective plots. This is an indication that the model selected may not be the most accurate representation of the data.

## Results

Our residuals are bounded on the lower end because our dependent variable, income, is bounded at zero. This introduces heteroskedasticity into the residuals. It also appears the residuals may have heteroskedasticity due to a poor model fit. This means that in further analysis, we will work to transform the dependent variable, and potentially other variables, to help make the residuals more homogeneous.

A histogram of our residuals shows that they are mostly centered around zero, with skew toward high residuals. Our predictions were made for four different workers:

* 36 year old, married, male, state government employee, masters: $79,180.58
* 26 year old, never married, female, state government employee, masters: $31,778.77
* 26 year old, never married, female, federal government employee, phd: $69,617.56
* 26 year old, never married, male, private company employee, phd: $71,839.48

It is interesting to note that our prediction confidence intervals are very large and include negative values of income.

## Obstacles

Question 2 required us to decode the SOCP variable in the data set. We reviewed the variable for patterns and discovered the occupational classes followed a unique pattern. Because of this pattern, we were able to transform the data to the generalized occupational classes. This transformation took a decent amount of time to understand and then code in R but it allowed us to accurately group together like jobs into generalized occupational classes.

The data contained missing values, top coded values, and bottom coded values. Depending on the question at hand certain values were omitted from the analysis to reduce the amount of leverage these outliers had on our model. For the most part we were interested in the general impact the explanatory variables had on an individual’s income. Since outliers skewed our results, they were removed.

During the initial data exploration, it was discovered educational attainment had two different effects on an individual’s income. If the individual did not graduate from high school, the number of years of education was irrelevant. If the individual graduated from high school, income increased with additional years of schooling. Since the relationship between income and educational attainment was non-linear, an indicator variable was created to add an additional coefficient to the models to account for this difference.

## Conclusion

From our analysis, we were able to determine that being employed by the federal government had the greatest effect on an individual’s earnings and that this increased wages earned by as much as 5.3% when compared to other sectors. We also found that an individual’s occupational classification has a significant impact on their yearly income however, other variables such as age and education level have a much greater impact on an individual’s yearly income.

We were also able to construct the following predictive model to estimate an individual’s earnings based on age, sex, marital status, employment sector, and level of education. Our residuals are bounded on the lower end because our dependent variable, income, is bounded at zero. Due to the presence of heteroskedasticity, in future analyses, we would spend more time investigating transformations of both the response and explanatory variables, in an effort to ensure the residuals are more homogenous.