A Take on Generalized Spatial Mismatch Hypothesis: Effects of Residency Away From Work on Income Inequality

Brian Kang

Department of Economics
University of Washington*
Seattle, WA 98105, USA
kangb98@u.washington.edu

May 22, 2020

Abstract

The Spatial Mismatch Hypothesis suggests that disadvantaged minorities in the labor market are physically disconnected from locations of abundant and high-demand job opportunities. This indicates aggravated income inequality between the rich and poor, especially today when suburbanization is most evident than ever in U.S. history. Additionally, sociodemographic diversity of the current labor population adds another layer of complexity to examining the factors of such economic disparity.

This paper aims to contribute to the literature of the Spatial Mismatch Hypothesis by investigating the potential effects of commuting characteristics on variation of wealth. Machine learning and econometrics are leveraged on data collected from the U.S. Census to quantify this relationship and feature select probable determinants of wealth. Analytic methods used in this research process include double LASSO regularization, Breusch-Pagan test for heteroskedasticity, and false discovery rate control using the Benjamini-Yekutieli procedure.

*I would like to thank Professor Gregory Duncan, my faculty advisor, for supporting and guiding me through this one-of-a-kind empirical research opportunity. Researching under your supervision encouraged to continuously challenge myself with high expectations. I could not have asked for a better learning experience.

Also, thank you Professor Alan Griffith and Professor Michelle Turnovsky for providing the opportunity and resources to participate and succeed in this honors sequence; Professor Dong-Jae Eun for assistance in modeling; Kathy (Heejung) Jung for motivating and upskilling R coding and ML knowledge; Thor Dodson for econometric consulting.

Contents

1 Introduction

How does society structurally respond to rapid growth? Observe rising cities in the Pacific Northwest, notably Seattle, WA and Portland, OR. During the past decade, their boom in technological and digital engineering (HiTech) accompanied a rapid population growth and local economic success, placing these urban cities on list of international hubs for development and investment.[1] As much as the robustness of these locations attract business and population, these MSAs are experiencing difficulties in resolving the negative spillovers caused by a dramatic increase in demand, and nearby suburban communities like Redmond and Spokane, WA are also affected by the repercussions of scarcity (Building Solutions 2017). Housing prices nearly doubled since 2012 (FHFA 2019). The homelessness crisis worsened (Greenstone 2019). Traffic congestion aggravated (Adam Millsap 2018). Then, rising MSAs like Seattle and Portland pooled money to these areas in order to increase investments for large scale public transportation and infrastructure construction projects. Hence, many who could not afford living close to MSAs or those who prefer to live away from large metropolitan cities are naturally crowded out from the HiTech hubs. On the other hand, those who can afford and have incentives to take advantage of the high amenities available in concentrated urban areas may crowd-in.

In light of this unique modern urbanization movement called 'suburbanization,' which was unseen during the industrial ages, I tackle the question *"What is the quantitative effect of residency away from workplace on income inequality?"* More commonly known as the Spatial Mismatch Hypothesis, it is "serious limitations on black residential choice, combined with the steady dispersal of jobs from central cities, are responsible for the low rates of employment and low earnings of Afro-American workers" (Kain 1994). This term has now been more generalized in terms of a mismatch between location of people's residencies and the location of job opportunities, and this mismatch can be seen as a repercussion of rapid urban development. From anecdotal experience, many parents have unwantedly been separated from family, unable to financially support the entire household. Many struggle to make ends meet even after working two full-time jobs, while some others indulge in extravagance and explosive wealth (Finio and Mishel 2013). But at who's expense? This economic unfairness not only undermines political stability, but also the economic system itself and the basic livelihoods of citizens (Piketty 2014). So, as much as the Spatial Mismatch Hypothesis is significant to assessing the current condition of financial inequality in our society, it is unfortunate to acknowledge that there is a lack of empirical research done on this issue.

This paper is an attempt, following a minority of researchers, to bring light into a small portion of the larger problem of widespread economic inequality and reaction to change. The

[1]The U.S. Office of Management and Budget defines these core locations as Metropolitan Statistical Area (MSA), "a core area containing a substantial population nucleus, together with adjacent communities having a high degree of economic and social integration with that core (Census 2010)."

approach I take to bring about an answer to my question is using algorithmic measures to deduce the determinants of wealth. Utilizing machine learning will return promising results as what societal factors may be related to variations in wealth. Following is further econometric analysis because machine generated results lack power and trust in interpretation and inference. What this paper does not aim to capture is the analysis and inference for reverse causality. As much as how physical distance between residency and labor may alter wealth, the outcome for wealth determining residential and labor location preferences is highly likely. My analysis limited to the previous case will be motivational for future attempts to answer the latter.

2 Literature Review

2.1 Income Inequality

With a more widely available and efficient transportation system than every before, it is logical to argue that the significance of physical proximity between residency and employment and difficulty in accessing nearby public amenities diminishes because advanced transportation eliminates the high stakes and costs for living far from MSAs. However, accordingly with the Spatial Mismatch Hypothesis, this is not true. This trend comes from various economic, political, and social factors of MSAs that are not clear at first sight. Notably, income distribution among the metropolitan population is directly affected by scale and rate of urban development, caused by factors such as robustness of business activities and population migration (Sanchez 2002)[2]. Related to the separation of wage is widening human talent per physical location. Unskilled and less specialized labor move to MSAs for opportunities to grow income only to be compensated with low wages, while the upper class that has skilled and/or specialized talent earn the income and choice to stay in either MSAs or emigrate outside (Sanchez 2002). Again, this is the spatial mismatch hypothesis in action.

To further expand on the social factors of MSA income inequality, both Sanchez and Frumkin 2002 describe a potential explanatory factor for income inequality that cannot be overshadowed by economic activity and politics: racial and ethic discrimination. Sanchez states that, "discrimination has direct and indirect implications on labour conditions in terms of skill/education levels, job opportunities and potential for advancement." Frumkin highlights that the groups that are exposed to such discrimination are more likely exposed to costly health and environment related issues, such as, air pollution, water quality, and mental illness. These groups are also more vulnerable to injuries from increased transportation vehicles. Do note that for a non-negligible number of Americans, health insurance is a luxury. According to Dubay, Holahan, and Allison 2006, in 2005, the Current Population Survey (CPS) showed that 46.1 million (and rising) non-elderly Americans were uninsured. This is

[2]Construction Spending on transportation infrastructure is at a all-time high (Census 2019b).

an addressable issue and will be included in our econometric model to observe its strength of correlation to income disparity.

In addition to apparent wage inequality and sociodemographic disparity, residential segregation is clear as well. In locations where public amenities are abundant, housing development often display specific price ranges. This leads to economic strata becoming clear between those who can afford some houses and those who cannot afford. Oftentimes, for some members the former group, this phenomenon is a profitable economic opportunity. But, for the latter, housing is an large economic burden as it takes up 30% - 40% of the average American's monthly income according to the U.S. Bureau of Labor Statistics 2020. Naturally, they are pushed out of certain neighborhoods defined by spatial stratification. This development pattern intensifies physical location based income inequality within MSAs (Frumkin 2002).

Durlauf 1992 concisely summarizes a key takeaway the physical characteristic of the Spatial Mismatch Theorem,

> "When redistribution from rich to poor in the urban center is large enough, wealthy agents (families) have an incentive to abandon the community and form their own neighborhoods... Stratification and the breaking up of the urban center are functions of realized income distribution."

Here, it is appropriate to address the "Inequality and the Measurement of Residential Segregation by Income in American Neighborhoods" paper which poses an interesting point to my question. It shows that income inequality and location-based stratification is indeed not an ideal societal outcome, but it addresses that economic segregation across neighborhoods is important. A lack of exposure to middle-class "role models" leads to urban unemployment and social issues (Watson 2009). This social behavior branches from an ambiguity for allocating public good and irregular commuting behavior for all income levels. Also, Watson shares,

> "Residential decisions (made by the individuals in segregated economies) have implications for commuting behavior and the allocation of public goods. If residential choice is sensitive to income distribution, economic policies that moderate or amplify income inequality may shape the cities in which we live."

She then quantifies this residential decision making by introducing a measure of income sorting, Centile Gap Index (CGI), based on income distributions of families in 216 U.S. cities, concluding that inequality can fully explain the rise in income segregation, a promising statement for my question.

2.2 Commuting and Residency

We move onto the topic of migration. Although migration behaviors are not the focus of this paper, because it is tightly intertwined with residency preferences and commuting

behaviors to work, it is discussed here. Population redistribution and residential preferences are widely studied topics, and a concentration on transportation economics has been a popular common interest for many researchers. For our purposes, let us separate relocation into rural-urban migration, intra-urban migration, and somewhere in between, including non-MSA cities adjacent to the MSAs and remote non-MSA counties.

Those who prefer non-MSA locations, but not as far as rural, accept the tradeoff between living in smaller cities versus compactly located benefits offerable in urban locations like work and public transportation, and the U.S. Census shows that the growth for these preferences have finally surpassed the demand for MSA locations in 1970 and continues to grow (Jong, F., and Sell 1977). They recognize that this suburbanization behavior is partially derived by spatial differentiation of population and land-use patterns following industrialization, so the spillover from MSAs naturally flow to non-MSAs that maintain some physical proximity while taking advantage of evolving transportation routes. This allows location-based industry specialization and encourages the physical division between work and family, which contributes to making migration decisions despite potentially longer commute to work.

On the other hand, as much as access to mass transit within MSAs are available today, it is as convenient to relocate within the MSAs. Assume that MSAs by definition are the core hubs that contain the most specialized and highest demand jobs. Brown and Moore 1970 show that the closer the proximity to work, that is probably shorter commutes, and the more access to public transportation is available, the higher income people earn in MSAs. This may imply an huge disadvantage for those with low demand jobs. They also note that metropolitan citizens are very sensitive to environment changes such as relocation of industrial sites, change in racial composition in neighborhood, and update in transportation technology. We extrapolate the fact that continuous external stimuli on individuals in MSAs constantly change the population behaviors because socioeconomic development or demographic change imply the requirement for individuals to quickly adapt. I hypothesize that this in essence is a setup for high barrier of entry for those who already reside outside MSAs. Not only is the labor market competitive but also this constant societal change may not be suited for those who are likely not accustomed to it, a potential factor for preference living away from MSAs.

We now discuss rural-urban migration occur in accordance to transportation policies. Zenou 2010 not only discusses transportation policies that improve the public transport system in cities, but also introduces the idea of an entry-cost imposing policy that "encourages investment in the city" and a migration-restricting policy that "imposes costs on migrants." This entry cost is similar to the notion of high barrier of entry above. The key takeaway from this paper is that policies that improve transportation across long distances increase employment. Also, it decreases large commuting costs, which increases rent prices and decreases urban wages. This may discourage immigration from rural to urban locations. However, this does not necessarily imply that investments in public transportation discourages immigra-

tion because the decrease in commuting costs cannot entirely be explained by transportation policies.

3 Research Methodology

3.1 Data

The cross sectional dataset is individual and household level data collected from 1-Year American Community Survey (ACS)[3], 5-Year Public Use Microdata Sample (PUMS)[4], and Current Population Survey (CPS)[5] compiled in year 2018. All are subsidiaries under the greater U.S. Census Bureau and they provide very specific information such as household poverty thresholds, income, location of residence, travel time to work, place of work, nearest history of migration, type of insurance, occupation type, race, etc.

The advantage of using low level micro data over larger macro level data is the ability to question and answer how specific decision makings, characteristics, and preferences of people or households affect a larger entity as a whole, e.g. "impact of residential preference on population dispersal migration (Jong, F., and Sell 1977)". Hence, instead of the Gini index, a widely used indicator for income inequality that is considered macro level data, we use the Income-Poverty Ratio as a proxy for the combinatorial effect of wealth and poverty, popularly practiced by the U.S. Census Bureau. (We will denote the Income-Poverty Ratio as Income/Poverty or IncPovRatio.) We will shift this ratio and adjust the values using logarithmic scale to normalize the data and make sure that the values are fit for analyses (negative IncPovRatio is present, i.e., household income is less than the poverty threshold). Refer to Figure 1. From the natural values of IncPovRatio, the skew is extremely right, evidence for the presence of income inequality.

$$log(IncPovRatio) = log((Income + Earnings)/PovertyThreshold)$$

The CPS provides specific guidelines on what types of money flow composites income and earnings. Poverty threshold varies by size of household and age of the members. This threshold is used throughout the U.S. and are updated annually for inflation using the Consumer Price Index for All Urban Consumers (CPI-U). The Census Bureau notes that "although the thresholds in some sense reflect a family's needs, they are intended for use as a statistical yardstick, not as a complete description of what people and families need to live" (Census 2019). Refer to Figure 2.

[3]Dataset downloaded from `https://www2.census.gov/programs-surveys/acs/data/pums/2018/1-Year/`

[4]Variable dictionary downloaded from `https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2018.txt`

[5]Poverty Threshold dataset downloaded from `https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html`

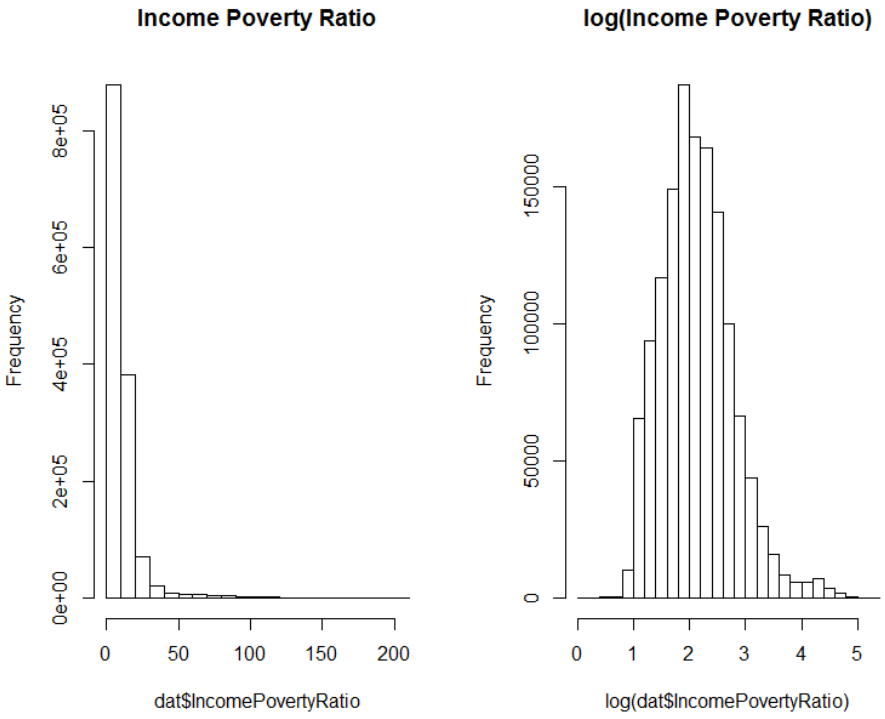


Figure 1: Income Poverty Ratio in two ways

Poverty Thresholds for 2018 by Size of Family and Number of Related Children Under 18 Years

Size of family unit	Weighted average thresholds	Related children under 18 years								
		None	One	Two	Three	Four	Five	Six	Seven	Eight or more
One person (unrelated individual):	12,784									
Under age 65	13,064	13,064								
Aged 65 and older	12,043	12,043								
Two people:	16,247									
Householder under age 65	16,889	16,815	17,308							
Householder aged 65 and older	15,193	15,178	17,242							
Three people	19,985	19,642	20,212	20,231						
Four people	25,701	25,900	26,324	25,465	25,554					
Five people	30,459	31,234	31,689	30,718	29,967	29,509				
Six people	34,533	35,925	36,068	35,324	34,612	33,553	32,925			
Seven people	39,194	41,336	41,594	40,705	40,085	38,929	37,581	36,102		
Eight people	43,602	46,231	46,640	45,800	45,064	44,021	42,696	41,317	40,967	
Nine people or more	51,393	55,613	55,883	55,140	54,516	53,491	52,082	50,807	50,491	48,546

Source: U.S. Census Bureau.

Figure 2: 2018 Poverty Threshold Table

The dataset is merged data from person (P) level and household (H) data and contains 3,214,539 observations and 286 variables. After omitting NA values, deleting multicollinear variables, removing unwanted levels or variables, and creating necessary dummies, we are left with 976,712 observations and 135 variables. Much of the dropped values and unwanted levels come from survey data's self reported nature. Various other scatterplots and histograms to understand some of our variables of interest are in Appendix A.1.

3.2 Modeling

The econometric model I used is the following,

$$\begin{aligned} log(IncomePovertyRatio) = \beta_0 + \beta_1 TravelTimeToWork \text{ (JWMNP)} + \beta_2 SameResidenceWorkplace \\ + \beta_3(TravelTimeToWork \text{ (JWMNP)} * SameResidenceWorkplace) \\ + \beta_4 TransitMeanToWork \text{ (JWTR)} \\ + \beta_5(TravelTimeToWork \text{ (JWMNP)} * TransitMeanToWork \text{ (JWTR)}) \\ + \vec{\beta} SocioDemographicCovariates \end{aligned}$$

The control feature variables of interest in this paper are: travel time to work (JWMNP), the deterministic variable in place for distance between residency and work place; SameResidenceWorkplace, a binary variable indicating if an individual's residency is equal/close to work place or not; transit mean to work (JWTR); and the interactions of SameResidency-Workplace and JWTR with JWMNP. The response variable is log(IncPovRatio). Refer to the below table for specifics on the control variables.

I first hypothesize that the longer the travel time to work, the lower the IncPovRatio. Brown and Moore 1970 showed that there is evidence that the closer located to work, the higher income. Related to this evidence, if the residence and workplace are equal, I expect IncPovRatio to be higher for an individual on average. Do note that for this variable, the sample size of people with different locations is relatively small compared to its counterpart (42,925 v.s. 933,787). For means of transportation, I hypothesize that private or commercialized shareable vehicles (cars, motorcycles, taxi, and bicycles) are positively correlated to IncPovRatio while the rest are negatively correlated. While the former maybe a possession or relatively pricey commuting method, the rest are mostly public transportation, a cheaper alternative for transportation widely accessible to people. Keeping these variables and its interactions as control, including other potential socidemographic covariates that can help explain the variation in IncPovRatio will return results on which are the significant determinants of wealth.[6]

[6]For the descriptions of all sociodemographic covariates, please refer to the U.S. Census variable dictionary `https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2018.txt`

Control Variables	**Description**
Travel Time to Work (JWMNP)	Numeric: 1 - 200 minutes.
SameResidenceWorkplace	Binary: 0 if residence and work location are far. 1 if close or equal.
Means of Transportation to Work (JWTR)	Factor: **01** Car, truck, or van; **02** Bus or trolley bus; **03** Streetcar or trolley car (carro publico in Puerto Rico); **04** Subway or elevated; **05** Railroad; **06** Ferryboat; **07** Taxicab; **08** Motorcycle; **09** Bicycle; **10** Walked; **11** Worked at home; **12** Other method

3.2.1 Double ML

Before introducing the algorithmic feature selection method I use, I first introduce the LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani 1996). This shrinkage model uses a L1 regularization hyperparameter that penalizes the coefficients $\hat{\beta}$ of a regression and "shrinks" them towards zero. At the end, some coefficients are equal to zero while others are not. By encouraging an approximately sparse solution like this, not only is mechanical computation sped up, but also the independent variables that should matter most in predicting the dependant variable are left (those with non-zero coefficients). This is how the LASSO does variable selection and prevents overfitting. This method not only works well in high-dimensional spaces with possibly more features than observations, but also has the potential to improve model prediction error by biasing the model a little with regularization but drastically decreasing variance.

Given data points $(x_1, y_1), \ldots, (x_n, y_n)$ the objective function of the LASSO is,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^k, \beta_0 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i^T \beta + \beta_0 - y_i)^2 + \lambda \sum_{j=1}^{k} |\beta_j|$$

which is the minimization of regularized squared error loss $l_i = (x_i^T \beta + \beta_0 - y_i)^2$ or sum

squared residuals (SSR). Use k-fold Cross Validation to get the optimal tuning parameter λ that does the minimization. I used 10-folds, the default for R's `cv.glmnet` command.

Now, the method I used is Double ML or Double LASSO regularization, used with ordinary least squares regression to finish off (Urminsky, Hansen, and Chernozhukov 2016). The double LASSO variable selection is a three step process (Belloni, Hansen, and Chernozhukov 2014).

Given that our original regression function has the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{i1} + \cdots + \beta_{k+1} W_{ik} + \epsilon_i$$

where X_i are the control variables and $W_{i1}, \ldots, W_{ik}$ are the potential covariates called "focal variables" (the sociodemographic variables),

Step 1: Fit a LASSO regression of focal variables $W_i k$ on the response Y_i. Keep the variables with nonzero coefficients $\alpha_i \neq 0$. This step insures robustness by selecting the covariates with strongest predictive power for determining the response variable.

$$Y_i = \alpha_0 + \alpha_1 W_{i1} + \cdots + \alpha_k W_{ik} + \epsilon_i$$

Step 2: Fit a LASSO regression of focal variables W_{ik} on each control X_i. Keep the variables with nonzero coefficients $\delta_i \neq 0$. This step makes sure to capture the key covariates that could support explaining the regression interest. Ideally, the residual variance will stay small even with the additional chosen covariates.

$$X_i = \delta_0 + \delta_1 W_{i1} + \cdots + \delta_k W_{ik} + \epsilon_i$$

Step 3: Run a linear regression on the response Y_i using the control variables X_i and union of kept focal variables from step 1 and step 2. Let the union be U.

$$Y_i = \beta_0 + \beta_1 X_i + \sum_{k \in U} \beta_{k+1} W_{ik} + \epsilon_i$$

3.2.2 Double ML v.s. LASSO and OLS

So why use a post-double-selection estimator over a post-single-selection estimator or least squares? Let us first discuss why double ML is preferred over OLS. The most notable issue of relying on OLS is Omitted Variable Bias. OLS fails to take into account that when covariates that have predictive power on the dependent variable and are correlated to the existing independent variables but are omitted from the regression, then the variation of model error depends on both dependent and omitted variables. This not only makes causal inference impossible but also biases the estimated parameters up or down (Belloni, Hansen, and Chernozhukov 2014). For example, refer to an example OLS in Appendix A.2 where

omitted variable bias is present due to naive variable selection. The current status of work (whether an individual is on layoff (NWLA), looking for work (NWLK), recalled (NWRE), etc.) will likely have an impact on income. It also likely is correlated to class of work (COW) and number of work weeks in a year (WKW), but that variable is missing in the regression, causing omitted variable bias. Also, in the case of high dimensionality, OLS performance and accuracy downfalls as more predictive features are added to the regression accordingly to the Curse of Dimensionality. The double LASSO prevents both omitted variable bias and running into high dimensionality issues like the LASSO. The first two steps ensure to capture the covariates that can predict the dependent variable and those that are correlated to the existing independent variables (the control group for this case). The properties of dimensionality reduction from the LASSO apply to the double LASSO. However, the LASSO itself has some issues.

The LASSO and double LASSO essentially have the same properties of regularization and dimensionality reduction. But, according to Urminsky, Hansen, and Chernozhukov 2016 the LASSO inflates Type I errors and may exclude non-zero coefficients with moderate effect, causing "significant regularization bias that adversely affects estimation and inference about β_1" and omitted variable bias (but not as much as blindly performing OLS). As this can lead to overfitting as well, an example is Appendix A.3. This process involves running a LASSO on log(IncPovRatio) using all the control variables and covariates, then fitting OLS using the remaining variables. Most variables are significant at almost 100% confidence level due to how LASSO selects these variables. Notice that the in-sample R-squared value is 0.9841, almost perfect positive correlation. With so many variables selected, this indicates that although in-sample prediction maybe decent, out-of-sample prediction can be disastrous due to overfitting. The double LASSO alleviates such issues by just following the three steps, "making it possible to perform robust/uniform inference after model selection" according to Belloni, Hansen, and Chernozhukov 2014, where the mechanical proof can be found. To summarize, the double LASSO insures the LASSO for potential overfitting and omitted variable bias by capturing what the LASSO might not catch on.

However, this does not mean that the double LASSO is fail proof from Type I errors or biases, let alone heteroskedasticity, since it simply uses data-driven optimization with the help of imperfect modeling using real data that are not perfectly randomly generated. This leads to the various tests and controls done. In the results section is discussed the model excluding all possible endogenous variables since double LASSO simply looks to minimize the loss function regardless of correlations between variables.

3.2.3 Control and Testing

Breusch-Pagan Test:

First, I did the Breusch-Pagan test for heteroskedasticity because if our regression turns out to be heteroskedastic, then our standard errors will be incorrect and so t-test and F-test results will be invalid. The null and alternative hypotheses are,

$$H_0 : \mathbb{V}(\epsilon|X_i) = \sigma^2$$

$$H_1 : \mathbb{V}(\epsilon|X_i) \neq \sigma^2$$

or given $\sigma^2 = h(\delta^T z)$ the auxiliary regression that explains the differences in the variances of the model, the null and alternative hypotheses are,

$$H_0 : \delta_1 = \cdots = \delta_p = 0$$

$$H_1 : \text{ at least one } \delta \neq 0$$

Now, get the Lagrange Multiplier (LM) statistic with the following steps,

1. Fit OLS on $Y = \beta^T X + \epsilon$ and get the residuals $\hat{u}$.
2. Fit the auxiliary regression $\hat{u}^2 = \delta^T z + \eta$.
3. Get the LM test statistic with sample size n and R-squared value from the auxiliary regression $LM = n * R^2 \sim \chi^2_{p-1}$.

If $Prob(LM \leq \chi^2_{p-1})$ is sufficiently small, we reject the null hypothesis in favor of the alternative, meaning that we have heteroskedasticity. This can be done in R using the `bptest` command.

F-Test:

If we have homoskedasticity, the regression results will return reliable individual significance for each feature variable and the joint significance across all feature variables used. If we have heteroskedasticity, use a heteroskedasticity robust regression. We will proceed even without this robustness due to lack of computational power required on our very large dataset and regression. Since we want to also know if our control variables (Travel Time to Work, SameResidenceWorkplace, Means of Transportation) truly have significant effects on log(IncPovRatio) as a whole, we do a F-test. Then, the null and alternative hypotheses are,

$$H_0 : \beta_{JWMNP} = \beta_{SameResidenceWorkplace} = \beta_{JWTR} = 0$$

$$H_1 : \text{ at least one } \beta \neq 0$$

and the F-statistic is,

$$F_{stat} = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n-k-1)} \sim F(q, df)$$

where R stands for the restricted model controlling all other variables as constant to test our three variables of interest

$$log(IncPovRatio) = \beta_0 + \beta_1 \text{JWMNP} + \beta_2 \text{SameResidenceWorkplace} + \beta_3 \text{JWTR}$$

This makes the UR the unrestricted model, which is the post-double LASSO OLS regression formula.

$$log(IncPovRatio) = \beta_0 + \beta_1 \text{JWMNP} + \beta_2 \text{SameResidenceWorkplace} + \beta_3 \text{JWTR} + \sum_{k \in U} \beta_{k+1} W_k$$

q = number of variables kept as control

k = number of independent variables

df = degrees of freedom = n-k-1

If $Prob(F_{stat} \leq F)$ is sufficiently small, we reject the null hypothesis in favor of the alternative, meaning that our three variables are jointly significant. This can be done in R using the `linearHypothesis` command.

False Discovery Rate Control: BH and BY Procedures:

When we test for individual significance, our null hypothesis is usually $H_0 : \beta_i = 0$. We calculate the test statistic and if this value is large or if the p-value is less than the critical value α, then the true coefficient β unlikely to be zero so we reject the null. Under the null hypothesis, the coefficient is normally distributed around zero, so if $p = 0.05$ we interpret it as "under the null hypothesis there is a likelihood of 0.05 of getting this estimate." Equivalently, this means 5% of the times we reject the null when actually it is true. This is a "false discovery" and it can happen purely by chance. This applies to when we test multiple hypotheses simultaneously for our regression where the probability of finding Type I error is almost 1, hence a "multiple comparison issue" and there are mathematical methods to control this.

The forerunner who first developed a method to control for false discovery rates is Carlo Emilio Bonferroni in 1936 with his Bonferroni procedure. Later Benjamini and Hochberg 1995 developed a method based on the Bonferroni procedure, called the Benjamini-Hochberg Procedure. Consider testing hypothesis $H_1, H_2, \ldots, H_m$ with corresponding p-values $P_1, P_2, \ldots, P_m$, so m = number of hypothesis tests done.

1. Sort the p-values from decreasing to increasing order so that $P_1 \leq P_2 \leq \cdots \leq P_m$. Let rank $i = 1$ for the lowest p-value and $i = 2$ for the second lowest p-value and so on.

2. Set the Benjamini-Hochberg critical value for a false discovery rate Q (commonly $Q =$ 10% or 20%). For each i, calculate

$$P_i < \frac{i}{m}Q$$

3. Get $\text{argmax}_{k \in i=\{1,\ldots,m\}} P_i$ satisfying the above inequality.

All null hypotheses H_i with index $i \leq k$ are rejected. The rest cannot be rejected, so it is probably best to use the BH-adjusted p-values.

$$P_i^* = \min\left(\frac{m}{i}P_i, P_{i+1}^*\right)$$

that is, the adjusted p-value for a test H_i is the smaller of the original p-value times m/i and the adjusted p-value for next test H_{i+1}. If this adjusted p-value is smaller than the false discovery rate, the hypothesis is rejected (McDonald 2014). This can be done in R using the `p.adjust(method = 'BH')` command.

Since the double LASSO is not fail proof from Type I error, we would apply Benjamini-Hochberg Procedure. But, do note that the BH method requires for each test to be independent of each other, which is likely not the case for us. To allow for dependency, we use the Benjamini–Yekutieli Procedure, which uses a similar equation,

$$P_i < \frac{i}{m * c(m)}Q$$

where $c(m) = \sum_{i=1}^{m} \frac{1}{i}$ (Benjamini and Yekutieli 2001). The BH method equals to the BY method when there is independence, $c(m) = 1$. This can be done in R using the `p.adjust(method = 'BY')` command. Other methods of taking care of dependence are bootstrapping or rerandomization.

R Squared

We focus on the out-of-sample R-Squared values. There are three methods for calculating them and all have their own characteristics (Duncan 2020). Given,

y = demean-ed actual vector

$\hat{y}$ = demean-ed predicted vector

$u = y - \hat{y}$ = residual vector,

$$R^2 = \frac{\hat{y}^T \hat{y}}{y^T y}$$
$$R^2 = 1 - \frac{u^T u}{y^T y} = 1 - \frac{SSR}{SST}$$
$$R^2 = corr(y, \hat{y})^2$$

The third equation is useful in the sense that it is the squared correlation coefficient of actual and fitted values of the response variable, so "this will always be between zero and 1 and if the actual and fitted are close, (R squared) will be close to 1 and if not close to zero." The second equation is perhaps the most referred equation but can be negative. The first equation can be greater than 1, so it has the power to explain more variation in the model than there actually is.

4 Results

4.1 Model One: Double LASSO

We follow the three steps of performing the double LASSO. In the first two steps, we do feature selection by doing a LASSO of log(IncPovRatio) on all potential covariates and another set of LASSO's of our control variables (Travel Time to Work, SameResidenceWorkplace, Means of Transportation) on the same set of all potential covariates. Before each LASSO do 10-fold cross validation to get the smallest tuning parameter λ that minimizes the Mean Squared Error. Use the λ as the hyperparameter of the LASSO.

For the LASSO of log(IncPovRatio) on covariates, we get $\lambda = 0.0001775132$ and are left with 157 covariates that we should keep.

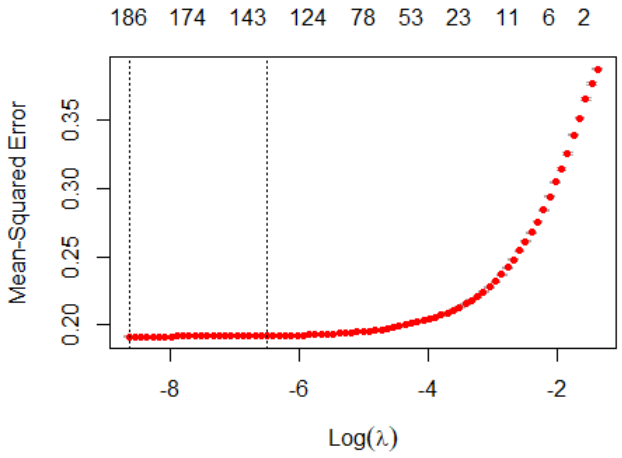


Figure 3: Tuning Parameter v.s. MSE

For the 3 LASSO's of controls on covariates, we get the following results. For code and plots refer to Appendix A.5.

LASSO function	λ **and number of covariates kept**
SameResidenceWorkplace $\sim$.	$\lambda = 6.405355 * 10^{-5}$ and $n = 74$
Travel Time to Work $\sim$.	$\lambda = 2.590354 * 10^{-4}$ and $n = 81$
Means of Transportation to Work (JWTR) $\sim$.	$\lambda = 3.31276 * 10^{-5}$ and $n = 213$

Do note that for JWTR a LASSO had to be done for each factor level, so 11 total (JWTR11 is dropped due to no observations), which explains why so seemingly many variables were selected. We take the union and we are left with final 219 covariates that will be used in the post-double LASSO regression. At first sight this may seem like an transparent overfitted model. However, considering that we originally had **5461** potential covariates, what the double LASSO returned is astounding because these variables were selected while preventing omitted variable bias and overfitting. We now present the regression result.

4.1.1 Model Output

Let us look at the control variables results, in particular the interaction terms, in the table below. Refer to table in section 3.2 above for the variable descriptions. The interaction SameResidenceWorkplace * Travel Time to Work is insignificance even though individually they are significant. This suggests that there is no evidence to say log(IncPovRatio) (or wealth) increases or decreases depending on travel time to work (or distance between residency and work) and if an individual's residency and work place is close or not. Because these two variables were proxy variables for distance between residency and work, our results suggests that either travel time to work has no correlation to wealth or it is a bad instrument in place for distance. But, we do have a say that compared to those who commute by car, truck, or van (JWTR01: reference group), with longer commute time, those who commute by bus (JWTR02), subway (JWTR04), train (JWTR05), ferryboat (JWTR06) are poorer on average. Those who walked to work (JWTR10) are wealthier. This suggests some evidence supporting the Spatial Mismatch Hypothesis. I assume those who walk to work can afford housing near work, which is generally more expensive as literature and recent evidence on housing prices suggest. Those who cannot afford are crowded out to locations far from labor and must commute by vehicle, which may or may not take longer than walking. But, it is reasonable to say those who use mass-public transit compared to private vehicles (e.g. cars) may have preferred the cheaper cost to commute, potentially implying a stricter wealth constraint. And from this pool of commuters, I assume those who get to work faster either are closer to work or use faster vehicles, both of which can be correlated to income or wealth. These conclusions deduced from our regression are simply correlational analyses.

	y = log(IncPovRatio)
SameResidenceWorkplaceTRUE	$-8.169E{-}02^{***}$
	$(4.054E{-}03)$
JWMNP	$1.431E{-}03^{***}$
	$(7.106E{-}05)$
JWTR02	$1.093E{-}02$
	$(1.481E{-}02)$
JWTR03	$3.399E{-}02$
	$(4.974E{-}02)$
JWTR04	$2.488E{-}01^{***}$
	$(1.612E{-}02)$
JWTR05	$3.626E{-}01^{***}$
	$(1.985E{-}02)$
JWTR06	$3.288E{-}01^{***}$
	$(4.727E{-}02)$
JWTR07	$8.782E{-}02^{***}$
	$(2.524E{-}02)$
JWTR08	$-2.325E{-}02$
	$(2.305E{-}02)$
JWTR09	$-5.182E{-}02^{**}$
	$(1.791E{-}02)$
JWTR10	$-6.349E{-}02^{***}$
	$(1.350E{-}02)$
JWTR12	$9.491E{-}03$
	$(1.469E{-}02)$
SameResidenceWorkplaceTRUE:JWMNP	$4.053E{-}06$
	$(7.287E{-}05)$
JWMNP:JWTR02	$-4.712E{-}04^{***}$
	$(1.278E{-}04)$
JWMNP:JWTR03	$-4.075E{-}04$
	$(1.036E{-}03)$
JWMNP:JWTR04	$-3.007E{-}03^{***}$
	$(1.743E{-}04)$
JWMNP:JWTR05	$-1.646E{-}03^{***}$
	$(1.913E{-}04)$
JWMNP:JWTR06	$-1.650E{-}03^{**}$
	$(6.101E{-}04)$
JWMNP:JWTR07	$-1.013E{-}03$
	$(7.045E{-}04)$
JWMNP:JWTR08	$3.503E{-}04$
	$(6.298E{-}04)$
JWMNP:JWTR09	$1.478E{-}03^{***}$
	$(4.447E{-}04)$
JWMNP:JWTR10	$4.182E{-}04^{\cdot}$
	$(2.490E{-}04)$
JWMNP:JWTR12	$-4.323E{-}05$
	$(1.213E{-}04)$

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

This leads to the question of correlation within the dependent variables and endogeneity. With so many covariates chosen purely in an algorithmic way, correlation between some covariates are very likely and even with the response variable log(IncPovRatio). For instance, consider the following,

DRIVESP2TRUE	$-6.237e{-}02^{***}$
	$(1.281e{-}02)$
DRIVESP3TRUE	$-5.937e{-}02^{***}$
	$(1.334e{-}02)$
DRIVESP4TRUE	$-5.789e{-}02^{***}$
	$(1.438e{-}02)$
DRIVESP5TRUE	$-3.096e{-}02^{*}$
	$(1.578e{-}02)$

DRIVESP is number of vehicles per person commuting. So if one drives alone, DRIVESP = 1. If one carpools with two more people, DRIVESP = 3 (DRIVESP3). This variable is very likely to be correlated to JWTR. If JWTR = 1, that is, one commutes by car, DRIVESP is probably high. If one commutes by public transit, DRIVESP is likely low. In terms of endogeneity, most dependent variables will probably cause reverse causality on wealth to some degree. Consider JWTR and DRIVESP. The wealthier one is, commute by car or even walking maybe likely and DRIVESP is likely higher. The regression involves many other covariates like SCIENGP (degree in STEM), SEX, WKW (number of work weeks per year), and MSP (current marriage status) that are likely correlated to one variable or another[7]. Such correlations affect test results and make the joint significance that the regression achieved less trustworthy. The double LASSO cannot judge such characteristics of each variable, so in the next regression we use econometrics and economic domain knowledge to drop variables that are likely endogenous.

The in-sample R-squared is 0.5123. The adjusted R-squared is 0.5122, which is very close to the R-squared value even after penalizing for including many variables in the OLS regression. The out-of-sample MSE is commendable 0.1894029.

4.1.2 Testing Output

The three out-of-sample R-squared values we got are: 0.5125579, 0.5089613, 0.5089676. Based on the third value, we can say that the predicted log(IncPovRatio) is not too close but also not too far from the actual. Also, observe that the first value is greater than the in-sample R-squared. This is possible because that method of calculating R-squared explains more variation than there is.

Now, based on the BP test result, we reject homoskedasticity in favor of heteroskedasticity. This shows that the post-double LASSO individual and joint tests are in fact invalid.

Proceeding with the test statistics we have, from the multiple comparisons we perform, it maybe likely to get a significant test result purely by chance. Controlling for false discovery rate using the BY method, we get the following, In the below plot 4 the black dots refer to hypotheses tests that can be rejected. The red dots refer to those that cannot be rejected. The cutoff depends on the false discovery rate what we will allow for (10%) and the potential dependencies between the hypotheses[8].

[7]For the full regression result go to middle of Appendix A.5.
[8]In bottom of Appendix A.5.

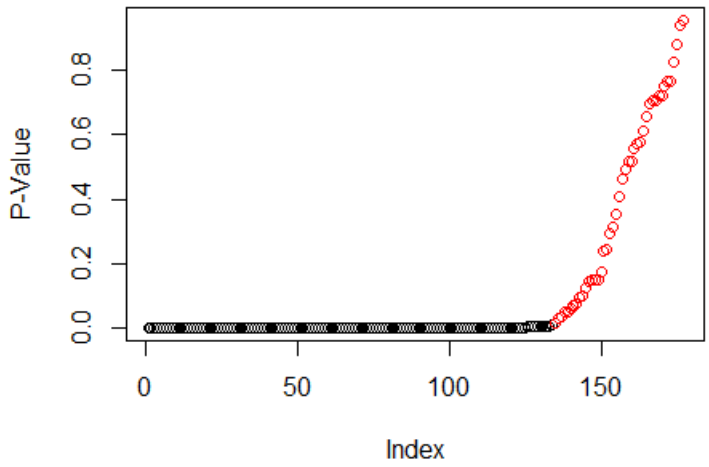


Figure 4: Which tests can be rejected or not?

A few hypotheses that cannot be rejected are significant at the 99% confidence level: Veteran status (VETERAN), citizenship status (CIT), race: pacific islander (RACPI) or other (RACSOR), and number of vehicles per commute persons (DRIVESP). If these hypotheses were rejected, they would be false discoveries. Also, this supports the evidence that certain means of transportation that were individually significant are indeed indicative of correlation to income.

We now observe the heteroskedasticity-robust F-tests on the control variables, excluding all covariates.

Hypothesis H_0	**F test stat**	**Significance**
$\beta_{SameResidenceWorkplace} = \beta_{JWMNP} = \beta_{JWTR}$	$F = 419.85$	$p < 2.2E-16^{***}$
$\beta_{JWMNP} = \beta_{JWTR} = \beta_{JWMNP*JWTR}$	$F = 154.04$	$p < 2.2E-16^{***}$
$\beta_{SameResidenceWorkplace} = \beta_{JWMNP} = \beta_{JWTR} = \beta_{SameResidenceWorkplace*JWMNP} = \beta_{JWMNP*JWTR}$	$F = 434.27$	$p < 2.2E-16^{***}$

The various F-tests show that various combinations of the control variables and/or their interactions are all jointly significant. This suggests that although individually the variables SameResidenceWorkplace and Travel Time to Work may not be important in explaining wealth, but as a whole with Means of Transportation, they are significant.

4.2 Model Two: Exogenous OLS

Presented is a linear model only using dependent variables that are not likely to be endogenous. This is a slight transition from answering *"what are the (correlative) effects of residency away from work"* to more of *"what are the determinants of wealth."* By dropping many variables that are likely endogenous with the use of economic knowledge, predictive power significantly decreases, but many of the remaining variables turn out to be significant in explaining characteristics of wealth. The following is a list of variables used in the OLS.

Selected Variables	Description
Weight (PWGTP)	Numeric: 1 - 9999 lbs
Age (AGEP)	Numeric: 0 - 99
Language other than English spoken at home (LANX)	Factor: **1** if Yes, speaks another language; **2** if No, speaks only English
Last Year Married (MARHYP):	Numeric: 1937 - 2018
Decade of entry to U.S. (DECADE)	Factor: **NA** (Born in the US); **1** Before 1950; **2** 1950 - 1959; **3** 1960 - 1969; **4** 1970 - 1979; **5** 1980 - 1989; **6** 1990 - 1999; **7** 2000 - 2009; **8** 2010 or later
Presence and age of own children (PAOC)	Factor: **NA** (male/female under 16 years old/GQ[9]); **1** Females with own children under 6 years only; **2** Females with own children 6 to 17 years only; **3** Females with own children under 6 years and 6 to 17 years; **4** Females with no own children
Quarter of birth (QTRBIR)	Factor: **1** January through March; **2** April through June; **3** July through September; **4** October through December
World area of birth (WAOB)	Factor: **1** US state; **2** PR and US Island Areas; **3** Latin America; **4** Asia; **5** Europe; **6** Africa; **7** Northern America; **8** Oceania and at Sea
Medicare (HINS3)	Factor: **1** if Yes; **2** if No
Age*Medicare (AGEP:HINS3)	Factor

4.2.1 Model Output

Below is the regression result (Appendix A.6). Most of the factors are related to personal characteristics like age, birth weight, which are not likely to cause wealth or vice versa, hence exogenous (although there are literature on the relationship between weight and income (Lähteenkorva, Silventoinen, and Lahelma 2004)). The in-sample R-squared value is low, 0.106. However, most are individually significant at least at the 99.9% confidence level except WAOB8, AGEP*HINS3, DECADE1.

[9]The U.S. Census defines Group Quarters as "noninstitutional living arrangements for groups not living in conventional housing units or groups living in housing units containing ten or more unrelated people or nine or more people unrelated to the person in charge." (Census 2019a)

	y = log(IncPovRatio)
PWGTP	$-1.671e{-}04^{***}$
	$(8.701e{-}06)$
AGEP	$8.362e{-}03^{***}$
	$(9.287e{-}05)$
LANX2TRUE	$5.025e{-}02^{***}$
	$(2.234e{-}03)$
MARHYP	$1.625e{-}03^{***}$
	$(7.909e{-}05)$
DECADE3TRUE	$1.416e{-}01^{***}$
	$(7.867e{-}03)$
DECADE4TRUE	$1.505e{-}01^{***}$
	$(5.346e{-}03)$
DECADE7TRUE	$2.341e{-}02^{***}$
	$(3.879e{-}03)$
DECADE8TRUE	$-5.670e{-}02^{***}$
	$(4.600e{-}03)$
PAOC1TRUE	$-1.810e{-}01^{***}$
	$(3.869e{-}03)$
PAOC2TRUE	$-2.695e{-}01^{***}$
	$(2.217e{-}03)$
PAOC4TRUE	$-3.561e{-}01^{***}$
	$(1.553e{-}03)$
QTRBIR3TRUE	$4.710e{-}03^{**}$
	$(1.610e{-}03)$
WAOB2TRUE	$-1.658e{-}01^{***}$
	$(9.589e{-}03)$
WAOB3TRUE	$-3.119e{-}01^{***}$
	$(3.194e{-}03)$
WAOB5TRUE	$8.521e{-}02^{***}$
	$(4.784e{-}03)$
WAOB6TRUE	$-1.018e{-}01^{***}$
	$(7.735e{-}03)$
WAOB7TRUE	$2.027e{-}01^{***}$
	$(1.084e{-}02)$
WAOB8TRUE	$4.730e{-}02^{*}$
	$(1.933e{-}02)$
AGEP:HINS31	$-2.175e{-}04$
	$(3.219e{-}04)$
DECADE5TRUE	$9.779e{-}02^{***}$
	$(4.130e{-}03)$
HINS32TRUE	$1.916e{-}01^{***}$
	$(2.207e{-}02)$
DECADE1TRUE	$7.548e{-}02^{\cdot}$
	$(4.239e{-}02)$
DECADE2TRUE	$7.733e{-}02^{***}$
	$(1.371e{-}02)$
PAOC3TRUE	$-2.705e{-}01^{***}$
	$(4.101e{-}03)$
QTRBIR2TRUE	$5.347e{-}03^{**}$
	$(1.649e{-}03)$

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

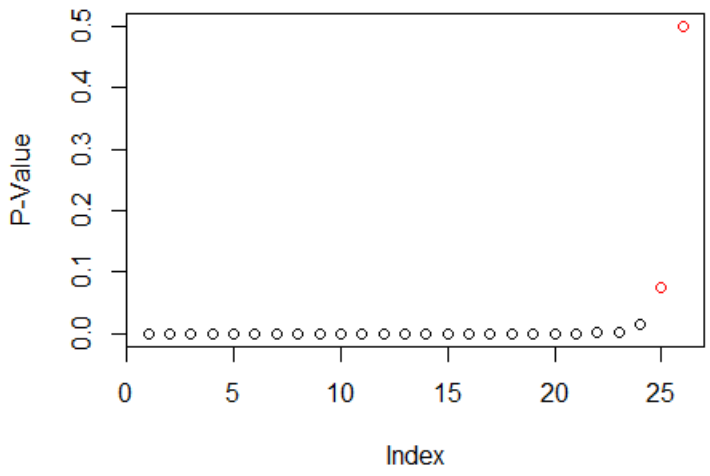


Figure 5: Which tests can be rejected or not?

These variables are jointly significance according to the F-test statistic. Based on such results, we say there is evidence suggesting that the exogenous sociodemographic variables selected do have explanatory power for wealth. Of those that are significant, We see that the coefficients of age related variables suggest higher age explains for higher log(IncPovRatio), look at AGEP, MARHYP, and DECADE. Variables with negative coefficients include PWGTP, PAOC, and WAOB, excluding WAOB5, which is European born persons. Although lacking evidence to prove causality in this paper, these variables with negative coefficients are some of the characteristics which citizens today in the U.S. are often discriminated by: external appearance, own child characteristics, and location of birth (often incorrectly associated with certain races).

4.2.2 Testing Output

The three out-of-sample R-squared values are 0.1064778, 0.1028011, 0.1028329. Values are quite similar to the in-sample R-squared just like post-double LASSO regression.

The BP test result suggests that heteroskedasticity exists, so the test results the OLS returns are in fact invalid.

We perform the BY procedure to control for false discovery rate at 10%. We discover that variables entered U.S. before 1950 (DECADE1) and age * Medicare (AGEP*HINS3) cannot be rejected, which were already statistically insignificant before controlling for Type I errors. Refer to the two red dots in Figure 5.

5 Conclusion

Income equality is an unfortunate byproduct of many factors including sociodemographic disparity and location-wise stratification which negatively affects the basic livelihood of the majority citizens in the U.S. As the Spatial Mismatch Hypothesis suggests, there is a misalignment in locations of labor and locations of residency and this paper provides some evidence that this behavior is related to income equality. Using the pretense if an individuals residency is close or far to work, using travel time to work as substitute for the physical distance to work, and using log of income poverty ratio as a proxy for wealth has shown that there is statistically significant relationship between the distancing of residency

and work with wealth, given certain types of means of transportation individuals use to commute. However, a critical limitation to this solution is that income and poverty brackets cannot serve as an all-encompassing proxy for wealth. Income does provide information for tangible properties, assets, or outside money inflow. It also cannot capture liabilities such as debt, which itself is unclear how to be measured. The poverty thresholds used to calculate the ratio, which depends on household size and age, do not capture the average requirement for sustenance of a household and is often noted as "not enough" by the public.

The economic modeling is imperfect as well. We have shown evidence supporting our hypothesis and have also shown potential exogenous sociodemographic characteristics that may determine wealth to some degree. However, for several reasons our results may be erroneous. First, our pretense if an individuals residency is close or far to work (SameResidenceWorkplace) is determined on a state basis. So if one works and lives in the same state that is a yes or no question. This approach is not exhaustive of the analysis for spatial mismatch based on location; however, it is also not disregardful since a non-negligible number of people work out-of-state from home (42,925 out of 933,788 as noted previously). Future studies should use the Public Use Microdata Area (PUMA) data available from the U.S. Census for a more thorough analysis. Second, we have shown that heteroskedasticity is present, so our test results may in fact be incorrect. Lastly, although we have chosen the double LASSO for algorithmic feature selection to prevent overfitting and omitted variable bias in the dataset as much as possible, randomness is always present and data is always lacking. The produced results could have occurs purely by chance. Also, do note that the U.S. Census dataset I used is not indicative of the population. The datasets provided to the public is a sample of a larger embodiment. And, the data can never include all information needed for study. For example, liabilities like debt are missing, hence omitted variable bias is inevitable for this paper. For future studies, using a larger dataset with more quantitative information on the financial statue of an individual and using different sampling methods like stratified sampling maybe useful. Additionally, since this paper focuses on 2018 data available, it is possible to compile previous years data to construct a time-series data based research question.

This paper is an attempt to bring additional attention to the issue of income inequality and the current state of our communities. Many people suffer financially and so have difficulty bettering family households. A relatively small proportion of research has been done on income equality with respect to very basic characteristics of these people, like commuting. The U.S. Census has released reports on similar issues (Census 2011a and Census 2011b), but no policy has caught my eye making significantly positive contribution to bettering the family livelihood. My hope is that the larger literature contributes more to identifying key factors that cause inequality in wealth, so that future policies alleviate such disparity in a fair manner without discriminating some in favor of the suffering majority.

References

Adam Millsap, Forbes. 2018. "Is Seattle Doomed?" Visited on 03/01/2019. https://www.forbes.com/sites/adammillsap/2018/05/21/is-seattle-doomed/#442ad39b74f4.

Belloni, Alexandre, Christian Hansen, and Victor Chernozhukov. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls". *The Review of Economic Studies* 81, no. 2 (): 608–650. doi:https://doi.org/10.1093/restud/rdt044.

Benjamini, Yaov, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing Under Dependency". *Annals of Statistics* 29 (4): 1165–1188. doi:https://doi.org/10.1214/aos/1013699998.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. ISSN: 00359246. http://www.jstor.org/stable/2346101.

Brown, Lawrence A., and Eric G. Moore. 1970. "The Intra-Urban Migration Process: A Perspective". *Geografiska Annaler. Series B, Human Geography* 52 (1): 1–13. ISSN: 04353684, 14680467. http://www.jstor.org/stable/490436.

Building Solutions, CRH Americas. 2017. "What's Driving the Pacific Northwest Economy?" Visited on 02/27/2019. https://www.buildingsolutions.com/industry-insights/whats-driving-the-pacific-northwest-economy.

Census, U.S. 2019. "How the Census Bureau Measures Poverty". Visited on 03/15/2020. https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html.

Census, U.S. Bureau of the. 2010. "Metropolitan and Micropolitan". Visited on 02/27/2019. https://www.census.gov/programs-surveys/metro-micro/about.html.

— . 2011a. "Commuting in the United States: 2009". https://www2.census.gov/library/publications/2011/acs/acs-15.pdf.

— . 2011b. "U.S. Neighborhood Income Inequality in the 2005–2009 Period". https://www2.census.gov/library/publications/2011/acs/acs-16.pdf.

— . 2019a. "Subject Definitions". https://www.census.gov/programs-surveys/cps/technical-documentation/subject-definitions.html.

— . 2019b. *Total Construction Spending: Transportation [TLTRANSCONS]*. Retrieved from FRED. Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/TLTRANSCONS.

Dubay, Lisa, John Holahan, and Cook Allison. 2006. "The Uninsured And The Affordability Of Health Insurance Coverage". *HealthAffairs* 25. doi:https://doi.org/10.1377/hlthaff.26.1.w22.

Duncan, Gregory. 2020. *Notes on Out of sample R.*

Durlauf, Steven N. 1992. "A Theory of Persistent Income Inequality". *NBER Working Paper Series*, no. 4056 ().

FHFA, U.S. Federal Housing Finance Agency. 2019. *All-Transactions House Price Index for Seattle-Bellevue-Kent, WA (MSAD) [ATNHPIUS42644Q].* Retrieved from FRED. Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/ATNHPIUS42644Q.

Finio, Nicholas, and Lawrence Mishel. 2013. https://www.epi.org/publication/ib347-earnings-top-one-percent-rebound-strongly/.

Frumkin, Howard. 2002. "Urban Sprawl and Public Health". *Public Health Reports* 117 (): 201–217.

Greenstone, Scott. 2019. "Is Seattle's homeless crisis the worst in the country?" https://www.seattletimes.com/seattle-news/homeless/is-seattles-homeless-crisis-the-worst-in-the-country/.

Jong, De, Gordon F., and Ralph R. Sell. 1977. "Population Redistribution, Migration, and Residential Preferences". *The Annals of the American Academy of Political and Social Science* 429:130–144. ISSN: 00027162. http://www.jstor.org/stable/1041580.

Kain, John F. 1994. "The Spatial Mismatch Hypothesis: Three Decades Later". *Housing Policy Debate* 3 (2): 371–392.

Labor Statistics, U.S. Bureau of. 2020. *CONSUMER EXPENDITURES MIDYEAR UPDATE–JULY 2018 THROUGH JUNE 2019 AVERAGE.* Visited on 05/10/2020. https://www.bls.gov/news.release/cesmy.nr0.htm.

Lähteenkorva, Sirpa, Karri Silventoinen, and Eero Lahelma. 2004. "Relative Weight and Income at Different Levels of Socioeconomic Status". *Am J Public Health* 94 (3): 468–472. doi:https://doi.org/10.2105/ajph.94.3.468.

McDonald, John H. 2014. *Handbook of Biological Statistics.* Baltimore, Maryland: Sparky House Publishing.

Piketty, Thomas. 2014. "Capital in the Twenty First Century".

Sanchez, Thomas W. 2002. "The Impact of Public Transport on US Metropolitan Wage Inequality". *Urban Studies* 39 (3): 423–436. doi:10.1080/00420980220112766.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso". *Royal Statistical Society.*

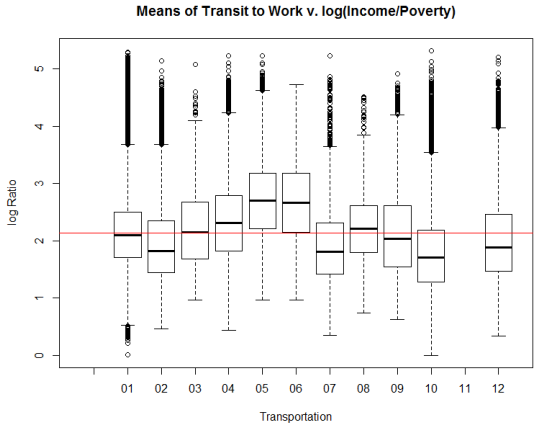
Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov. 2016. "Using Double-Lasso Regression for Principled Variable Selection".

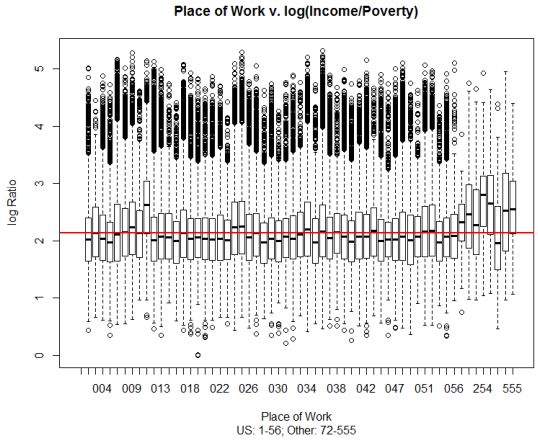
Watson, Tara. 2009. "INEQUALITY AND THE MEASUREMENT OF RESIDENTIAL SEGREGATION BY INCOME IN AMERICAN NEIGHBORHOODS". *Review of Income and Wealth* 55 (3): 820–844.

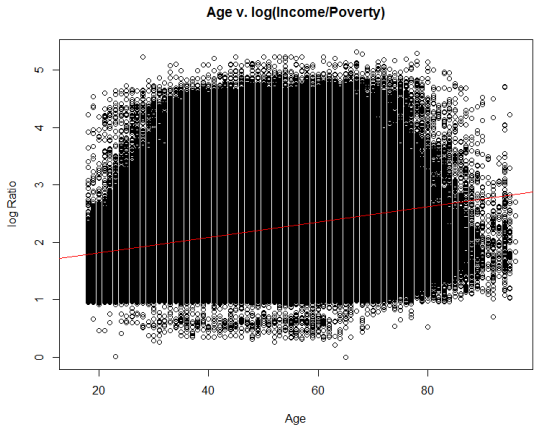
Zenou, Yves. 2010. "Search, migration, and urban land use: The case of transportation policies". `http://hdl.handle.net/10419/51747`.

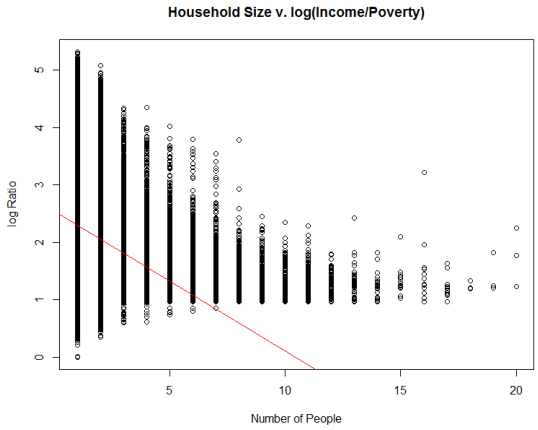
A Appendix

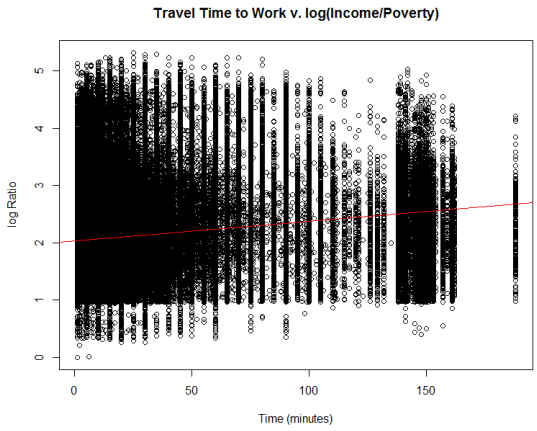
A.1 Data Exploration











The red lines in the top two box plots are the mean of log(IncPovRatio).
The red lines in the scatterplots are the linear fits.

A.2 OLS Example

```
OLS with control and sociodemographic variables

Call:
lm(formula = formula4, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6867 -0.2591 -0.0307  0.2170  3.7607

Coefficients:
                                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)                     1.272e+00  7.424e-03  171.315  < 2e-16 ***
SameResidenceWorkplaceTRUE      6.955e-02  2.250e-02    3.091 0.001992 **
JWMNP                           1.446e-03  1.768e-05   81.814  < 2e-16 ***
JWTR02                          2.418e-02  4.630e-03    5.221 1.78e-07 ***
JWTR03                          1.030e-01  2.877e-02    3.579 0.000344 ***
JWTR04                          2.459e-01  6.268e-03   39.236  < 2e-16 ***
JWTR05                          3.344e-01  1.133e-02   29.504  < 2e-16 ***
JWTR06                          3.582e-01  3.311e-02   10.817  < 2e-16 ***
JWTR07                          1.026e-01  1.290e-02    7.959 1.74e-15 ***
JWTR08                          1.487e-02  1.395e-02    1.066 0.286536
JWTR09                          5.440e-03  7.857e-03    0.692 0.488676
JWTR10                         -1.624e-02  2.979e-03   -5.451 5.02e-08 ***
JWTR12                          1.105e-02  5.096e-03    2.168 0.030193 *
MIG2                           -7.284e-02  5.838e-03  -12.478  < 2e-16 ***
MIG3                           -2.535e-02  1.118e-03  -22.671  < 2e-16 ***
AGEP                            6.416e-03  3.645e-05  176.040  < 2e-16 ***
SPORDER                        -1.511e-01  3.932e-04 -384.384  < 2e-16 ***
COW2                           -4.088e-02  1.344e-03  -30.412  < 2e-16 ***
COW3                           -6.203e-02  1.442e-03  -43.002  < 2e-16 ***
COW4                           -7.189e-02  1.735e-03  -41.433  < 2e-16 ***
COW5                            1.256e-01  2.378e-03   52.815  < 2e-16 ***
COW6                           -9.278e-02  1.734e-03  -53.519  < 2e-16 ***
COW7                            1.459e-01  2.050e-03   71.171  < 2e-16 ***
COW8                           -1.569e-01  8.643e-03  -18.154  < 2e-16 ***
OC0                             1.598e-01  3.350e-03   47.715  < 2e-16 ***
MAR2                           -4.369e-02  2.727e-03  -16.021  < 2e-16 ***
MAR3                           -3.605e-02  1.233e-03  -29.242  < 2e-16 ***
MAR4                           -6.801e-02  2.857e-03  -23.806  < 2e-16 ***
MAR5                           -9.702e-02  1.018e-03  -95.329  < 2e-16 ***
WKHP                            1.354e-02  3.492e-05  387.783  < 2e-16 ***
WKW2                           -6.885e-02  2.569e-03  -26.804  < 2e-16 ***
WKW3                           -1.540e-01  1.670e-03  -92.238  < 2e-16 ***
WKW4                           -2.676e-01  1.888e-03 -141.745  < 2e-16 ***
WKW5                           -3.643e-01  2.392e-03 -152.272  < 2e-16 ***
WKW6                           -4.637e-01  2.412e-03 -192.216  < 2e-16 ***
```

```
HINS12                             -1.670e-01  1.060e-03 -157.507  < 2e-16 ***
HINS22                             -1.702e-02  1.277e-03  -13.334  < 2e-16 ***
HINS32                             -3.675e-02  1.718e-03  -21.388  < 2e-16 ***
HINS42                              6.789e-02  1.523e-03   44.581  < 2e-16 ***
HINS52                             -3.567e-02  2.810e-03  -12.694  < 2e-16 ***
HINS62                              6.426e-02  3.152e-03   20.385  < 2e-16 ***
MIL2                                1.813e-01  5.705e-03   31.777  < 2e-16 ***
MIL3                                1.600e-01  6.501e-03   24.617  < 2e-16 ***
MIL4                                1.421e-01  5.655e-03   25.121  < 2e-16 ***
SCH2                               -2.897e-02  1.579e-03  -18.350  < 2e-16 ***
SCH3                               -3.179e-02  2.578e-03  -12.333  < 2e-16 ***
SEX2                               -1.609e-01  7.859e-04 -204.788  < 2e-16 ***
NATIVITY2                          -4.411e-02 "message"  -36.595  < 2e-16 ***
RAC1P2                             -7.036e-02  1.329e-03  -52.959  < 2e-16 ***
RAC1P3                             -5.855e-02  4.415e-03  -13.262  < 2e-16 ***
RAC1P4                              1.023e-01  1.757e-02    5.822 5.83e-09 ***
RAC1P5                             -6.698e-02  1.047e-02   -6.395 1.60e-10 ***
RAC1P6                              6.592e-02  1.756e-03   37.543  < 2e-16 ***
RAC1P7                             -2.177e-02  8.779e-03   -2.480 0.013155 *
RAC1P8                             -4.580e-02  1.998e-03  -22.919  < 2e-16 ***
RAC1P9                              6.368e-04  2.402e-03    0.265 0.790916
SCIENGRLP1                          4.104e-01  1.907e-03  215.249  < 2e-16 ***
SCIENGRLP2                          3.656e-01  8.467e-04  431.820  < 2e-16 ***
SameResidenceWorkplaceTRUE:JWMNP -9.941e-04  3.671e-04   -2.708 0.006771 **
JWMNP:JWTR02                       -6.293e-04  8.084e-05   -7.785 6.99e-15 ***
JWMNP:JWTR03                       -1.385e-03  5.776e-04   -2.397 0.016517 *
JWMNP:JWTR04                       -2.612e-03  1.140e-04  -22.906  < 2e-16 ***
JWMNP:JWTR05                       -1.020e-03  1.429e-04   -7.137 9.52e-13 ***
JWMNP:JWTR06                       -1.458e-03  4.458e-04   -3.272 0.001069 **
JWMNP:JWTR07                       -1.353e-03  4.417e-04   -3.063 0.002190 **
JWMNP:JWTR08                       -6.767e-05  4.550e-04   -0.149 0.881773
JWMNP:JWTR09                        5.685e-04  2.917e-04    1.949 0.051308 .
JWMNP:JWTR10                       -8.489e-05  1.532e-04   -0.554 0.579492
JWMNP:JWTR12                        2.255e-04  9.071e-05    2.486 0.012936 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4291 on 1379179 degrees of freedom
  (15943 observations deleted due to missingness)
Multiple R-squared:  0.5441,Adjusted R-squared:  0.5441
F-statistic: 2.421e+04 on 68 and 1379179 DF,  p-value: < 2.2e-16
```

A.3 LASSO Example

```
Summary of LASSO:

Call:
rlasso.formula(formula = formula, data = data, post = post, intercept = intercept,
    model = model, control = control)

Post-Lasso Estimation:  FALSE

Total number of variables: 250
Number of selected variables: 103

Residuals:
      Min        1Q    Median        3Q       Max
-92.67302  -0.48111  -0.01726   0.55912  27.01928


Post-LASSO OLS Regression output:

Call:
lm(formula = formula, data = dattemp)

Residuals:
    Min      1Q  Median      3Q     Max
-92.323  -0.482  -0.015   0.568  26.917

Coefficients: (11 not defined because of singularities)
                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)     4.205e+00  3.639e-02  115.554  < 2e-16 ***
DIVISION6TRUE   3.341e-02  7.148e-03    4.674 2.96e-06 ***
DIVISION7TRUE   1.471e-02  8.850e-03    1.662 0.096555 .
SPORDER        -5.102e-01  3.890e-03 -131.174  < 2e-16 ***
REGION2TRUE     1.302e-02  4.020e-03    3.238 0.001204 **
REGION3TRUE     1.039e-02  4.725e-03    2.200 0.027823 *
ST04TRUE        3.158e-02  1.078e-02    2.930 0.003391 **
ST08TRUE       -1.729e-02  1.092e-02   -1.583 0.113412
ST12TRUE        2.783e-02  7.117e-03    3.911 9.21e-05 ***
ST25TRUE       -3.826e-02  9.951e-03   -3.845 0.000121 ***
ST48TRUE        2.823e-03  9.366e-03    0.301 0.763129
ST50TRUE       -4.237e-02  3.080e-02   -1.376 0.168923
PWGTP          -8.990e-05  1.917e-05   -4.691 2.72e-06 ***
AGEP            3.982e-03  2.000e-04   19.906  < 2e-16 ***
CIT3TRUE       -4.542e-02  1.476e-02   -3.076 0.002096 **
CIT5TRUE        7.255e-02  8.051e-03    9.011  < 2e-16 ***
COW2TRUE       -1.438e-02  5.205e-03   -2.763 0.005731 **
COW3TRUE       -2.364e-02  5.412e-03   -4.367 1.26e-05 ***
COW4TRUE       -1.960e-02  6.584e-03   -2.977 0.002910 **
```

```
COW6TRUE       -1.321e-01  7.140e-03  -18.499  < 2e-16 ***
FER1TRUE        9.337e-02  1.446e-02    6.456 1.08e-10 ***
FER2TRUE        7.017e-02  5.880e-03   11.935  < 2e-16 ***
GCL1TRUE       -2.140e-01  1.218e-02  -17.578  < 2e-16 ***
GCL2TRUE       -1.814e-01  6.963e-03  -26.051  < 2e-16 ***
GCM5TRUE       -2.302e-02  2.292e-02   -1.004 0.315365
HINS12TRUE     -5.952e-02  5.203e-03  -11.439  < 2e-16 ***
HINS32TRUE     -2.257e-01  1.433e-02  -15.742  < 2e-16 ***
HINS42TRUE      1.929e-01  1.382e-02   13.956  < 2e-16 ***
HINS52TRUE     -4.397e-02  1.024e-02   -4.294 1.76e-05 ***
HINS62TRUE      3.034e-01  1.418e-02   21.398  < 2e-16 ***
HINS72TRUE      3.545e-02  2.122e-02    1.670 0.094844 .
JWMNP          -2.891e-04  6.185e-05   -4.675 2.94e-06 ***
MAR2TRUE       -4.096e-01  9.378e-03  -43.684  < 2e-16 ***
MAR3TRUE       -2.586e-01  4.837e-03  -53.454  < 2e-16 ***
MAR4TRUE       -3.086e-01  9.891e-03  -31.206  < 2e-16 ***
MARHT2TRUE      1.668e-02  3.815e-03    4.372 1.23e-05 ***
MIG3TRUE       -2.049e-02  4.978e-03   -4.116 3.85e-05 ***
NWAV3TRUE      -6.834e-02  2.135e-02   -3.201 0.001369 **
NWLA3TRUE      -2.435e-02  5.443e-03   -4.474 7.68e-06 ***
OIP            -5.635e-06  5.153e-07  -10.933  < 2e-16 ***
PAP            -2.867e-05  4.475e-06   -6.407 1.48e-10 ***
RELP01TRUE     -1.903e+00  5.148e-03 -369.648  < 2e-16 ***
RELP02TRUE     -1.070e+00  1.373e-02  -77.971  < 2e-16 ***
RELP03TRUE     -1.071e+00  6.373e-02  -16.801  < 2e-16 ***
RELP04TRUE     -9.403e-01  4.462e-02  -21.074  < 2e-16 ***
RELP05TRUE     -1.240e+00  2.277e-02  -54.465  < 2e-16 ***
RELP06TRUE     -1.238e+00  2.196e-02  -56.352  < 2e-16 ***
RELP07TRUE     -9.612e-01  4.560e-02  -21.080  < 2e-16 ***
RELP08TRUE     -1.318e+00  4.307e-02  -30.604  < 2e-16 ***
RELP09TRUE     -1.148e+00  2.547e-02  -45.064  < 2e-16 ***
RELP10TRUE     -1.184e+00  2.420e-02  -48.912  < 2e-16 ***
RELP11TRUE     -1.363e+00  3.103e-02  -43.921  < 2e-16 ***
RELP12TRUE     -1.338e+00  2.088e-02  -64.108  < 2e-16 ***
RELP13TRUE     -1.492e+00  1.372e-02 -108.700  < 2e-16 ***
RELP15TRUE     -1.350e+00  2.333e-02  -57.857  < 2e-16 ***
RELP17TRUE     -4.254e-01  3.198e-02  -13.303  < 2e-16 ***
SCH2TRUE        3.941e-02  1.601e-02    2.461 0.013858 *
SCHG14TRUE      1.553e-01  5.019e-02    3.094 0.001973 **
SCHG15TRUE     -7.934e-03  1.587e-02   -0.500 0.617087
SCHG16TRUE     -4.042e-02  1.513e-02   -2.672 0.007550 **
SEMP            6.773e-05  1.053e-07  643.445  < 2e-16 ***
SEX2TRUE       -7.105e-02  8.507e-03   -8.352  < 2e-16 ***
SSIP           -1.455e-05  2.055e-06   -7.081 1.43e-12 ***
SSP             5.676e-06  4.594e-07   12.356  < 2e-16 ***
WAGP            6.454e-05  8.293e-08  778.294  < 2e-16 ***
WKHP           -3.437e-03  1.423e-04  -24.153  < 2e-16 ***
```

```
WKW3TRUE        5.603e-02  6.843e-03    8.187 2.68e-16 ***
WKW4TRUE        1.098e-01  8.232e-03   13.342  < 2e-16 ***
WKW5TRUE        1.360e-01  1.131e-02   12.026  < 2e-16 ***
WKW6TRUE        1.247e-01  1.153e-02   10.820  < 2e-16 ***
WRK1TRUE        5.300e-02  6.821e-03    7.770 7.84e-15 ***
ANC2TRUE       -2.895e-02  3.562e-03   -8.129 4.34e-16 ***
ANC4TRUE       -1.583e-02  4.652e-03   -3.403 0.000667 ***
DECADE2TRUE     1.281e-01  2.932e-02    4.369 1.25e-05 ***
DECADE7TRUE     4.404e-02  8.442e-03    5.217 1.82e-07 ***
DECADE8TRUE     1.229e-01  1.088e-02   11.297  < 2e-16 ***
DIS2TRUE        5.066e-02  6.118e-03    8.281  < 2e-16 ***
ESR4TRUE       -6.295e-02  2.132e-02   -2.953 0.003152 **
HICOV2TRUE      1.679e-01  1.007e-02   16.678  < 2e-16 ***
MSP2TRUE       -1.845e-01  9.556e-03  -19.306  < 2e-16 ***
NATIVITY2TRUE   6.933e-03  5.483e-03    1.265 0.206048
PAOC2TRUE       3.558e-02  7.697e-03    4.623 3.79e-06 ***
PAOC4TRUE       8.357e-02  7.733e-03   10.807  < 2e-16 ***
PINCP           7.651e-05  7.705e-08  992.925  < 2e-16 ***
PRIVCOV2TRUE   -7.261e-02  8.985e-03   -8.081 6.42e-16 ***
PUBCOV2TRUE    -3.374e-01  1.460e-02  -23.108  < 2e-16 ***
QTRBIR2TRUE    -8.179e-03  3.359e-03   -2.435 0.014898 *
RAC1P2TRUE     -3.158e-02  5.886e-03   -5.365 8.11e-08 ***
SCIENGP2TRUE   -3.824e-02  3.740e-03  -10.224  < 2e-16 ***
SCIENGRLP1TRUE -3.718e-02  7.436e-03   -5.000 5.73e-07 ***
SFN1TRUE       -3.761e-01  2.015e-02  -18.663  < 2e-16 ***
SFR2TRUE       -9.101e-02  2.510e-02   -3.626 0.000288 ***
SFR3TRUE        5.012e-01  2.716e-02   18.457  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.424 on 976619 degrees of freedom
Multiple R-squared:  0.9841,Adjusted R-squared:  0.9841
F-statistic: 6.576e+05 on 92 and 976619 DF,  p-value: < 2.2e-16
```

A.4 Data Preparation

```
knitr::opts_chunk$set(echo = TRUE)
#+ setup, error=TRUE
```

Data Preparation

```
# Brian Kang
# Must use 64-bit version of R
#knitr::opts_chunk$set(error=TRUE)
#rm(list = ls())

setwd("H:/Honors Research")
#chooseCRANmirror(graphics=FALSE, ind=1)
#install.packages(c("hdm","dplyr","stringr","glmnet","gdata","fastDummies"))
library(hdm)
library(dplyr)

library(stringr)
library(glmnet)

library(gdata)

library(fastDummies)
library(lmtest)

library(car)

# Import dataset ------------------------------------------------------------------
temp <- read.csv("psam_pusb.csv", header = T, nrows = 1)
# columns that don't import as factors using default settings
fnf <- c("DIVISION","PUMA","REGION","ST","ADJINC","CIT","COW","DDRS","DEAR",
         "DEYE","DOUT","DPHY","DRAT","DRATX","DREM","ENG","FER","GCL","GCM","GCR",
         "HINS1","HINS2","HINS3","HINS4","HINS5","HINS6","HINS7","JWTR","LANX",
         "MAR","MARHD","MARHM","MARHT","MARHW","MIG","MIL","MLPA","MLPB","MLPCD",
         "MLPE","MLPFG","MLPH","MLPI","MLPJ","MLPK","NWAB","NWAV","NWLA","NWLK",
         "NWRE","RELP","SCH","SCHG","SCHL","SEX","WKL","WKW","WRK","ANC","ANC1P",
         "ANC2P","DECADE","DIS","DRIVESP","ESP","ESR","FOD1P","FOD2P","HICOV",
         "HISP","INDP","JWAP","JWDP","LANP","MIGPUMA","MIGSP","MSP","NATIVITY",
         "NOP","OC","OCCP","PAOC","POBP","POWPUMA","POWSP","PRIVCOV","PUBCOV",
         "QTRBIR","RAC1P","RAC2P","RAC3P","RACAIAN","RACASN","RACBLK","RACNH",
         "RACNUM","RACPI","RACSOR","RACWHT","RC","SCIENGP","SCIENGRLP","SFN",
         "SFR","VPS","WAOB")
# columns that do import as factors using default setting
fif <- c("RT","SERIALNO","NAICSP","SOCP")
# all columns that are factors
fcf <- append(fnf,fif)
fcf <- append(fcf,names(temp[,c(131:286)]))
# vector of classes of data columns
colclass <- ifelse(colnames(temp) %in% fcf, 'factor', 'numeric')

temp1 <- read.csv("psam_pusa.csv", header = T, colClasses = colclass)    # U.S. PUMS data
temp2 <- read.csv("psam_pusb.csv", header = T, colClasses = colclass)
dat <- rbind(temp1,temp2)
dat <- dat[,-c(1,2)] # drop unecessary IDs
dat <- dat[,-c(129:284)] # drop unnecessary flag vars

# US Census: "Income used to calculate poverty status includes PERNP (earnings) and PINCP (income)"

# Calculate Income-Poverty ratio --------------------------------------------------
# (POVPIP only shows NA or <0.5 or >=0.5 so calculate actual ratio)
PovertyThreshold <- rep(NA, nrow(dat))
getThreshold <- function(threshold) { # values from CPS 2018
  for (i in 1:nrow(dat)) {
    if (dat$AGEP[i] < 18) { # under 18 yrs
      threshold[i] <- NA
    } else if (dat$SPORDER[i]==1 & dat$AGEP[i] < 65) { # individual
      threshold[i] <- 13064
    } else if (dat$SPORDER[i]==1 & dat$AGEP[i] >= 65) {
      threshold[i] <- 12043
```

```
      # AGEP doesn't have NA values
    } else if (dat$SPORDER[i]==2 & dat$AGEP[i] < 65 & dat$OC[i]==0) { # two people
      threshold[i] <- 16815
    } else if (dat$SPORDER[i]==2 & dat$AGEP[i] < 65 & dat$OC[i]==1) {
      threshold[i] <- 17308
    } else if (dat$SPORDER[i]==2 & dat$AGEP[i] >= 65 & dat$OC[i]==0) {
      threshold[i] <- 15178
    } else if (dat$SPORDER[i]==2 & dat$AGEP[i] >= 65 & dat$OC[i]==1) {
      threshold[i] <- 17242
    } else if (dat$SPORDER[i]==2) { # OC is NA value
      threshold[i] <- 16247
    } else if (dat$SPORDER[i]==3 & dat$OC[i]==0) { # three people
      threshold[i] <- 19642
    } else if (dat$SPORDER[i]==3 & dat$OC[i]==1) {
      threshold[i] <- (20212+20231)/2
    } else if (dat$SPORDER[i]==3) { # OC is NA value
      threshold[i] <- 19985
    } else if (dat$SPORDER[i]==4 & dat$OC[i]==0) { # four people
      threshold[i] <- 25900
    } else if (dat$SPORDER[i]==4 & dat$OC[i]==1) {
      threshold[i] <- (26324+25465+25554)/3
    } else if (dat$SPORDER[i]==4) { # OC is NA value
      threshold[i] <- 25701
    } else if (dat$SPORDER[i]==5 & dat$OC[i]==0) { # five people
      threshold[i] <- 31234
    } else if (dat$SPORDER[i]==5 & dat$OC[i]==1) {
      threshold[i] <- (31689+30718+29967+29509)/4
    } else if (dat$SPORDER[i]==5) { # OC is NA value
      threshold[i] <- 30459
    } else if (dat$SPORDER[i]==6 & dat$OC[i]==0) { # six people
      threshold[i] <- 35925
    } else if (dat$SPORDER[i]==6 & dat$OC[i]==1) {
      threshold[i] <- (36068+35324+34612+33553+32925)/5
    } else if (dat$SPORDER[i]==6) { # OC is NA value
      threshold[i] <- 34533
    } else if (dat$SPORDER[i]==7 & dat$OC[i]==0) { # seven people
      threshold[i] <- 41336
    } else if (dat$SPORDER[i]==7 & dat$OC[i]==1) {
      threshold[i] <- (4159+40705+40085+38929+37581+36102)/6
    } else if (dat$SPORDER[i]==7) { # OC is NA value
      threshold[i] <- 39194
    } else if (dat$SPORDER[i]==8 & dat$OC[i]==0) { # eight people
      threshold[i] <- 46231
    } else if (dat$SPORDER[i]==8 & dat$OC[i]==1) {
      threshold[i] <- (46640+45800+45064+44021+42696+41317+40967)/7
    } else if (dat$SPORDER[i]==8) { # OC is NA value
      threshold[i] <- 43602
    } else if (dat$SPORDER[i]>=9 & dat$OC[i]==0) { # nine or more people
      threshold[i] <- 55613
    } else if (dat$SPORDER[i]>=9 & dat$OC[i]==1) {
      threshold[i] <- (55883+55140+54516+53491+52082+50807+50491+48546)/8
    } else if (dat$SPORDER[i]>=9) { # OC is NA value
      threshold[i] <- 51393
    } else {
      threshold[i] <- NA
    }
  } # individually assign poverty threshold
  return(threshold)
}
PovertyThreshold <- getThreshold(PovertyThreshold)
dat$IncomePovertyRatio <- (dat$PERNP + dat$PINCP)/PovertyThreshold

dat$IncomePovertyRatio <- dat$IncomePovertyRatio + 1 + abs(min(dat$IncomePovertyRatio, na.rm=na.omit)) # e
nsure all values are positive
```

```r
# function for catching error in rlasso()
myTryCatch <- function(expr) {
  warn <- err <- NULL
  value <- withCallingHandlers(
    tryCatch(expr, error=function(e) {
      err <<- e
      NULL
    }), warning=function(w) {
      warn <<- w
      invokeRestart("muffleWarning")
    })
  list(error=err)
}
```

A.5 Double LASSO Regression

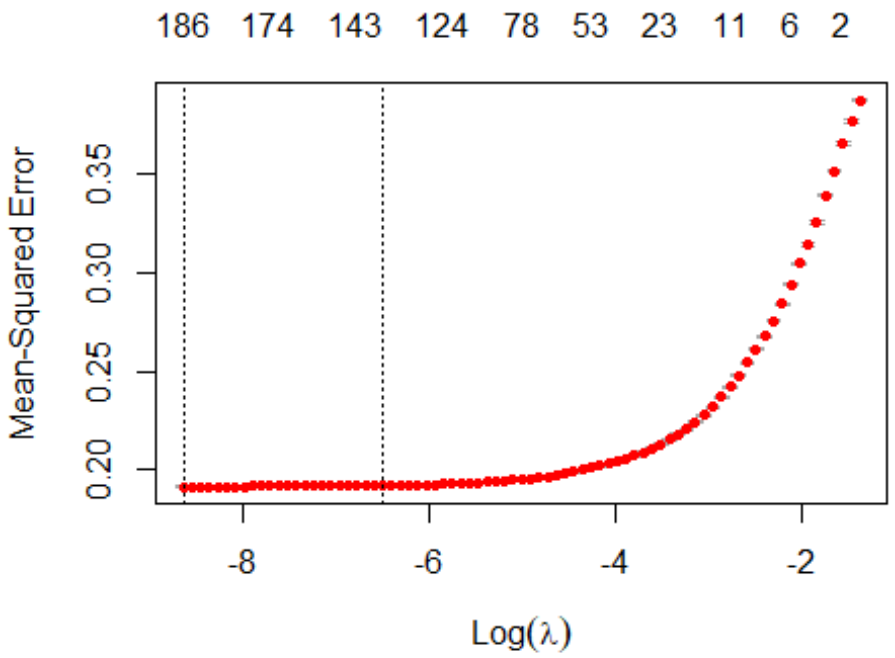
Double ML and Regression

```
## Initial Model 2: Double ML (Double LASSO) ------------------------------------
temp <- dat
dat <- temp # recover original dataset

# DATA CLEANING FOR DOUBLE LASSO
nums <- unlist(lapply(dat, is.factor))
dattemp <- dat[,nums] # all the factors of dataset
# delete variables have >51 factor levels
get1 <- names(which(sapply(dattemp, function(x) length(unique(x))>51)))
delete <- which(names(dat) %in% get1[-16])
dat <- dat[,-c(delete)]
dat <- dat[,-c(9,43:52,82,90,108,95,100)] # delete CITWP,MIL~MLPK,HISP,POVPIP,SFR,RAC1P,RACNUM
dat <- dat[,-c(1,3,76,77)] # delete DIVISION,REGION,PERNP,PINCP
# delete NA rows for Income-Poverty ratio
dat <- dat[!is.na(dat$IncomePovertyRatio),]
dat <- dat[,-c(which(colnames(dat)=="JWRIP"),which(colnames(dat)=="YOEP"))]
# delete NA rows for JWMNP (travel time to work)
dat <- dat[!is.na(dat$JWMNP),]
# delete NA rows for MARHYP: for factor <2 error
dat <- dat[!is.na(dat$MARHYP),]
# delete NA rows for WKHP (usual hours worked per week)
dat <- dat[!is.na(dat$WKHP),]


# delete/add vars that should be deleted/added (from correlation/causal inference)
dat <- dat[, -which(names(dat) %in% c("INTP","OIP","PAP","RETP","SEMP","SSIP","SSP","WAGP"))]   # drop val
ues related to Income
dat <- dat[, -which(names(dat) %in% c("SCH","SCHG"))]   # drop weird educ vars
dat <- dat[, -which(names(dat) %in% c("ANC"))]  # drop ancestry
dat <- droplevels(dat)
#str(dat)
dat <- dat[, -which(names(dat) %in% c("ESR","ESP"))]  # drop meaningless labor vars
dat <- dat[, -which(names(dat) %in% c("HICOV","PRIVCOV","PUBCOV"))]  # drop redundant insurance vars
dat <- dat[, -which(names(dat) %in% c("OC","RC"))]  # drop child vars. Lack data
dat <- dat[, -which(names(dat) %in% c("SFN"))]  # drop "subfamily number"
dat <- dat[, -which(names(dat) %in% c("WRK"))]  # drop "worked last week" (we don't know when is "last wee
k")

dat <- cbind(dat, model.matrix(~(AGEP:HINS3), dat)[,-1])  # age*medicare
dat <- cbind(dat, model.matrix(~(SCIENGP:SCHL), dat)[,-1])  # stem degree*attained degree
dat <- cbind(dat, model.matrix(~(SCIENGRLP:SCHL), dat)[,-1])  # stem related degree*attained degree
dat$VETERAN <- ifelse((dat$DRATX %in% c("1","2")), 1, 0)  # veteran or not
dat <- dat[, -which(names(dat) %in% c("DRATX","VPS","DRAT"))]  # drop veteran related vars
dat <- cbind(dat, model.matrix(~(AGEP:VETERAN), dat)[,2])  # age*veteran or not
names(dat)[ncol(dat)] <- "AGEP_VETERAN"
dat <- cbind(dat, model.matrix(~(AGEP:GCL), dat)[,-1])  # age*grandparent living with grandchild
dat <- cbind(dat, model.matrix(~(AGEP:GCR), dat)[,-1])  # age*grandparent responsible grandchild


# logical dummy for ST==POWSP
dat$SameResidenceWorkplace <- (as.numeric(dat$ST)==as.numeric(dat$POWSP))
dat <- dat[,-which(names(dat) %in% c("ST","POWSP"))]  # delete ST,POWSP
# get which variables have <2 factor levels
get2 <- which(sapply(dat, function(x) length(unique(x)))<2)
dat <- dat[,-get2]
names(dat) <- str_replace(names(dat), ":", "_") # reformat interaction term names

# STEP 1: FIRST LASSO: LOG(IncPovRatio) on ALL POTENTIAL VARIATES (i.e. y on focals)
varnames <- paste(c(names(dat[,-c(which(names(dat) %in%
                                        c("IncomePovertyRatio","SameResidenceWorkplace",
                                          "JWMNP","JWTR")))])), collapse = "+")
# throw in everything and see what happens with this LASSO
formula <- paste(c("log(IncomePovertyRatio)",varnames), collapse = "~")

# Split data into train and test for K-fold CV
set.seed(497)
```

```r
train <- sample(1:nrow(dat), nrow(dat)*0.8)  # 80% for training

# get which variables have <2 factor levels AFTER SUBSETTING
get <- which(sapply(dat[train,], function(x) length(unique(x))<2))

# get train and test datasets
# takeout intercept
xtrain <- model.matrix(as.formula(formula), data = dat[train,])[,-1]
ytrain <- log(dat[train,]$IncomePovertyRatio)

# cross validation then fit LASSO
cv.lasso.1 <- cv.glmnet(xtrain, ytrain, alpha = 1)  # 1 for lasso
cv.lambda.1 <- cv.lasso.1$lambda.min  # get smallest tuning parameter
cv.lambda.1
```

```
## [1] 0.0001775132
```

```r
plot(cv.lasso.1)
```

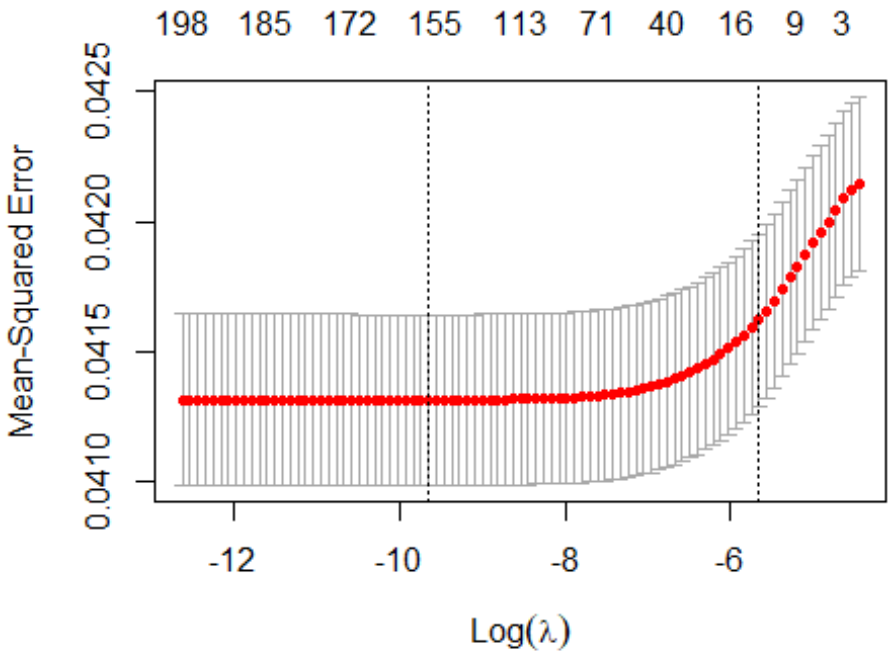



```r
# run lasso and get the necessary focal variables
dlasso.1 <- rlasso(formula, data = dat[train,],
                   lambda.start = cv.lambda.1, post = F)
#summary(lasso.2, all = F)
control <- which(coef(dlasso.1)[-1]!=0)
length(control)
```

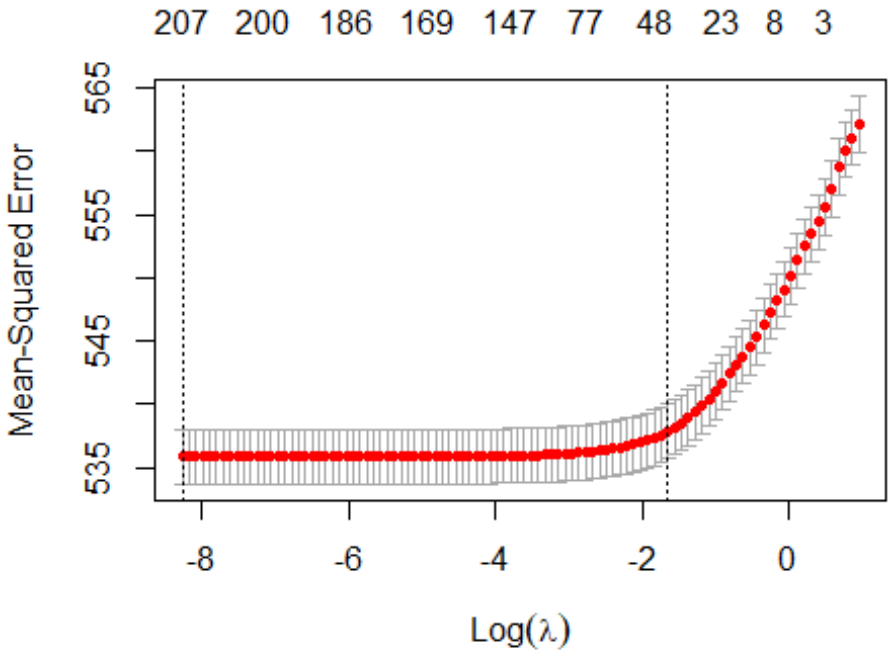
```
## [1] 157
```

```r
# STEP 2: SECOND LASSO: Core vars on ALL POTENTIAL VARIATES (i.e. controls on focals)
formula2.1 <- paste(c("SameResidenceWorkplace",varnames), collapse = "~")
# k-fold cv
xtrain <- model.matrix(as.formula(formula2.1), data = dat[train,])[,-1]
ytrain <- dat[train,]$SameResidenceWorkplace
cv.lasso.2.1 <- cv.glmnet(xtrain, ytrain, alpha = 1)  # 1 for lasso
```

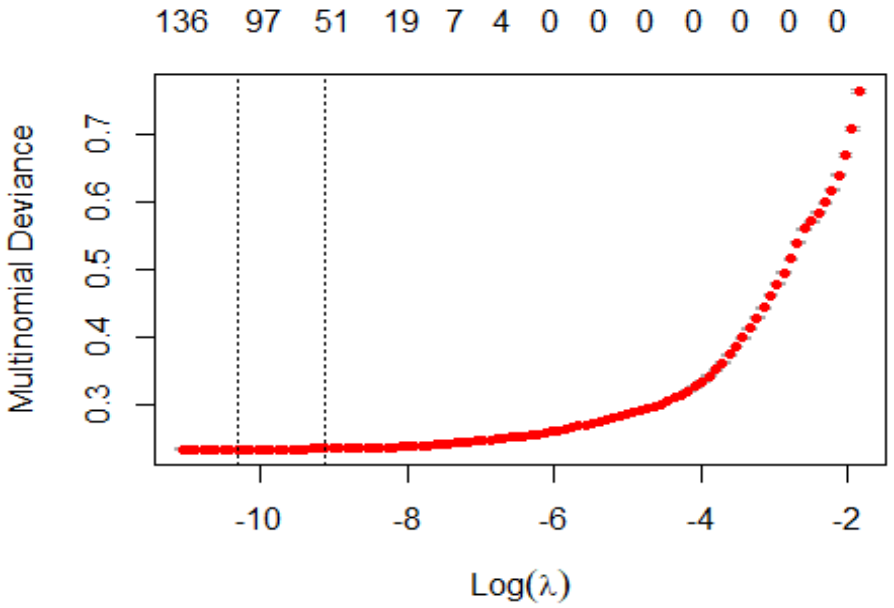
```
cv.lambda.2.1 <- cv.lasso.2.1$lambda.min  # get smallest tuning parameter
cv.lambda.2.1

## [1] 6.405355e-05

plot(cv.lasso.2.1)
```




```
# lasso
dlasso.2.1 <- rlasso(formula2.1, data = dat[train,],
                     lambda.start = cv.lambda.2.1, post = F)
#summary(dlasso.2.1, all = F)
focal1 <- which(coef(dlasso.2.1)[-1]!=0)
length(focal1)

## [1] 74

# travel time
formula2.2 <- paste(c("JWMNP",varnames), collapse = "~")
# k-fold cv
xtrain <- model.matrix(as.formula(formula2.2), data = dat[train,])[,-1]
ytrain <- dat[train,]$JWMNP
cv.lasso.2.2 <- cv.glmnet(xtrain, ytrain, alpha = 1)  # 1 for lasso
cv.lambda.2.2 <- cv.lasso.2.2$lambda.min  # get smallest tuning parameter
cv.lambda.2.2

## [1] 0.0002590354

plot(cv.lasso.2.2)
```




```
# lasso
dlasso.2.2 <- rlasso(formula2.2, data = dat[train,],
                     lambda.start = cv.lambda.2.2, post = F)
#summary(dlasso.2.2, all = F)
focal2 <- which(coef(dlasso.2.2)[-1]!=0)
length(focal2)

## [1] 81

formula2.4 <- paste(c("JWTR",varnames), collapse = "~")
# k-fold cv
xtrain <- model.matrix(as.formula(formula2.4), data = dat[train,])[,-1]
ytrain <- drop.levels(dat[train,]$JWTR) # factor level "11" has 0 observations, use drop.levels()
cv.lasso.2.4 <- cv.glmnet(xtrain, ytrain, alpha = 1, family = "multinomial", nfolds = 3) # 1 for lasso
cv.lambda.2.4 <- cv.lasso.2.4$lambda.min  # get smallest tuning parameter
cv.lambda.2.4

## [1] 3.31276e-05

plot(cv.lasso.2.4)
```




```
# Lasso
tempdat2 <- fastDummies::dummy_cols(dat)
focal4 <- c()
for (ii in 1:11) {
  if (ii<10) {
    formula2.4.n <- paste(c(paste("JWTR_0",ii,sep=""),varnames), collapse = "~")
    dlasso.2.4.n <- rlasso(formula2.4.n, data = tempdat2[train,],
                           lambda.start = cv.lambda.2.4, post = F)
    focal4.n <- which(coef(dlasso.2.4.n)[-1]!=0)
    focal4 <- unique(c(focal4, names(focal4.n)))
  } else if (ii==10) {
    formula2.4.n <- paste(c("JWTR_10",varnames), collapse = "~")
    dlasso.2.4.n <- rlasso(formula2.4.n, data = tempdat2[train,],
                           lambda.start = cv.lambda.2.4, post = F)
    focal4.n <- which(coef(dlasso.2.4.n)[-1]!=0)
    focal4 <- unique(c(focal4, names(focal4.n)))
  } else if (ii==11) {
    formula2.4.n <- paste(c("JWTR_12",varnames), collapse = "~")
    dlasso.2.4.n <- rlasso(formula2.4.n, data = tempdat2[train,],
                           lambda.start = cv.lambda.2.4, post = F)
    focal4.n <- which(coef(dlasso.2.4.n)[-1]!=0)
    focal4 <- unique(c(focal4, names(focal4.n)))
  }
}
length(focal4)

## [1] 213

# STEP 3: Take union of all remainder potential variates
union <- c(names(control), names(focal1), names(focal2), focal4)
if (any(duplicated(union))==T) {
  union <- unique(union)
}

# Total number of feature variables kept from Double Lasso
length(union)

## [1] 219
```

```r
# STEP 3 Continued: do OLS of y on focals and kept potential variates
unionf <- paste(c("SameResidenceWorkplace*JWMNP+JWTR*JWMNP",union), collapse = "+")
formula <- paste(c("log(IncomePovertyRatio)", unionf), collapse = "~")

# name all extra variables created from doing LASSO
dattemp <- dat
for (i in 1:500) { # look at formula and count how many new vars need to be made
  error <- myTryCatch(olsDLasso1<- lm(formula, data = dattemp)) # CAUTION
  newvars <- substr(error[[1]], 45, str_length(error[[1]])-12)
  existingvars <- names(dattemp)[which(str_detect(newvars, names(dattemp)))]
  existingchars <- sub(existingvars, "", newvars)
  dattemp[,newvars] <- dattemp[,which(names(dattemp)==existingvars)]==existingchars
}

# start with declaring the new vars
which(colSums(is.na(dattemp))==nrow(dattemp))

## SCIENGRLP1 SCIENGRLP2
##        238        239

dattemp$SCIENGRLP1 <- dattemp$SCIENGRLP == "1"
dattemp$SCIENGRLP2 <- dattemp$SCIENGRLP == "2"

which(lapply(dattemp, class)=="matrix")

## MARHD2 MARHT2 MARHT3 MARHD8
##    160    162    163    244

dattemp$MARHT3 <- dattemp$MARHT == "3"
dattemp$MARHD2 <- dattemp$MARHD == "2"
dattemp$MARHD8 <- dattemp$MARHD == "8"
dattemp$MARHT2 <- dattemp$MARHT == "2"

# multicolinearity: get which variables have <2 unique values
multicol <- names(which(sapply(dattemp[train,], function(x) length(unique(x))<2)))
# manually delete some of the rest (NA values in summary of lm, multicollinearity)
multicol <- c(multicol,
              "MSP3","MSP4","MSP5","ENG1","SCHL21","DRIVESP6",
              "NATIVITY2","SCHL18","DECADE6","WAOB4")
union <- union[-which(union %in% multicol)] # delete them from formula
aliased <- which(summary(lm(formula, data = dattemp[train,]))$aliased)
union <- union[-which(union %in% names(aliased))]

# rewrite formula for OLS
unionf <- paste(c("SameResidenceWorkplace*JWMNP+JWTR*JWMNP",union), collapse = "+")
formula <- paste(c("log(IncomePovertyRatio)", unionf), collapse = "~")

# Training OLS regression post double LASSO
olsDLasso1 <- lm(formula, data = dattemp[train,])
DMLresult <- summary(olsDLasso1)
DMLresult

##
## Call:
## lm(formula = formula, data = dattemp[train, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0233 -0.2576 -0.0282  0.2212  3.5316
##
## Coefficients:
##                                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                    1.918e+00  1.461e-01   13.124  < 2e-16 ***
## SameResidenceWorkplaceTRUE    -8.169e-02  4.054e-03  -20.149  < 2e-16 ***
## JWMNP                          1.431e-03  7.106e-05   20.141  < 2e-16 ***
## JWTR02                         1.093e-02  1.481e-02    0.737 0.460822
## JWTR03                         3.399e-02  4.974e-02    0.683 0.494321
## JWTR04                         2.488e-01  1.612e-02   15.434  < 2e-16 ***
```

```
## JWTR05                        3.626e-01  1.985e-02   18.263  < 2e-16 ***
## JWTR06                        3.288e-01  4.727e-02    6.955 3.53e-12 ***
## JWTR07                        8.782e-02  2.524e-02    3.479 0.000503 ***
## JWTR08                       -2.325e-02  2.305e-02   -1.009 0.313043
## JWTR09                       -5.182e-02  1.791e-02   -2.894 0.003805 **
## JWTR10                       -6.349e-02  1.350e-02   -4.705 2.54e-06 ***
## JWTR12                        9.491e-03  1.469e-02    0.646 0.518130
## SPORDER                      -6.923e-02  1.313e-03  -52.723  < 2e-16 ***
## PWGTP                         4.407e-05  6.578e-06    6.700 2.09e-11 ***
## AGEP                          4.456e-03  4.849e-04    9.191  < 2e-16 ***
## CIT3TRUE                     -3.097e-02  6.628e-03   -4.672 2.98e-06 ***
## CIT4TRUE                      9.808e-03  4.480e-03    2.189 0.028572 *
## CIT5TRUE                     -1.488e-02  4.987e-03   -2.983 0.002857 **
## COW2TRUE                     -8.112e-02  1.805e-03  -44.949  < 2e-16 ***
## COW3TRUE                     -8.743e-02  1.869e-03  -46.793  < 2e-16 ***
## COW4TRUE                     -1.141e-01  2.277e-03  -50.089  < 2e-16 ***
## COW5TRUE                      6.980e-02  2.941e-03   23.732  < 2e-16 ***
## COW6TRUE                     -1.073e-01  2.194e-03  -48.883  < 2e-16 ***
## COW7TRUE                      9.960e-02  2.500e-03   39.844  < 2e-16 ***
## COW8TRUE                     -1.725e-01  1.103e-02  -15.645  < 2e-16 ***
## DDRS2TRUE                    -2.697e-02  8.954e-03   -3.012 0.002594 **
## DEYE2TRUE                     2.110e-02  5.392e-03    3.913 9.13e-05 ***
## DPHY2TRUE                     2.578e-02  5.036e-03    5.119 3.07e-07 ***
## DREM2TRUE                     3.432e-02  5.451e-03    6.296 3.06e-10 ***
## ENG2TRUE                     -9.786e-02  2.844e-03  -34.408  < 2e-16 ***
## ENG3TRUE                     -1.272e-01  3.656e-03  -34.784  < 2e-16 ***
## ENG4TRUE                     -1.138e-01  6.350e-03  -17.920  < 2e-16 ***
## FER1TRUE                      4.442e-02  5.013e-03    8.861  < 2e-16 ***
## FER2TRUE                      2.342e-02  2.036e-03   11.503  < 2e-16 ***
## GCL2TRUE                      2.570e-01  2.083e-02   12.342  < 2e-16 ***
## GCR2TRUE                      8.356e-02  4.472e-02    1.869 0.061677 .
## HINS12TRUE                   -1.520e-01  1.500e-03 -101.372  < 2e-16 ***
## HINS22TRUE                   -1.308e-02  1.728e-03   -7.569 3.76e-14 ***
## HINS42TRUE                    9.252e-02  2.248e-03   41.164  < 2e-16 ***
## HINS52TRUE                   -2.228e-02  3.320e-03   -6.711 1.94e-11 ***
## HINS62TRUE                    5.436e-02  3.758e-03   14.465  < 2e-16 ***
## HINS72TRUE                    3.229e-03  8.499e-03    0.380 0.703987
## LANX2TRUE                     1.159e-02  2.012e-03    5.758 8.50e-09 ***
## MAR2TRUE                     -2.623e-02  3.210e-03   -8.173 3.02e-16 ***
## MAR3TRUE                     -4.197e-02  1.670e-03  -25.131  < 2e-16 ***
## MAR4TRUE                     -4.869e-02  3.359e-03  -14.493  < 2e-16 ***
## MARHD2TRUE                   -2.075e-02  4.490e-03   -4.621 3.81e-06 ***
## MARHT2TRUE                   -5.534e-03  1.472e-03   -3.758 0.000171 ***
## MARHT3TRUE                   -2.639e-02  2.677e-03   -9.860  < 2e-16 ***
## MARHYP                       -3.978e-04  6.967e-05   -5.710 1.13e-08 ***
## MIG2TRUE                     -5.720e-02  8.815e-03   -6.489 8.64e-11 ***
## MIG3TRUE                     -1.420e-02  1.707e-03   -8.323  < 2e-16 ***
## NWAB2TRUE                    -2.978e-02  9.740e-03   -3.057 0.002234 **
## NWAV5TRUE                     7.657e-03  5.268e-03    1.454 0.146062
## NWLA3TRUE                     3.213e-03  1.069e-02    0.301 0.763636
## NWLK3TRUE                     8.442e-02  7.974e-03   10.586  < 2e-16 ***
## NWRE2TRUE                     6.009e-02  1.167e-02    5.149 2.62e-07 ***
## RELP01TRUE                   -1.579e-01  1.740e-03  -90.721  < 2e-16 ***
## RELP02TRUE                   -2.422e-01  4.139e-03  -58.517  < 2e-16 ***
## RELP03TRUE                   -2.222e-01  2.175e-02  -10.218  < 2e-16 ***
## RELP04TRUE                   -2.501e-01  1.523e-02  -16.422  < 2e-16 ***
## RELP05TRUE                   -2.344e-01  7.721e-03  -30.357  < 2e-16 ***
## RELP06TRUE                   -2.428e-01  7.102e-03  -34.192  < 2e-16 ***
## RELP07TRUE                   -2.017e-01  1.581e-02  -12.757  < 2e-16 ***
## RELP08TRUE                   -3.037e-01  1.467e-02  -20.705  < 2e-16 ***
## RELP09TRUE                   -2.874e-01  7.891e-03  -36.425  < 2e-16 ***
## RELP10TRUE                   -2.442e-01  8.130e-03  -30.035  < 2e-16 ***
## RELP11TRUE                   -2.412e-01  1.051e-02  -22.950  < 2e-16 ***
## RELP12TRUE                   -2.277e-01  7.150e-03  -31.849  < 2e-16 ***
## RELP13TRUE                   -2.062e-01  4.674e-03  -44.118  < 2e-16 ***
## RELP15TRUE                   -2.442e-01  7.903e-03  -30.895  < 2e-16 ***
## RELP17TRUE                   -2.786e-01  1.332e-02  -20.911  < 2e-16 ***
## SCHL04TRUE                   -1.043e-01  3.053e-02   -3.416 0.000635 ***
## SCHL05TRUE                   -1.342e-01  2.178e-02   -6.160 7.27e-10 ***
```

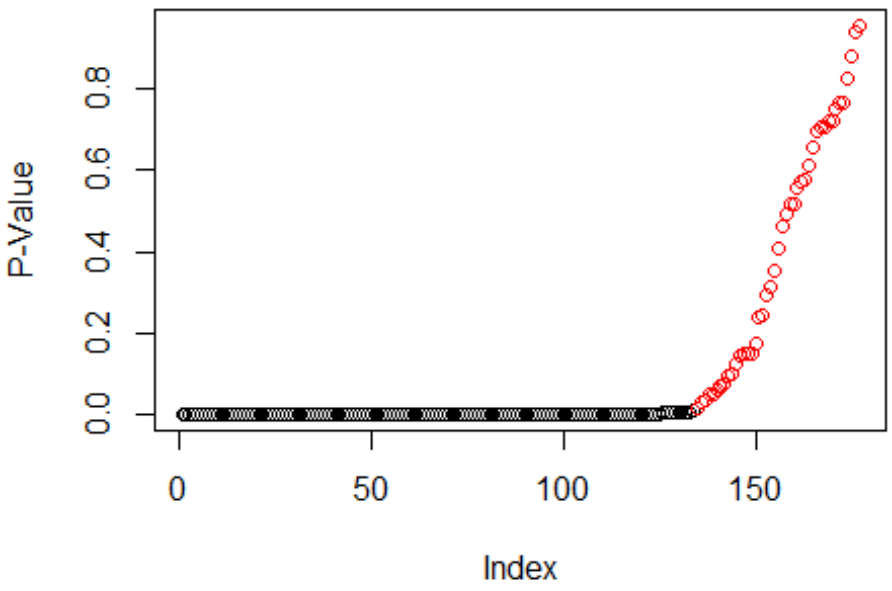
```
## SCHL06TRUE                  -1.007e-01  1.467e-02   -6.863 6.73e-12 ***
## SCHL07TRUE                  -1.055e-01  1.708e-02   -6.175 6.61e-10 ***
## SCHL08TRUE                  -1.064e-01  1.370e-02   -7.771 7.82e-15 ***
## SCHL09TRUE                  -1.012e-01  6.870e-03  -14.725  < 2e-16 ***
## SCHL10TRUE                  -1.151e-01  1.224e-02   -9.400  < 2e-16 ***
## SCHL11TRUE                  -8.077e-02  7.093e-03  -11.386  < 2e-16 ***
## SCHL12TRUE                  -1.061e-01  5.992e-03  -17.709  < 2e-16 ***
## SCHL13TRUE                  -1.301e-01  5.561e-03  -23.397  < 2e-16 ***
## SCHL14TRUE                  -1.185e-01  5.176e-03  -22.892  < 2e-16 ***
## SCHL15TRUE                  -9.016e-02  4.420e-03  -20.397  < 2e-16 ***
## SCHL16TRUE                  -6.246e-02  2.171e-03  -28.772  < 2e-16 ***
## SCHL17TRUE                  -8.702e-02  3.278e-03  -26.545  < 2e-16 ***
## SCHL19TRUE                   3.397e-02  2.304e-03   14.741  < 2e-16 ***
## SCHL20TRUE                   6.445e-02  2.424e-03   26.591  < 2e-16 ***
## SCHL22TRUE                   1.233e-01  2.478e-03   49.751  < 2e-16 ***
## SCHL23TRUE                   3.437e-01  8.596e-03   39.985  < 2e-16 ***
## SCHL24TRUE                   2.175e-01  6.802e-03   31.977  < 2e-16 ***
## SEX2TRUE                    -7.976e-02  2.289e-02   -3.484 0.000493 ***
## WKHP                         1.365e-02  4.765e-05  286.546  < 2e-16 ***
## WKW2TRUE                    -7.342e-02  3.505e-03  -20.948  < 2e-16 ***
## WKW3TRUE                    -1.582e-01  2.339e-03  -67.629  < 2e-16 ***
## WKW4TRUE                    -2.789e-01  2.820e-03  -98.897  < 2e-16 ***
## WKW5TRUE                    -4.105e-01  3.867e-03 -106.155  < 2e-16 ***
## WKW6TRUE                    -5.210e-01  3.940e-03 -132.240  < 2e-16 ***
## DECADE3TRUE                  3.593e-02  6.193e-03    5.801 6.58e-09 ***
## DECADE4TRUE                  2.350e-02  4.386e-03    5.359 8.37e-08 ***
## DECADE7TRUE                 -2.040e-02  3.331e-03   -6.125 9.10e-10 ***
## DECADE8TRUE                 -5.498e-02  4.133e-03  -13.302  < 2e-16 ***
## DIS2TRUE                     4.190e-02  4.841e-03    8.655  < 2e-16 ***
## DRIVESP1TRUE                -1.330e-02  1.269e-02   -1.049 0.294402
## DRIVESP2TRUE                -6.237e-02  1.281e-02   -4.870 1.12e-06 ***
## DRIVESP3TRUE                -5.937e-02  1.334e-02   -4.451 8.55e-06 ***
## DRIVESP4TRUE                -5.789e-02  1.438e-02   -4.025 5.70e-05 ***
## DRIVESP5TRUE                -3.096e-02  1.578e-02   -1.962 0.049801 *
## MSP2TRUE                    -1.459e-02  3.143e-03   -4.642 3.45e-06 ***
## PAOC1TRUE                   -9.965e-02  2.304e-02   -4.324 1.53e-05 ***
## PAOC2TRUE                   -1.329e-01  2.291e-02   -5.803 6.53e-09 ***
## PAOC4TRUE                   -1.321e-01  2.288e-02   -5.773 7.80e-09 ***
## QTRBIR3TRUE                  3.229e-03  1.189e-03    2.715 0.006626 **
## RACAIAN1TRUE                -2.852e-02  5.081e-03   -5.612 2.00e-08 ***
## RACASN1TRUE                  6.301e-02  4.437e-03   14.203  < 2e-16 ***
## RACBLK1TRUE                 -6.564e-02  4.108e-03  -15.980  < 2e-16 ***
## RACPI1TRUE                  -2.557e-02  1.224e-02   -2.088 0.036787 *
## RACWHT1TRUE                  2.257e-02  3.879e-03    5.817 6.00e-09 ***
## SCIENGRLP1TRUE               3.276e-01  3.554e-03   92.186  < 2e-16 ***
## SCIENGRLP2TRUE               2.506e-01  2.330e-03  107.536  < 2e-16 ***
## WAOB2TRUE                    1.641e-02  5.217e-02    0.315 0.753068
## WAOB3TRUE                   -1.879e-02  4.167e-03   -4.508 6.54e-06 ***
## WAOB5TRUE                    6.072e-02  4.846e-03   12.528  < 2e-16 ***
## WAOB6TRUE                   -2.510e-02  6.657e-03   -3.771 0.000163 ***
## WAOB7TRUE                    1.233e-01  8.684e-03   14.196  < 2e-16 ***
## WAOB8TRUE                    1.105e-01  1.535e-02    7.198 6.13e-13 ***
## AGEP_HINS31                  1.562e-03  2.448e-04    6.382 1.75e-10 ***
## SCIENGP_SCHL01              -9.545e-02  5.379e-03  -17.745  < 2e-16 ***
## SCIENGP1_SCHL22              1.123e-01  3.115e-03   36.039  < 2e-16 ***
## SCIENGP1_SCHL23              2.355e-01  5.968e-03   39.456  < 2e-16 ***
## SCIENGP1_SCHL24              1.742e-01  7.863e-03   22.154  < 2e-16 ***
## SCIENGRLP1_SCHL22            7.894e-03  5.833e-03    1.353 0.175985
## SCIENGRLP2_SCHL23            1.390e-02  9.687e-03    1.435 0.151276
## SCIENGRLP1_SCHL24            7.851e-02  1.269e-02    6.188 6.08e-10 ***
## AGEP_VETERAN                 5.122e-04  1.336e-04    3.834 0.000126 ***
## AGEP_GCL                     6.289e-03  9.141e-04    6.880 5.99e-12 ***
## DOUT2TRUE                   -7.836e-03  6.693e-03   -1.171 0.241662
## MARHD8TRUE                  -5.994e-03  1.348e-02   -0.445 0.656588
## NWAB3TRUE                    8.029e-04  1.048e-02    0.077 0.938949
## RACNH1TRUE                  -1.854e-03  1.226e-02   -0.151 0.879789
## RACSOR1TRUE                  8.821e-03  4.479e-03    1.969 0.048897 *
## AGEP_GCL2                    6.915e-04  4.779e-04    1.447 0.147896
## CIT2TRUE                    -4.371e-02  5.260e-02   -0.831 0.406013
```

```
## DEAR2TRUE                          1.772e-03  5.003e-03    0.354 0.723115
## GCL1TRUE                           1.786e-01  4.101e-02    4.355 1.33e-05 ***
## NWLA2TRUE                         -1.020e-02  1.103e-02   -0.925 0.354919
## DECADE5TRUE                        1.040e-03  3.506e-03    0.297 0.766786
## SCIENGP1_SCHL21                    7.754e-02  2.292e-03   33.827  < 2e-16 ***
## VETERAN                           -1.758e-02  7.315e-03   -2.404 0.016228 *
## GCM1TRUE                           1.106e-02  1.711e-02    0.646 0.518256
## GCM2TRUE                          -6.377e-03  1.695e-02   -0.376 0.706680
## GCM4TRUE                           2.230e-02  1.348e-02    1.654 0.098059 .
## HINS32TRUE                         4.964e-02  1.664e-02    2.983 0.002857 **
## NWAV3TRUE                         -2.263e-02  8.308e-03   -2.723 0.006466 **
## SCHL02TRUE                        -3.983e-02  3.363e-02   -1.184 0.236227
## SCHL03TRUE                        -6.848e-02  3.793e-02   -1.805 0.071006 .
## DECADE1TRUE                        1.600e-02  3.138e-02    0.510 0.610265
## DECADE2TRUE                        5.866e-03  1.042e-02    0.563 0.573573
## PAOC3TRUE                         -1.105e-01  2.305e-02   -4.795 1.63e-06 ***
## AGEP_GCR1                          1.164e-03  7.613e-04    1.529 0.126288
## GCM3TRUE                           2.665e-03  1.223e-02    0.218 0.827452
## NWAV2TRUE                         -8.294e-03  1.418e-02   -0.585 0.558620
## NWLK2TRUE                          7.811e-02  7.118e-03   10.973  < 2e-16 ***
## NWRE3TRUE                          3.179e-02  1.178e-02    2.698 0.006973 **
## QTRBIR2TRUE                        2.164e-03  1.219e-03    1.776 0.075757 .
## SameResidenceWorkplaceTRUE:JWMNP   4.053e-06  7.287e-05    0.056 0.955644
## JWMNP:JWTR02                      -4.712e-04  1.278e-04   -3.688 0.000226 ***
## JWMNP:JWTR03                      -4.075e-04  1.036e-03   -0.393 0.693954
## JWMNP:JWTR04                      -3.007e-03  1.743e-04  -17.252  < 2e-16 ***
## JWMNP:JWTR05                      -1.646e-03  1.913e-04   -8.605  < 2e-16 ***
## JWMNP:JWTR06                      -1.650e-03  6.101e-04   -2.704 0.006852 **
## JWMNP:JWTR07                      -1.013e-03  7.045e-04   -1.439 0.150256
## JWMNP:JWTR08                       3.503e-04  6.298e-04    0.556 0.578017
## JWMNP:JWTR09                       1.478e-03  4.447e-04    3.325 0.000885 ***
## JWMNP:JWTR10                       4.182e-04  2.490e-04    1.680 0.093044 .
## JWMNP:JWTR12                      -4.323e-05  1.213e-04   -0.356 0.721470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.435 on 781192 degrees of freedom
## Multiple R-squared:  0.5123, Adjusted R-squared:  0.5122
## F-statistic:  4662 on 176 and 781192 DF,  p-value: < 2.2e-16

# Test Prediction
pred.olsDLasso.1 <- predict(olsDLasso1, newdata = dattemp[-train,])
summary(pred.olsDLasso.1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.3127  1.9878  2.2644  2.2562  2.5408  4.2571

length(na.omit(pred.olsDLasso.1)) # count remaining observations

## [1] 195343

# test error
mse.1 <- mean((pred.olsDLasso.1-log(dattemp[-train,]$IncomePovertyRatio))^2, na.rm=T)
mse.1

## [1] 0.1894029
```

Result 1: Analysis & Hypothesis Testing

```
# 3 Ways of getting Test R2
y <- log(dattemp[-train,]$IncomePovertyRatio)-mean(log(dattemp[-train,]$IncomePovertyRatio))
yhat <- pred.olsDLasso.1-mean(pred.olsDLasso.1)
u <- y - yhat
# 1:
# R2 = yhat*yhat/yTy
r2_1 <- (yhat %*% yhat)/(y %*% y)
r2_1
```

```
##           [,1]
## [1,] 0.5125579

# 2:
# R2 = 1- SSR/SST = 1- uTu/yTy
r2_2 <- 1 - (u %*% u)/(y %*% y)
r2_2

##           [,1]
## [1,] 0.5089613

# 3:
# R2 = corr(y, yhat)^2, "fair r-squared"
r2_3 <- cor.test(y, yhat, use = "complete.obs")
# now, square the correlation coefficient
r2_3

##
##  Pearson's product-moment correlation
##
## data:  y and yhat
## t = 449.97, df = 195341, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7112352 0.7155903
## sample estimates:
##       cor
## 0.7134197

r2_3$estimate^2

##       cor
## 0.5089676

# False Discovery Rate control
p <- as.data.frame(DMLresult$coefficients[,4])
sigcode <- cut(p[,1], breaks = c(-Inf, 0.001, 0.01, 0.05, 0.1, 1),
               labels = c("***", "**", "*", ".", " "))
p$"" <- sigcode

# sort by increasing p-value
p <- p[order(p$`DMLresult$coefficients[, 4]`),]
p$BY <- 0
m <- nrow(p)
Q = 0.10  # 10%
cm=0
for (ii in 1:m) {
  cm = cm + 1/ii
  p[ii,3] <- ii/m/cm*Q
}
noreject <- (!(p[,1] < p[,3]))
plot(p$`DMLresult$coefficients[, 4]`,ylab="P-Value", col = ifelse(noreject,'red','black'))
```




```
noreject <- which(noreject)
p <- p[noreject,]   # these one's we cannot reject the null
names(p) <- c("p-value","Sig. Level","BY Stat")
p
```

```
##                                  p-value Sig. Level    BY Stat
## VETERAN                       0.01622773           * 0.01390240
## CIT4TRUE                      0.02857226           * 0.01398663
## RACPI1TRUE                    0.03678683           * 0.01407078
## RACSOR1TRUE                   0.04889671           * 0.01415484
## DRIVESP5TRUE                  0.04980121           * 0.01423881
## GCR2TRUE                      0.06167725           . 0.01432270
## SCHL03TRUE                    0.07100579           . 0.01440650
## QTRBIR2TRUE                   0.07575731           . 0.01449022
## JWMNP:JWTR10                  0.09304410           . 0.01457386
## GCM4TRUE                      0.09805948           . 0.01465741
## AGEP_GCR1                     0.12628818             0.01474089
## NWAV5TRUE                     0.14606246             0.01482428
## AGEP_GCL2                     0.14789649             0.01490759
## JWMNP:JWTR07                  0.15025585             0.01499082
## SCIENGRLP2_SCHL23             0.15127602             0.01507397
## SCIENGRLP1_SCHL22             0.17598520             0.01515704
## SCHL02TRUE                    0.23622737             0.01524004
## DOUT2TRUE                     0.24166238             0.01532296
## DRIVESP1TRUE                  0.29440172             0.01540580
## JWTR08                        0.31304290             0.01548857
## NWLA2TRUE                     0.35491891             0.01557126
## CIT2TRUE                      0.40601337             0.01565387
## JWTR02                        0.46082218             0.01573642
## JWTR03                        0.49432147             0.01581889
## JWTR12                        0.51813036             0.01590128
## GCM1TRUE                      0.51825607             0.01598361
## NWAV2TRUE                     0.55862008             0.01606586
## DECADE2TRUE                   0.57357341             0.01614804
## JWMNP:JWTR08                  0.57801715             0.01623016
## DECADE1TRUE                   0.61026469             0.01631220
## MARHD8TRUE                    0.65658784             0.01639417
```

```
## JWMNP:JWTR03                      0.69395359           0.01647607
## HINS72TRUE                        0.70398697           0.01655791
## GCM2TRUE                          0.70667974           0.01663968
## JWMNP:JWTR12                      0.72146983           0.01672138
## DEAR2TRUE                         0.72311468           0.01680301
## WAOB2TRUE                         0.75306776           0.01688458
## NWLA3TRUE                         0.76363569           0.01696608
## DECADE5TRUE                       0.76678592           0.01704751
## GCM3TRUE                          0.82745213           0.01712888
## RACNH1TRUE                        0.87978946           0.01721019
## NWAB3TRUE                         0.93894933           0.01729143
## SameResidenceWorkplaceTRUE:JWMNP 0.95564425           0.01737261
```

```
# get BY-adjusted p-values
pBY <- as.data.frame(p.adjust(p[,1], method = "BY"))   #Benjamini-Yekutieli
rownames(pBY) <- rownames(p)
adjsigcode <- cut(pBY[,1], breaks = c(-Inf, 0.001, 0.01, 0.05, 0.1, 1),
                  labels = c("***", "**", "*", ".", " "))
pBY$"" <- adjsigcode

# compare p-values for non-rejected
fdr <- cbind.data.frame(p[,c(1,2)], pBY)
colnames(fdr) <- c("Original","Sig. Level", "FDR Adj.","Sig. Level")
fdr
```

```
##                                   Original Sig. Level FDR Adj. Sig. Level
## VETERAN                           0.01622773          *        1
## CIT4TRUE                          0.02857226          *        1
## RACPI1TRUE                        0.03678683          *        1
## RACSOR1TRUE                       0.04889671          *        1
## DRIVESP5TRUE                      0.04980121          *        1
## GCR2TRUE                          0.06167725          .        1
## SCHL03TRUE                        0.07100579          .        1
## QTRBIR2TRUE                       0.07575731          .        1
## JWMNP:JWTR10                      0.09304410          .        1
## GCM4TRUE                          0.09805948          .        1
## AGEP_GCR1                         0.12628818                   1
## NWAV5TRUE                         0.14606246                   1
## AGEP_GCL2                         0.14789649                   1
## JWMNP:JWTR07                      0.15025585                   1
## SCIENGRLP2_SCHL23                 0.15127602                   1
## SCIENGRLP1_SCHL22                 0.17598520                   1
## SCHL02TRUE                        0.23622737                   1
## DOUT2TRUE                         0.24166238                   1
## DRIVESP1TRUE                      0.29440172                   1
## JWTR08                            0.31304290                   1
## NWLA2TRUE                         0.35491891                   1
## CIT2TRUE                          0.40601337                   1
## JWTR02                            0.46082218                   1
## JWTR03                            0.49432147                   1
## JWTR12                            0.51813036                   1
## GCM1TRUE                          0.51825607                   1
## NWAV2TRUE                         0.55862008                   1
## DECADE2TRUE                       0.57357341                   1
## JWMNP:JWTR08                      0.57801715                   1
## DECADE1TRUE                       0.61026469                   1
## MARHD8TRUE                        0.65658784                   1
## JWMNP:JWTR03                      0.69395359                   1
## HINS72TRUE                        0.70398697                   1
## GCM2TRUE                          0.70667974                   1
## JWMNP:JWTR12                      0.72146983                   1
## DEAR2TRUE                         0.72311468                   1
## WAOB2TRUE                         0.75306776                   1
## NWLA3TRUE                         0.76363569                   1
## DECADE5TRUE                       0.76678592                   1
## GCM3TRUE                          0.82745213                   1
## RACNH1TRUE                        0.87978946                   1
## NWAB3TRUE                         0.93894933                   1
## SameResidenceWorkplaceTRUE:JWMNP 0.95564425                   1
```

```r
# BP test for heteroskedasticity
bpres1 <- bptest(olsDLasso1, data = dattemp[-train,]) #reject homoskedasticity if p-value is small
bpres1

##
##  studentized Breusch-Pagan test
##
## data:  olsDLasso1
## BP = 58372, df = 176, p-value < 2.2e-16

# F-test
null = c("SameResidenceWorkplaceTRUE","JWMNP",
         "JWTR02","JWTR03","JWTR04","JWTR05","JWTR06","JWTR07","JWTR08",
         "JWTR09","JWTR10","JWTR12")
if (bpres1$p.value >= 0.001) {   # homoskedastic
  linearHypothesis(olsDLasso1, null, vcov = hccm(olsDLasso1, type = "hc0")) # classical White VCOV
} else {
  linearHypothesis(olsDLasso1, null) # default homoskedastic error
}


## Hypothesis:
## SameResidenceWorkplaceTRUE = 0
## JWMNP = 0
## JWTR02 = 0
## JWTR03 = 0
## JWTR04 = 0
## JWTR05 = 0
## JWTR06 = 0
## JWTR07 = 0
## JWTR08 = 0
## JWTR09 = 0
## JWTR10 = 0
## JWTR12 = 0
##
## Model 1: restricted model
## Model 2: log(IncomePovertyRatio) ~ SameResidenceWorkplace * JWMNP + JWTR *
##     JWMNP + SPORDER + PWGTP + AGEP + CIT3 + CIT4 + CIT5 + COW2 +
##     COW3 + COW4 + COW5 + COW6 + COW7 + COW8 + DDRS2 + DEYE2 +
##     DPHY2 + DREM2 + ENG2 + ENG3 + ENG4 + FER1 + FER2 + GCL2 +
##     GCR2 + HINS12 + HINS22 + HINS42 + HINS52 + HINS62 + HINS72 +
##     LANX2 + MAR2 + MAR3 + MAR4 + MARHD2 + MARHT2 + MARHT3 + MARHYP +
##     MIG2 + MIG3 + NWAB2 + NWAV5 + NWLA3 + NWLK3 + NWRE2 + RELP01 +
##     RELP02 + RELP03 + RELP04 + RELP05 + RELP06 + RELP07 + RELP08 +
##     RELP09 + RELP10 + RELP11 + RELP12 + RELP13 + RELP15 + RELP17 +
##     SCHL04 + SCHL05 + SCHL06 + SCHL07 + SCHL08 + SCHL09 + SCHL10 +
##     SCHL11 + SCHL12 + SCHL13 + SCHL14 + SCHL15 + SCHL16 + SCHL17 +
##     SCHL19 + SCHL20 + SCHL22 + SCHL23 + SCHL24 + SEX2 + WKHP +
##     WKW2 + WKW3 + WKW4 + WKW5 + WKW6 + DECADE3 + DECADE4 + DECADE7 +
##     DECADE8 + DIS2 + DRIVESP1 + DRIVESP2 + DRIVESP3 + DRIVESP4 +
##     DRIVESP5 + MSP2 + PAOC1 + PAOC2 + PAOC4 + QTRBIR3 + RACAIAN1 +
##     RACASN1 + RACBLK1 + RACPI1 + RACWHT1 + SCIENGRLP1 + SCIENGRLP2 +
##     WAOB2 + WAOB3 + WAOB5 + WAOB6 + WAOB7 + WAOB8 + AGEP_HINS31 +
##     SCIENGP_SCHL01 + SCIENGP1_SCHL22 + SCIENGP1_SCHL23 + SCIENGP1_SCHL24 +
##     SCIENGRLP1_SCHL22 + SCIENGRLP2_SCHL23 + SCIENGRLP1_SCHL24 +
##     AGEP_VETERAN + AGEP_GCL + DOUT2 + MARHD8 + NWAB3 + RACNH1 +
##     RACSOR1 + AGEP_GCL2 + CIT2 + DEAR2 + GCL1 + NWLA2 + DECADE5 +
##     SCIENGP1_SCHL21 + VETERAN + GCM1 + GCM2 + GCM4 + HINS32 +
##     NWAV3 + SCHL02 + SCHL03 + DECADE1 + DECADE2 + PAOC3 + AGEP_GCR1 +
##     GCM3 + NWAV2 + NWLK2 + NWRE3 + QTRBIR2
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1 781204 148776
## 2 781192 147823 12    953.35 419.85 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
null = c("JWMNP",
         "JWTR02","JWTR03","JWTR04","JWTR05","JWTR06","JWTR07","JWTR08",
         "JWTR09","JWTR10","JWTR12",
         "JWMNP:JWTR02","JWMNP:JWTR03","JWMNP:JWTR04","JWMNP:JWTR05","JWMNP:JWTR06",
         "JWMNP:JWTR07","JWMNP:JWTR08","JWMNP:JWTR09","JWMNP:JWTR10","JWMNP:JWTR12")
if (bpres1$p.value >= 0.001) {   # homoskedastic
  linearHypothesis(olsDLasso1, null, vcov = hccm(olsDLasso1, type = "hc0")) # classical White VCOV
} else {
  linearHypothesis(olsDLasso1, null)
}
```

```
## Hypothesis:
## JWMNP = 0
## JWTR02 = 0
## JWTR03 = 0
## JWTR04 = 0
## JWTR05 = 0
## JWTR06 = 0
## JWTR07 = 0
## JWTR08 = 0
## JWTR09 = 0
## JWTR10 = 0
## JWTR12 = 0
## JWMNP:JWTR02 = 0
## JWMNP:JWTR03 = 0
## JWMNP:JWTR04 = 0
## JWMNP:JWTR05 = 0
## JWMNP:JWTR06 = 0
## JWMNP:JWTR07 = 0
## JWMNP:JWTR08 = 0
## JWMNP:JWTR09 = 0
## JWMNP:JWTR10 = 0
## JWMNP:JWTR12 = 0
##
## Model 1: restricted model
## Model 2: log(IncomePovertyRatio) ~ SameResidenceWorkplace * JWMNP + JWTR *
##     JWMNP + SPORDER + PWGTP + AGEP + CIT3 + CIT4 + CIT5 + COW2 +
##     COW3 + COW4 + COW5 + COW6 + COW7 + COW8 + DDRS2 + DEYE2 +
##     DPHY2 + DREM2 + ENG2 + ENG3 + ENG4 + FER1 + FER2 + GCL2 +
##     GCR2 + HINS12 + HINS22 + HINS42 + HINS52 + HINS62 + HINS72 +
##     LANX2 + MAR2 + MAR3 + MAR4 + MARHD2 + MARHT2 + MARHT3 + MARHYP +
##     MIG2 + MIG3 + NWAB2 + NWAV5 + NWLA3 + NWLK3 + NWRE2 + RELP01 +
##     RELP02 + RELP03 + RELP04 + RELP05 + RELP06 + RELP07 + RELP08 +
##     RELP09 + RELP10 + RELP11 + RELP12 + RELP13 + RELP15 + RELP17 +
##     SCHL04 + SCHL05 + SCHL06 + SCHL07 + SCHL08 + SCHL09 + SCHL10 +
##     SCHL11 + SCHL12 + SCHL13 + SCHL14 + SCHL15 + SCHL16 + SCHL17 +
##     SCHL19 + SCHL20 + SCHL22 + SCHL23 + SCHL24 + SEX2 + WKHP +
##     WKW2 + WKW3 + WKW4 + WKW5 + WKW6 + DECADE3 + DECADE4 + DECADE7 +
##     DECADE8 + DIS2 + DRIVESP1 + DRIVESP2 + DRIVESP3 + DRIVESP4 +
##     DRIVESP5 + MSP2 + PAOC1 + PAOC2 + PAOC4 + QTRBIR3 + RACAIAN1 +
##     RACASN1 + RACBLK1 + RACPI1 + RACWHT1 + SCIENGRLP1 + SCIENGRLP2 +
##     WAOB2 + WAOB3 + WAOB5 + WAOB6 + WAOB7 + WAOB8 + AGEP_HINS31 +
##     SCIENGP_SCHL01 + SCIENGP1_SCHL22 + SCIENGP1_SCHL23 + SCIENGP1_SCHL24 +
##     SCIENGRLP1_SCHL22 + SCIENGRLP2_SCHL23 + SCIENGRLP1_SCHL24 +
##     AGEP_VETERAN + AGEP_GCL + DOUT2 + MARHD8 + NWAB3 + RACNH1 +
##     RACSOR1 + AGEP_GCL2 + CIT2 + DEAR2 + GCL1 + NWLA2 + DECADE5 +
##     SCIENGP1_SCHL21 + VETERAN + GCM1 + GCM2 + GCM4 + HINS32 +
##     NWAV3 + SCHL02 + SCHL03 + DECADE1 + DECADE2 + PAOC3 + AGEP_GCR1 +
##     GCM3 + NWAV2 + NWLK2 + NWRE3 + QTRBIR2
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1 781213 148435
## 2 781192 147823 21    612.12 154.04 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
null = c("SameResidenceWorkplaceTRUE","JWMNP",
         "JWTR02","JWTR03","JWTR04","JWTR05","JWTR06","JWTR07","JWTR08",
```

```
         "JWTR09","JWTR10","JWTR12",
         "SameResidenceWorkplaceTRUE:JWMNP",
         "JWMNP:JWTR02","JWMNP:JWTR03","JWMNP:JWTR04","JWMNP:JWTR05","JWMNP:JWTR06",
         "JWMNP:JWTR07","JWMNP:JWTR08","JWMNP:JWTR09","JWMNP:JWTR10","JWMNP:JWTR12")
if (bpres1$p.value >= 0.001) {   # homoskedastic
  linearHypothesis(olsDLasso1, null, vcov = hccm(olsDLasso1, type = "hc0")) # classical White VCOV
} else {
  linearHypothesis(olsDLasso1, null)
}
```

```
## Hypothesis:
## SameResidenceWorkplaceTRUE = 0
## JWMNP = 0
## JWTR02 = 0
## JWTR03 = 0
## JWTR04 = 0
## JWTR05 = 0
## JWTR06 = 0
## JWTR07 = 0
## JWTR08 = 0
## JWTR09 = 0
## JWTR10 = 0
## JWTR12 = 0
## SameResidenceWorkplaceTRUE:JWMNP = 0
## JWMNP:JWTR02 = 0
## JWMNP:JWTR03 = 0
## JWMNP:JWTR04 = 0
## JWMNP:JWTR05 = 0
## JWMNP:JWTR06 = 0
## JWMNP:JWTR07 = 0
## JWMNP:JWTR08 = 0
## JWMNP:JWTR09 = 0
## JWMNP:JWTR10 = 0
## JWMNP:JWTR12 = 0
##
## Model 1: restricted model
## Model 2: log(IncomePovertyRatio) ~ SameResidenceWorkplace * JWMNP + JWTR *
##     JWMNP + SPORDER + PWGTP + AGEP + CIT3 + CIT4 + CIT5 + COW2 +
##     COW3 + COW4 + COW5 + COW6 + COW7 + COW8 + DDRS2 + DEYE2 +
##     DPHY2 + DREM2 + ENG2 + ENG3 + ENG4 + FER1 + FER2 + GCL2 +
##     GCR2 + HINS12 + HINS22 + HINS42 + HINS52 + HINS62 + HINS72 +
##     LANX2 + MAR2 + MAR3 + MAR4 + MARHD2 + MARHT2 + MARHT3 + MARHYP +
##     MIG2 + MIG3 + NWAB2 + NWAV5 + NWLA3 + NWLK3 + NWRE2 + RELP01 +
##     RELP02 + RELP03 + RELP04 + RELP05 + RELP06 + RELP07 + RELP08 +
##     RELP09 + RELP10 + RELP11 + RELP12 + RELP13 + RELP15 + RELP17 +
##     SCHL04 + SCHL05 + SCHL06 + SCHL07 + SCHL08 + SCHL09 + SCHL10 +
##     SCHL11 + SCHL12 + SCHL13 + SCHL14 + SCHL15 + SCHL16 + SCHL17 +
##     SCHL19 + SCHL20 + SCHL22 + SCHL23 + SCHL24 + SEX2 + WKHP +
##     WKW2 + WKW3 + WKW4 + WKW5 + WKW6 + DECADE3 + DECADE4 + DECADE7 +
##     DECADE8 + DIS2 + DRIVESP1 + DRIVESP2 + DRIVESP3 + DRIVESP4 +
##     DRIVESP5 + MSP2 + PAOC1 + PAOC2 + PAOC4 + QTRBIR3 + RACAIAN1 +
##     RACASN1 + RACBLK1 + RACPI1 + RACWHT1 + SCIENGRLP1 + SCIENGRLP2 +
##     WAOB2 + WAOB3 + WAOB5 + WAOB6 + WAOB7 + WAOB8 + AGEP_HINS31 +
##     SCIENGP_SCHL01 + SCIENGP1_SCHL22 + SCIENGP1_SCHL23 + SCIENGP1_SCHL24 +
##     SCIENGRLP1_SCHL22 + SCIENGRLP2_SCHL23 + SCIENGRLP1_SCHL24 +
##     AGEP_VETERAN + AGEP_GCL + DOUT2 + MARHD8 + NWAB3 + RACNH1 +
##     RACSOR1 + AGEP_GCL2 + CIT2 + DEAR2 + GCL1 + NWLA2 + DECADE5 +
##     SCIENGP1_SCHL21 + VETERAN + GCM1 + GCM2 + GCM4 + HINS32 +
##     NWAV3 + SCHL02 + SCHL03 + DECADE1 + DECADE2 + PAOC3 + AGEP_GCR1 +
##     GCM3 + NWAV2 + NWLK2 + NWRE3 + QTRBIR2
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1 781215 149713
## 2 781192 147823 23    1890.1 434.27 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A.6 Exogenous OLS Regression

Regression Without Potential Endogeneity

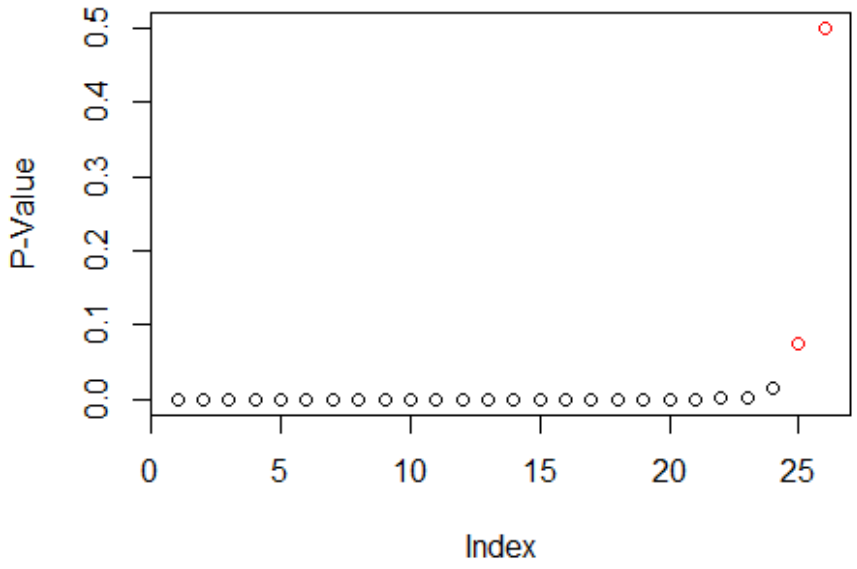
```
## Model 2 Post-Double ML OLs with exogenous variable selection --------------------
# manually delete potentially endogenous variables
# Can X affect or cause Income or Poverty of Both?
endog <- c("SPORDER",    # household size
           "CIT3","CIT4","CIT5",     # citizenship status
           "COW2","COW3","COW4","COW5","COW6","COW7","COW8",   # class of worker
           "DDRS2","DEYE2","DPHY2","DREM2",     # disability
           "ENG2","ENG3","ENG4","FER1","FER2",     # level of english and child birth
           "GCL2","GCR2",     # grandparents with grandchildren
           "HINS12","HINS22","HINS42","HINS52","HINS62","HINS72",    # insurance
           "MAR2","MAR3","MAR4","MARHD2","MARHT2","MARHT3",   # marriage
           "MIG2","MIG3",    # migration
           "NWAB2","NWAV5","NWLA3","NWLK3","NWRE2",   # current work status
           "RELP01","RELP02","RELP03","RELP04","RELP05","RELP06","RELP07",    # relationship in household
           "RELP08","RELP09","RELP10","RELP11","RELP12","RELP13","RELP15","RELP17"
           )
endog2 <- c(61:79,80,81:86,    # degree, sex, work
            91,92:96,97,    # disability, num cars per ppl, marriage status,
            102:106,107:108,    # race, stem degree
            116:122,123:124,125,    # stem*degree, age*stuff, disability
            126:134,    # marriage, work, race, age*stuff, citizenship, disability, work
            136:140,    # school, veteran, grandparents with grandchild
            142:144,    # insurance, work, school
            148:152    # age*stuff, grandparents with grandchild, work
            )

union <- union[-endog2]
union <- union[-which(union %in% endog)] # delete them from formula

# rewrite formula for OLS
exogunionf <- paste(union, collapse = "+")
exogformula <- paste(c("log(IncomePovertyRatio)", exogunionf), collapse = "~")

# Training OLS regression post LASSO
olsDLasso2 <- lm(exogformula, data = dattemp[train,])
DMLresult2 <- summary(olsDLasso2)
# Post-Double LASSO OLS only on Exogeneous vars Result
DMLresult2

##
## Call:
## lm(formula = exogformula, data = dattemp[train, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34876 -0.39110 -0.04425  0.33410  3.11596
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -1.428e+00  1.631e-01   -8.758  < 2e-16 ***
## PWGTP       -1.671e-04  8.701e-06  -19.200  < 2e-16 ***
## AGEP         8.362e-03  9.287e-05   90.045  < 2e-16 ***
## LANX2TRUE    5.025e-02  2.234e-03   22.496  < 2e-16 ***
## MARHYP       1.625e-03  7.909e-05   20.547  < 2e-16 ***
## DECADE3TRUE  1.416e-01  7.867e-03   18.003  < 2e-16 ***
## DECADE4TRUE  1.505e-01  5.346e-03   28.160  < 2e-16 ***
## DECADE7TRUE  2.341e-02  3.879e-03    6.034  1.6e-09 ***
## DECADE8TRUE -5.670e-02  4.600e-03  -12.326  < 2e-16 ***
## PAOC1TRUE   -1.810e-01  3.869e-03  -46.789  < 2e-16 ***
## PAOC2TRUE   -2.695e-01  2.217e-03 -121.533  < 2e-16 ***
## PAOC4TRUE   -3.561e-01  1.553e-03 -229.339  < 2e-16 ***
## QTRBIR3TRUE  4.710e-03  1.610e-03    2.926  0.00344 **
## WAOB2TRUE   -1.658e-01  9.589e-03  -17.294  < 2e-16 ***
## WAOB3TRUE   -3.119e-01  3.194e-03  -97.644  < 2e-16 ***
## WAOB5TRUE    8.521e-02  4.784e-03   17.810  < 2e-16 ***
## WAOB6TRUE   -1.018e-01  7.735e-03  -13.166  < 2e-16 ***
```

```
## WAOB7TRUE    2.027e-01  1.084e-02   18.696  < 2e-16 ***
## WAOB8TRUE    4.730e-02  1.933e-02    2.448  0.01439 *
## AGEP_HINS31 -2.175e-04  3.219e-04   -0.676  0.49922
## DECADE5TRUE  9.779e-02  4.130e-03   23.680  < 2e-16 ***
## HINS32TRUE   1.916e-01  2.207e-02    8.682  < 2e-16 ***
## DECADE1TRUE  7.548e-02  4.239e-02    1.781  0.07499 .
## DECADE2TRUE  7.733e-02  1.371e-02    5.640  1.7e-08 ***
## PAOC3TRUE   -2.705e-01  4.101e-03  -65.946  < 2e-16 ***
## QTRBIR2TRUE  5.347e-03  1.649e-03    3.241  0.00119 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5889 on 781343 degrees of freedom
## Multiple R-squared:  0.106,  Adjusted R-squared:  0.106
## F-statistic:  3708 on 25 and 781343 DF,  p-value: < 2.2e-16

# Test Prediction
pred.olsDLasso.2 <- predict(olsDLasso2, newdata = dattemp[-train,])
summary(pred.olsDLasso.2)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.402   2.113   2.234   2.258   2.431   2.950

length(na.omit(pred.olsDLasso.2)) # count remaining observations

## [1] 195343

# test error
mse.2 <- mean((pred.olsDLasso.2-log(dattemp[-train,]$IncomePovertyRatio))^2, na.rm=T)
mse.2

## [1] 0.34607
```

Result 2: Analysis & Hypothesis Testing

```
# 3 Ways of getting Test R2
y2 <- log(dattemp[-train,]$IncomePovertyRatio)-mean(log(dattemp[-train,]$IncomePovertyRatio))
yhat2 <- pred.olsDLasso.2-mean(pred.olsDLasso.2)
u2 <- y2 - yhat2
# 1:
# R2 = yhat*y/yTy
r2_1_2 <- (yhat2 %*% yhat2)/(y2 %*% y2)
r2_1_2

##           [,1]
## [1,] 0.1064778

# 2:
# R2 = 1- SSR/SST = 1- uTu/yTy
r2_2_2 <- 1 - (u2 %*% u2)/(y2 %*% y2)
r2_2_2

##           [,1]
## [1,] 0.1028011

# 3:
# R2 = corr(y, yhat)^2, "fair r-squared"
r2_3_2 <- cor.test(y2, yhat2, use = "complete.obs")
# now, square the correlation coefficient
r2_3_2

##
##  Pearson's product-moment correlation
##
## data:  y2 and yhat2
## t = 149.63, df = 195341, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.3166914 0.3246485
## sample estimates:
##       cor
## 0.3206756

r2_3_2$estimate^2

##       cor
## 0.1028329

# BP test for heteroskedasticity
bpres2 <- bptest(olsDLasso2, data = dattemp[-train,]) #reject homoskedasticity if p-value is small
bpres2

##
##  studentized Breusch-Pagan test
##
## data:  olsDLasso2
## BP = 13044, df = 25, p-value < 2.2e-16

# False Discovery Rate control
p2 <- as.data.frame(DMLresult2$coefficients[,4])
sigcode2 <- cut(p2[,1], breaks = c(-Inf, 0.001, 0.01, 0.05, 0.1, 1),
                labels = c("***", "**", "*", ".", " "))
p2$"" <- sigcode2

# sort by increasing p-value
p2 <- p2[order(p2$`DMLresult2$coefficients[, 4]`),]
p2$BY <- 0
m2 <- nrow(p2)
Q = 0.10  # 10%
cm=0
for (ii in 1:m2) {
  cm = cm + 1/ii
  p2[ii,3] <- ii/m2/cm*Q
}
noreject2 <- (!(p2[,1] < p2[,3]))
plot(p2$`DMLresult2$coefficients[, 4]`,ylab="P-Value", col = ifelse(noreject2,'red','black'))
```



```r
noreject2 <- which(noreject2)
p2 <- p2[noreject2,]   # these one's we cannot reject the null
names(p2) <- c("p-value","Sig. Level","BY Stat")
p2
```

```
##                p-value Sig. Level    BY Stat
## DECADE1TRUE 0.07498765          . 0.02519782
## AGEP_HINS31 0.49921838            0.02594424
```

```r
# get BY-adjusted p-values
pBY2 <- as.data.frame(p.adjust(p2[,1], method = "BY"))   #Benjamini-Yekutieli
rownames(pBY2) <- rownames(p2)
adjsigcode <- cut(pBY2[,1], breaks = c(-Inf, 0.001, 0.01, 0.05, 0.1, 1),
                  labels = c("***", "**", "*", ".", " "))
pBY2$"" <- adjsigcode

# compare p-values for non-rejected
fdr2 <- cbind.data.frame(p2[,c(1,2)], pBY2)
colnames(fdr2) <- c("Original","Sig. Level", "FDR Adj.","Sig. Level")
fdr2
```

```
##               Original Sig. Level  FDR Adj. Sig. Level
## DECADE1TRUE 0.07498765          . 0.2249630
## AGEP_HINS31 0.49921838            0.7488276
```