Stat 605 Final Project

Project Group 4

November 2020

Authors

Ji, Qingchuan, qji5
Kou, Huitong, hkou2
Rui, Peibin, prui
Shi, Daiyi, dshi42
Su, Haohao, hsu69

1 Abstract

This project aim at making an exploratory data analysis on Total Person's Income (PINCP) based on the ACS PUMS 2015-2019 datasets. After variable selection and data cleaning for missing values,an annual dataset contains 70 variables and units with size range between 2300000 and 2600000. With these data, line charts are drawn to reflect the time trend of median personal income as well as its variance, and heat maps are used to show the regional imbalance of personal income in each states intuitively. Besides, regression trees are established to help find out the factors with important impact on PINCP, so that we can draw the characters of people with different levels of personal income, which implies the considerable income gap among all groups of people.

2 Data Introduction

The Public Use Microdata Sample (PUMS) contains data samples of actual response to the annual American Community Survey (ACS) run by United States Census Bureau, covering appromately 1% of the population of residents in United States. The annual dataset has two parts of records, household-level and person-level, of which we are going to focus on 2015-2019 person-level records.

These datasets cover the most common fields of daily-life of nearly 1% of American people, which can help us to find out the potential factors affecting personal income. With these factors, we are able to prove the existence of income gap and track its source.

For the details for data and variables, please check the 2019 ACS PUMS Data Dictionary.

3 Statistical Analysis

3.1 Data Pre-processing

As many other social surveys, missing data problem is unavoidable and occurs in these datasets. Besides, not all about 300 numeric and categorical variables are closely related to our aim. Thus, data pre-processing is an essential prerequisite for further statistical analysis.

Totally, 69 variables, including 13 numeric variables and 56 categorical variables, are subjectively considered to be correlated to total person's income (PINCP) and work as explanatory variables for PINCP.

For convenience, our study objects will be limited to adults (no less than 18 years old) living in 50 states and Washington DC (**in the rest part of this report, we will simply regard Washington DC as a "state"**) of the United States. This can help cut down the missing rate, since many missing values occur due to low age. Besides, some allocations have already been made by staff of U.S. Census Bureau and we will accept these allocations considering the reliability. As for the rest missing values, they all occur in categorical variables and most of them contain some information but not pure vacancy. These "NA" would be better to be assigned as an extra "class" of the corresponding factor.

Now, only the variable OC (own child) still suffers from missing data problem. We just simply remove these units and after the steps above, we have datasets containing 70 variables and 2385595-2503750 units, without any missing value.

3.2 Statistical Methods

In this project, we will use statistical tools to show how the average personal income as well as income gap grows and differs at different scales. Since the mean can be easily affect by extreme points, median is regarded as a robust measure for average personal income. Besides, variance can show how observed values of a variable scatter, so it works as measurement of personal income gap in this project. That is, large variance of PINCP indicates potential large gap of personal income.

We believe that "graphs can speak louder than words" in statistical analysis. Thus, several types of graphs would be used. Line charts for the median and variance of PINCP can show how the average personal income as well as its gap grows from 2015 to 2019.

Totally 10 heat maps are drawn for median and variance of PINCP, which reflect not only regional differences of average personal income and income gap, but also the trend of these regional differences.

Finally, regression trees are going to be built for PINCP, where the other 69 numeric and categorical variables will work as explanatory variables. These trees can help filter the important factors for personal income.

3.3 Computation

Firstly, 5 parallel jobs to do data pre-processing on five 1-year ACS PUMS datasets from 2015 to 2019, with average file size of about 2.8 GB, were submitted and ran on HTC cluster, each of which took up on average 13 minutes running time, 1.6 GB memory and 4.9 GB disk.

After the 5 data cleaning jobs mentioned above, we got 5 clean datasets, and with these clean data, another 5 parallel jobs ran to draw heat maps and build regression trees. Technically, heat maps are drawn by R with help of packages like ggplot2 and regression tree building is achieved by GUIDE. GUIDE, standing for Generalized, Unbiased, Interaction Detection and Estimation, is a multi-purpose machine learning algorithm for constructing classification and regression trees designed and maintained by Wei-Yin Loh at the University of Wisconsin, Madison. By this algorithm, we can establish and prune regression trees with cross-validation SE automatically and it has been achieved by an executable file, which can be transferred together with data to the slot for running. (You can check here for details of GUIDE). On average, each of these jobs took up nearly 2.5 hours running time, 30 GB memory and 3.3 GB disk.

3.4 Findings and Discussion

3.4.1 Time Trend

First, a line chart for median PINCP from 2015 to 2019 is shown as follows. The red line indicates the growth of median PINCP values of all USA units, while the green line and blue line are for units in ”wealthiest” and ”poorest” states of 2019 data respectively. To clarify, we simply define ”wealthy” as ”with high personal income” and ”poor” as ”with low personal income”. So, under this definition, Washington DC is the wealthiest state of 2019 while Mississippi is the poorest.

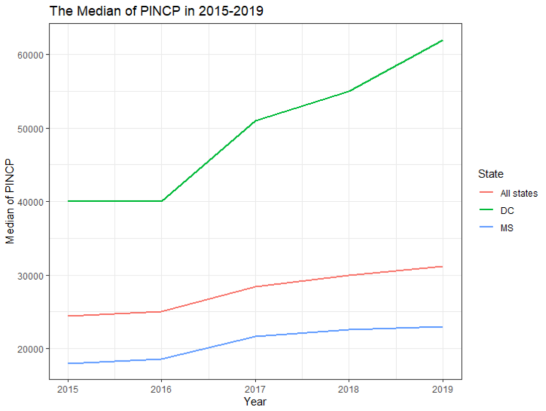


Figure 1: Time of Median Person's Income

From this line chart, we can find that the growth rate of median personal income in Washington DC surprisingly reaches more than 50%, much above the the growth rate of the whole America. In contrast, the increase of median personal income in the poorest state Mississippi is slower than which of the overall and what's worse, this increase has been slowed down since 2017, together with the increase of the whole America. This difference in median personal income increase results in the gap of these two states has expanded nearly one time in the period from 2015 to 2019.

Then, let's focus on the line chart for variance of PINCP.

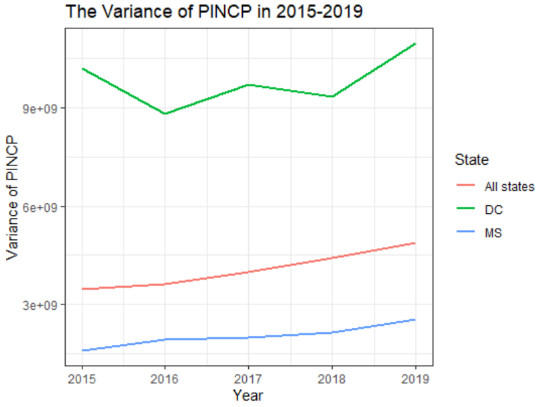
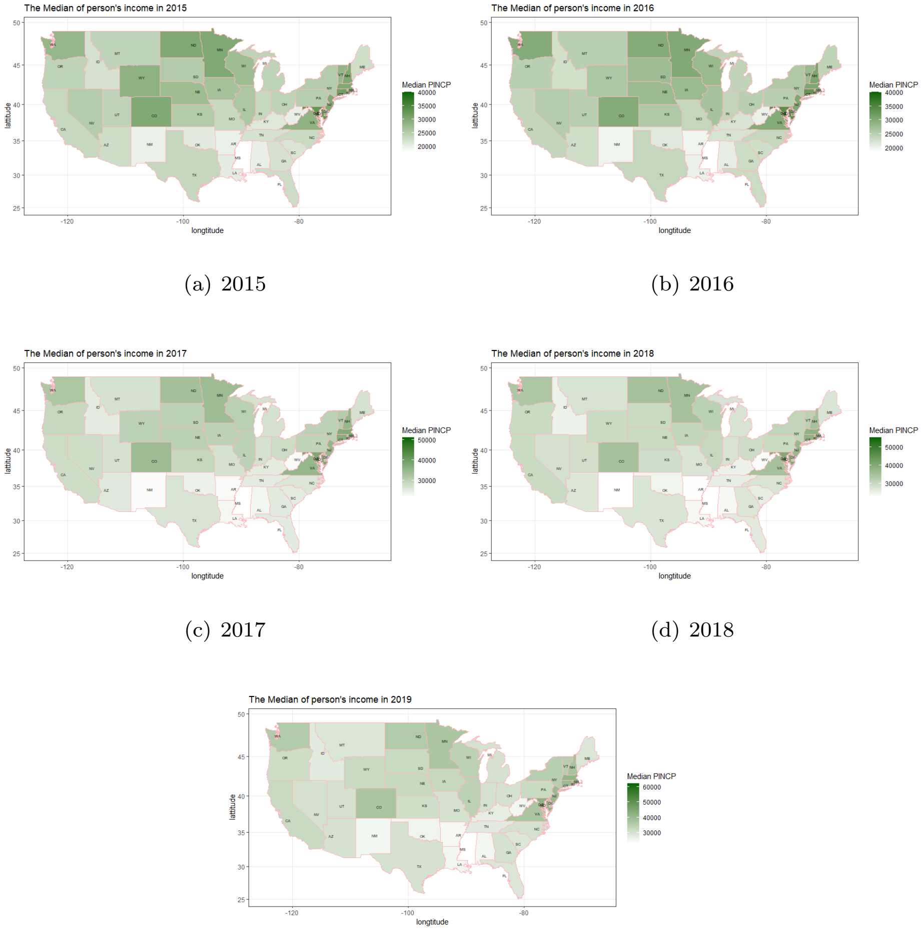


Figure 2: Time of Median Person's Income

It is obvious that the variance of personal income in DC is much greater than that in Mississippi, which implies the wealthiest state has much larger personal income gap than the poorest state. On the basis of this finding, we guess that higher median personal income may come with higher income gap.

3.4.2 Heat Maps

To dig deeper on regional difference of personal income, heat maps are drawn for median PINCP on data of each year. Here, the darkness of color is positive correlated to the median personal income, which means the state in darker green has higher median personal income.



(a) 2015

(b) 2016

(c) 2017

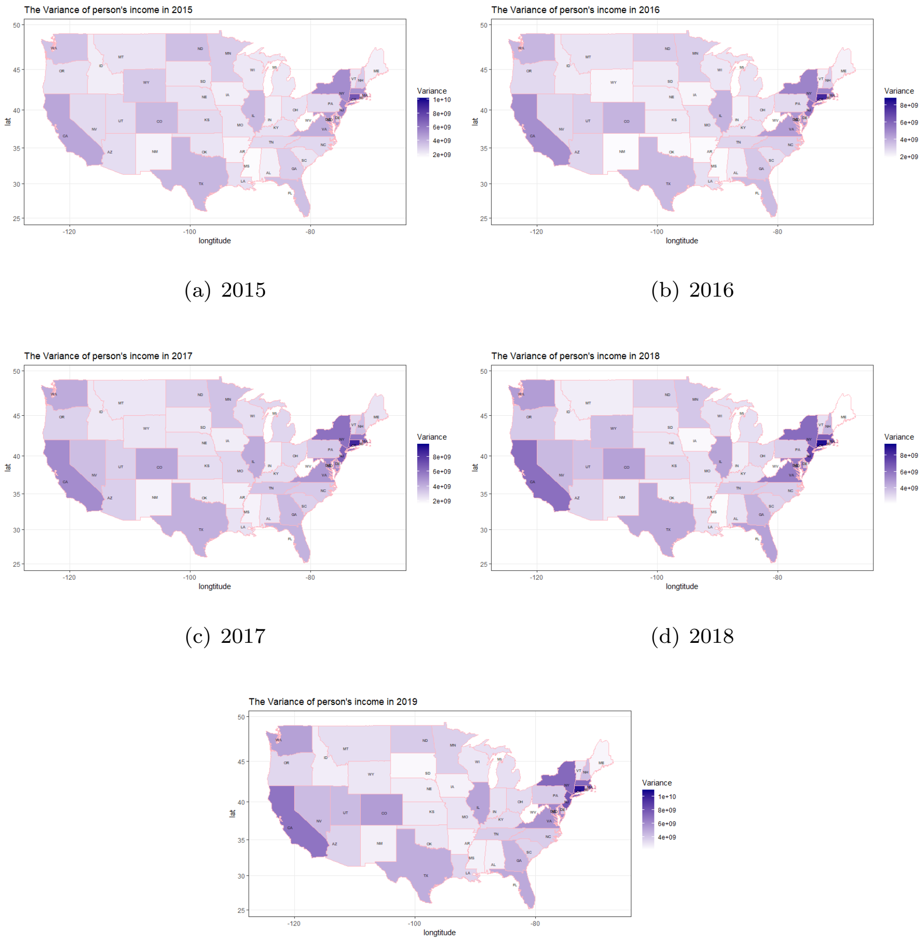
(d) 2018

(e) 2019

Figure 3: Heat Maps of Median Person's Income

Generally, the north has higher median personal income than the south; the west and east coast regions have higher median personal than the middle part. What's interesting is that the brightness contrast has gone down from 2015 to 2019. It implies that the personal income gap has reduced regionally, which surprisingly contradicts the fact from the last section that the gap between wealthiest and poorest states are becoming greater.

Then, go to the heat maps for variance of PINCP.


The Variance of person's income in 2015
lat
longtitude
Variance
1e+10
8e+09
6e+09
4e+09
2e+09
The Variance of person's income in 2016
Variance
8e+09
6e+09
4e+09
2e+09
The Variance of person's income in 2017
Variance
8e+09
6e+09
4e+09
2e+09
The Variance of person's income in 2018
Variance
8e+09
6e+09
4e+09
The Variance of person's income in 2019
Variance
1e+10
8e+09
6e+09
4e+09
-120
-100
-80
25
30
35
40
45
50

(a) 2015

(b) 2016

(c) 2017

(d) 2018

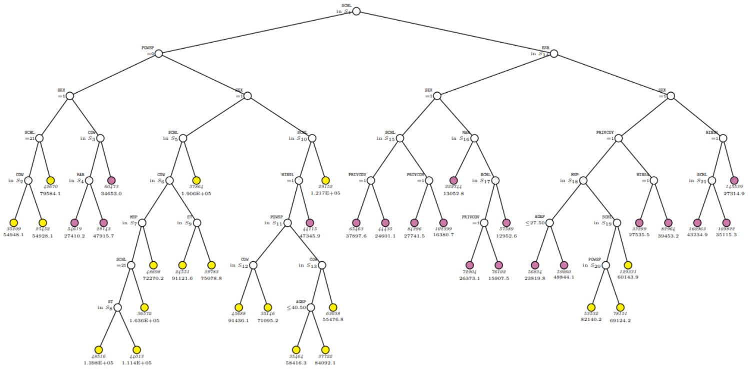
(e) 2019

Figure 4: Heat Maps of Person's Income Variance

Basically, the states with high median personal income are consistent with states with large variance of PINCP, which demonstrates our conjecture in the last section that higher median personal income may come with higher income gap. Besides, for "dark" states, their color is becoming "darker". That is, in the states with large personal income gap, their gaps are teared to be greater.

3.4.3 Regression Trees

Five regression trees are established for PINCP based on the data from 2015 to 2019. Roughly, these regression trees have the similar structure and split variables. And mostly, they are all very large even after pruning by cross-validation. Therefore, we are going to mainly show and discuss the regression tree on 2019 data and **for the clear images of all five trees, please turn to the Appendix**.



GUIDE v.36.1 0.50-SE piecewise linear least-squares regression tree with stepwise variable selection for predicting PINCP. Tree constructed with 2452515 observations. Maximum number of split levels is 30 and minimum node sample size is 24525. At each split, an observation goes to the left branch if and only if the condition is satisfied. Set S_1 = {21, 22, 23, 24}. Set S_2 = {0, 2, 8, 9}. Set S_3 = {0, 8, 9}. Set S_4 = {1, 4}. Set S_5 = {21, 22}. Set S_6 = {1, 5, 7}. Set S_7 = {1, 3, 4}. Set S_8 = {11, 13, 17, 24, 25, 34, 36, 48, 51, 53, 6, 8, 9}. Set S_9 = {10, 11, 2, 24, 25, 32, 34, 36, 44, 51, 53, 6, 9}. Set S_{10} = {21, 22}. Set S_{11} = {11, 24, 25, 251, 303, 34, 36, 44, 51, 53, 555, 6, 72, 9}. Set S_{12} = {1, 5, 7, 8}. Set S_{13} = {1, 5, 7}. Set S_{14} = {3, 6}. Set S_{15} = {18, 19, 20}. Set S_{16} = {1, 4}. Set S_{17} = {16, 18, 19, 20}. Set S_{18} = {2, 6}. Set S_{19} = {18, 19, 20}. Set S_{20} = {10, 11, 2, 22, 24, 25, 251, 254, 301, 32, 33, 34, 36, 399, 44, 48, 51, 53, 555, 56, 6, 72, 8, 9}. Set S_{21} = {18, 19, 2, 20}. Sample size (*in italics*) and mean of PINCP printed below nodes. Terminal nodes with means above and below value of 4.955E+04 at root node are colored yellow and purple, respectively. Second best split variable at root node is SCIENGP.

Figure 5: 2019 PINCP Regression Tree

This is a piecewise linear least-squares regression tree with stepwise variable selection for predicting PINCP. At each split of the tree, an observation goes to the left branch if and only if the condition is satisfied. The following are considered as top split variables for PINCP.

Variable	Type	Meaning
SCHL	Categorical with 24 classes	Educational attainment
POWSP	Categorical with 60 classes	Place of work
ESR	Categorical with 6 classes	Employment status
SEX	Categorical with 2 classes	Sex
COW	Categorical with 10 classes	Class of worker
MAR	Categorical with 5 classes	Marital status
AGEP	Numeric	Age
MSP	Categorical with 6 classes	Married, spouse present/spouse absent
PRICOV	Categorical with 2 classes	Private health insurance coverage

Among these, SCHL, standing for Educational attainment, works as the first split variable by dividing all units into two groups, those who have at least one bachelor's degree or above and those who do not. Generally, people from the former group have personal income higher than people from the latter on average, consistent with our common sense in education and indicating that education does help improve personal income.

There is another fact relating education and personal income, which however is not shown on this tree, that is the second best split variable at root node is SCIENGP. SCIENGP is a binary categorical variable for people with at least one bachelor's degree or above (NA for those who do not) reflecting whether the degree of the unit is in the field of science and engineering. And people with at least one degree in the field of science and engineering can have higher average personal income.

According to the regression tree, there exist some other interesting and meaningful findings, like the factor SEX. we can imply from the tree that males are more likely to have high personal income than females, implying the social phenomenon of gender inequality; people with private health insurance have higher average personal income than those without one, which may actually lead to a reversed casual relationship that people with higher personal income are inclined to have private health insurance; divorced or widowed women at work may earn more personal income than married women at work.

4 Conclusion

From the exploratory analysis on PINCP variable from ACS PUMS data, it is clear to found many facts related to personal income gaps on region, gender, education background and so on.

Regionally, higher median personal income may still come with higher income gap, suggesting the imbalance of development, which should be taken seriously;

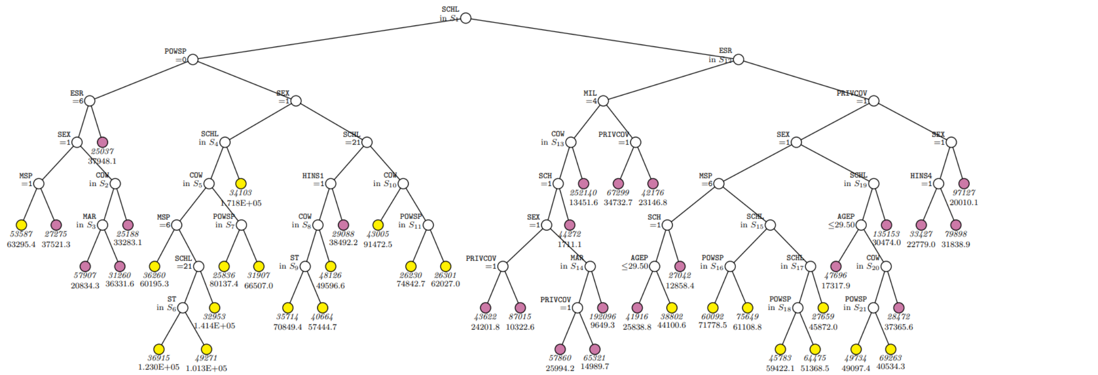
The problem of gender inequality is still severe based the the gap of personal income;

And among all factors found and considered to have impact on person's income, education may always be the first one with decisive effect;

......

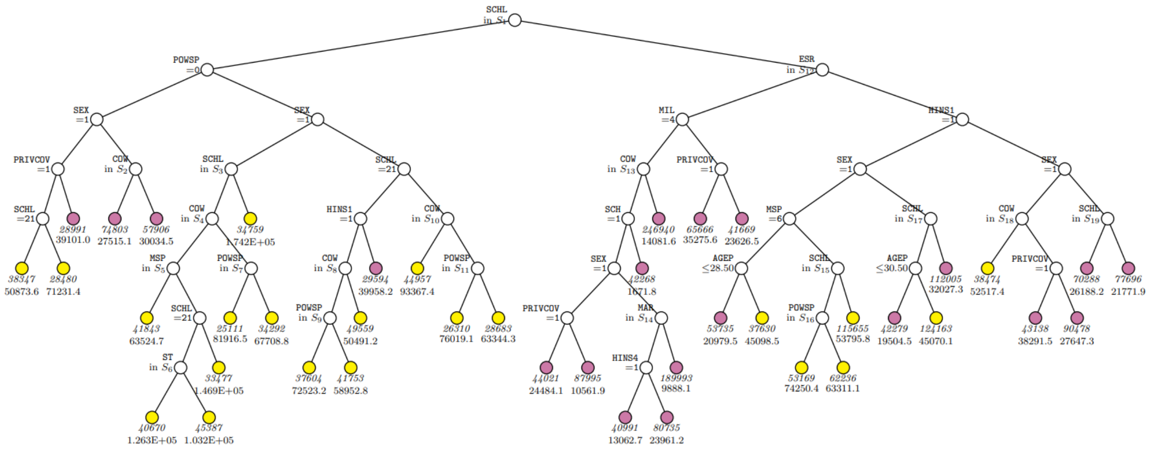
From these findings, we can learn much to reexamine the social development.

Appendix



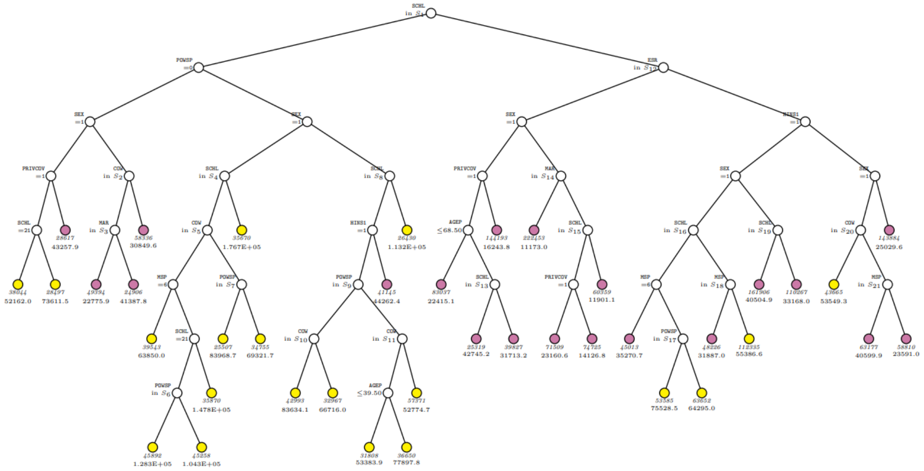
GUIDE v.36.1 0.50-SE piecewise linear least-squares regression tree with stepwise variable selection for predicting PINCP. Tree constructed with 2490616 observations. Maximum number of split levels is 30 and minimum node sample size is 24906. At each split, an observation goes to the left branch if and only if the condition is satisfied. Set S_1 = {21, 22, 23, 24}. Set S_2 = {0, 1, 8}. Set S_3 = {1, 4}. Set S_4 = {21, 22}. Set S_5 = {1, 5, 7}. Set S_6 = {11, 24, 25, 33, 34, 36, 48, 51, 53, 6, 9}. Set S_7 = {10, 11, 17, 2, 24, 25, 254, 301, 303, 32, 34, 36, 399, 44, 51, 53, 6, 8, 9}. Set S_8 = {1, 5, 7}. Set S_9 = {11, 2, 24, 25, 34, 36, 48, 51, 53, 6, 8, 9}. Set S_{10} = {1, 5, 7}. Set S_{11} = {10, 11, 17, 2, 24, 25, 254, 26, 27, 34, 36, 399, 42, 44, 53, 555, 6, 9}. Set S_{12} = {3, 6}. Set S_{13} = {0, 8, 9}. Set S_{14} = {2, 3}. Set S_{15} = {18, 19, 20}. Set S_{16} = {10, 11, 166, 17, 2, 22, 24, 25, 251, 254, 301, 303, 34, 36, 38, 399, 48, 51, 53, 555, 56, 6, 72, 8, 9}. Set S_{17} = {16, 17}. Set S_{18} = {10, 11, 17, 2, 22, 24, 25, 254, 27, 30, 301, 303, 34, 36, 38, 399, 48, 49, 53, 54, 555, 56, 6, 8, 9}. Set S_{19} = {18, 19, 2, 20, 3}. Set S_{20} = {1, 2, 5, 7}. Set S_{21} = {10, 11, 15, 166, 2, 24, 25, 251, 254, 33, 34, 36, 4, 41, 44, 48, 50, 51, 53, 555, 6, 8, 9}. Sample size (*in italics*) and mean of PINCP printed below nodes. Terminal nodes with means above and below value of 4.008E+04 at root node are colored yellow and purple, respectively. Second best split variable at root node is SCIENGP.

Figure 6: 2015 PINCP Regression Tree



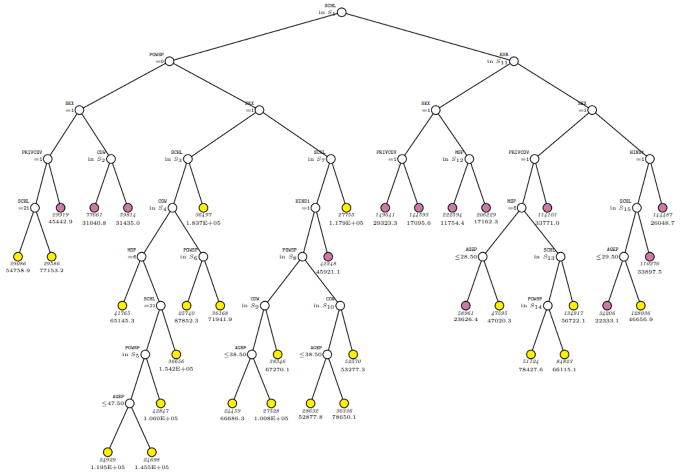
GUIDE v.36.1 0.50-SE piecewise linear least-squares regression tree with stepwise variable selection for predicting `PINCP`. Tree constructed with 2503750 observations. Maximum number of split levels is 30 and minimum node sample size is 25037. At each split, an observation goes to the left branch if and only if the condition is satisfied. Set $S_1 = \{21, 22, 23, 24\}$. Set $S_2 = \{0, 8, 9\}$. Set $S_3 = \{21, 22\}$. Set $S_4 = \{1, 5, 7\}$. Set $S_5 = \{2, 6\}$. Set $S_6 = \{10, 11, 17, 24, 25, 33, 34, 36, 48, 51, 53, 6, 9\}$. Set $S_7 = \{11, 166, 17, 2, 24, 25, 301, 32, 34, 36, 399, 44, 51, 53, 555, 6, 9\}$. Set $S_8 = \{1, 5, 7\}$. Set $S_9 = \{10, 11, 166, 24, 25, 254, 303, 34, 35, 36, 399, 48, 51, 53, 6, 8, 9\}$. Set $S_{10} = \{1, 5, 7\}$. Set $S_{11} = \{10, 11, 15, 17, 2, 24, 25, 251, 254, 27, 301, 34, 36, 399, 42, 44, 53, 555, 6, 72, 9\}$. Set $S_{12} = \{3, 6\}$. Set $S_{13} = \{0, 8, 9\}$. Set $S_{14} = \{2, 3\}$. Set $S_{15} = \{18, 19, 20\}$. Set $S_{16} = \{10, 11, 15, 166, 17, 2, 22, 24, 25, 251, 254, 301, 32, 34, 36, 38, 399, 44, 48, 51, 53, 555, 56, 6, 8, 9\}$. Set $S_{17} = \{18, 19, 2, 20\}$. Set $S_{18} = \{6, 7\}$. Set $S_{19} = \{18, 19, 20\}$. Sample size (*in italics*) and mean of `PINCP` printed below nodes. Terminal nodes with means above and below value of 4.133E+04 at root node are colored yellow and purple, respectively. Second best split variable at root node is `SCIENGP`.

Figure 7: 2016 PINCP Regression Tree



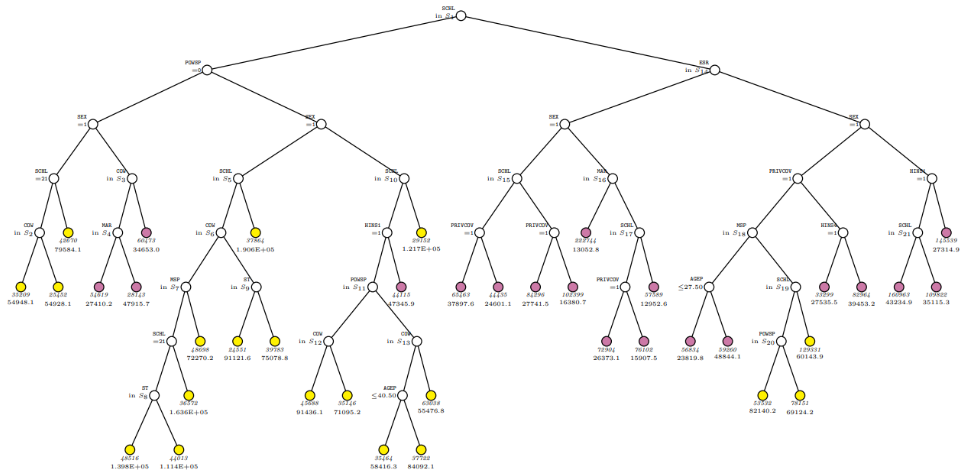
GUIDE v.36.1 0.50-SE piecewise linear least-squares regression tree with stepwise variable selection for predicting PINCP. Tree constructed with 2385595 observations. Maximum number of split levels is 30 and minimum node sample size is 23855. At each split, an observation goes to the left branch if and only if the condition is satisfied. Set $S_1 = \{21, 22, 23, 24\}$. Set $S_2 = \{0, 8, 9\}$. Set $S_3 = \{1, 4\}$. Set $S_4 = \{21, 22\}$. Set $S_5 = \{1, 5, 7\}$. Set $S_6 = \{11, 13, 166, 17, 24, 25, 251, 254, 303, 34, 36, 399, 48, 51, 53, 555, 6, 72, 9\}$. Set $S_7 = \{10, 11, 15, 17, 2, 24, 25, 254, 301, 34, 36, 399, 44, 51, 53, 6, 9\}$. Set $S_8 = \{21, 22\}$. Set $S_9 = \{10, 11, 166, 2, 24, 25, 251, 254, 301, 34, 36, 399, 44, 51, 53, 555, 6, 9\}$. Set $S_{10} = \{1, 5, 7\}$. Set $S_{11} = \{1, 5, 7\}$. Set $S_{12} = \{3, 6\}$. Set $S_{13} = \{18, 19, 20\}$. Set $S_{14} = \{1, 4\}$. Set $S_{15} = \{16, 18, 19, 20\}$. Set $S_{16} = \{18, 19, 20\}$. Set $S_{17} = \{11, 15, 166, 17, 2, 22, 24, 25, 251, 254, 301, 303, 32, 34, 36, 38, 399, 44, 48, 51, 53, 555, 56, 6, 8, 9\}$. Set $S_{18} = \{2, 6\}$. Set $S_{19} = \{18, 19, 2, 20\}$. Set $S_{20} = \{5, 6, 7\}$. Set $S_{21} = \{1, 3, 4\}$. Sample size (*in italics*) and mean of PINCP printed below nodes. Terminal nodes with means above and below value of 4.477E+04 at root node are colored yellow and purple, respectively. Second best split variable at root node is SCIENGP.

Figure 8: 2017 PINCP Regression Tree



GUIDE v.36.1 0.50-SE piecewise linear least-squares regression tree with stepwise variable selection for predicting PINCP. Tree constructed with 2414823 observations. Maximum number of split levels is 30 and minimum node sample size is 24148. At each split, an observation goes to the left branch if and only if the condition is satisfied. Set S_1 = {21, 22, 23, 24}. Set S_2 = {0, 8, 9}. Set S_3 = {21, 22}. Set S_4 = {1, 5, 7}. Set S_5 = {10, 11, 13, 166, 17, 24, 25, 251, 254, 301, 33, 34, 36, 399, 44, 48, 51, 53, 555, 6, 8, 9}. Set S_6 = {11, 166, 17, 2, 24, 25, 251, 254, 301, 303, 34, 36, 399, 44, 51, 53, 555, 6, 72, 9}. Set S_7 = {21, 22}. Set S_8 = {10, 11, 166, 17, 2, 24, 25, 251, 254, 301, 34, 36, 41, 44, 51, 53, 555, 6, 72, 9}. Set S_9 = {1, 5, 7}. Set S_{10} = {1, 5, 7}. Set S_{11} = {3, 6}. Set S_{12} = {1, 2, 5}. Set S_{13} = {18, 19, 20}. Set S_{14} = {11, 166, 2, 22, 24, 25, 251, 254, 301, 303, 34, 36, 38, 44, 48, 51, 53, 555, 56, 6, 72, 9}. Set S_{15} = {18, 19, 2, 20}. Sample size (*in italics*) and mean of PINCP printed below nodes. Terminal nodes with means above and below value of 4.659E+04 at root node are colored yellow and purple, respectively. Second best split variable at root node is SCIENGP.

Figure 9: 2018 PINCP Regression Tree



GUIDE v.36.1 0.50-SE piecewise linear least-squares regression tree with stepwise variable selection for predicting PINCP. Tree constructed with 2452515 observations. Maximum number of split levels is 30 and minimum node sample size is 24525. At each split, an observation goes to the left branch if and only if the condition is satisfied. Set $S_1 = \{21, 22, 23, 24\}$. Set $S_2 = \{0, 2, 8, 9\}$. Set $S_3 = \{0, 8, 9\}$. Set $S_4 = \{1, 4\}$. Set $S_5 = \{21, 22\}$. Set $S_6 = \{1, 5, 7\}$. Set $S_7 = \{1, 3, 4\}$. Set $S_8 = \{11, 13, 17, 24, 25, 34, 36, 48, 51, 53, 6, 8, 9\}$. Set $S_9 = \{10, 11, 2, 24, 25, 32, 34, 36, 44, 51, 53, 6, 9\}$. Set $S_{10} = \{21, 22\}$. Set $S_{11} = \{11, 24, 25, 251, 303, 34, 36, 44, 51, 53, 555, 6, 72, 9\}$. Set $S_{12} = \{1, 5, 7, 8\}$. Set $S_{13} = \{1, 5, 7\}$. Set $S_{14} = \{3, 6\}$. Set $S_{15} = \{18, 19, 20\}$. Set $S_{16} = \{1, 4\}$. Set $S_{17} = \{16, 18, 19, 20\}$. Set $S_{18} = \{2, 6\}$. Set $S_{19} = \{18, 19, 20\}$. Set $S_{20} = \{10, 11, 2, 22, 24, 25, 251, 254, 301, 32, 33, 34, 36, 399, 44, 48, 51, 53, 555, 56, 6, 72, 8, 9\}$. Set $S_{21} = \{18, 19, 2, 20\}$. Sample size (*in italics*) and mean of PINCP printed below nodes. Terminal nodes with means above and below value of 4.955E+04 at root node are colored yellow and purple, respectively. Second best split variable at root node is SCIENGP.

Figure 10: 2019 PINCP Regression Tree