Examining the Association between Socioeconomic Status(SES) and Air Pollution in the United States

Yunbi Nam, Yuhan Wang, Sirui Liu

1 Introduction

Major current environmental issues include air pollution, global warming, and resource depletion, etc. Air pollution is the release of pollutants into the air mainly generated by human activities or natural processes. We can no longer afford to ignore air pollution problems due to their severity and associations with several health risks including stroke, lung cancer, and respiratory infections. According to the World Health Organization, 4.2 million deaths every year are caused by exposure to outdoor air pollution. Furthermore, several social studies have shown that racial minorities are considered to be more vulnerable to the process of urbanization and more susceptible to air pollution. It applies to the United States where there are substantial racial disparities in living conditions and health outcomes.

Various research papers have addressed environmental inequalities and environmental justice over the years. According to Hajat et al. (2015), "Most North American studies have shown that areas where low socioeconomic status (SES) communities dwell experience higher concentrations of criteria air pollutants". Therefore, conducting the association between SES and air quality is important in understanding the causes of disparities in health outcomes related to air pollution. A few studies have concluded that exposures to particular pollutants differed by race/ethnicity, age, and SES (Bell and Ebisu 2012). Our project aimed to investigate the county-level association between the socioeconomic status index and the air quality index in the United States, adjusting for race/ethnicity and age. To construct the SES index, the top three principal components were selected by applying principal component analysis with various socioeconomic variables. We also built a predictive model for the air quality index (AQI) from several population characteristics.

2 Data

We adopted two public sets of data, 2014-2018 American Community Survey (ACS) data and 2018 United States Environmental Protection Agency (EPA) annual summary data, to conduct our analysis. Both of them can be downloaded online.

Socioeconomic status is a measure of the economic and social status of an individual or group of individuals based on education, income, occupation, and other relevant indicators. It is a complex and multidimensional concept and there are a lot of ways to measure socioeconomic status. To take into account several aspects rather than considering only one variable, we create an SES index by applying principal component analysis (PCA) with a pre-defined set

of variables from 2014-2018 ACS data.

2014-2018 ACS data include county-level statistics averaged over 5 years on social, economic, housing, and demographic characteristics from the Census Bureau Survey. The data file contains 3220 county-level records including counties, boroughs, census areas, the District of Columbia, and other equivalent entities. For each county, we considered the following population characteristics:

- Education: percent of persons 25 or older with at least high school education, percent of persons 25 or older with at least a Bachelor's degree
- Income: median household income, percent of households with household income $50,000 or more, percent of persons below the poverty level
- Occupation: percent unemployed among civilians 16 and over in the labor force, percent of civilians 16 and over not in the labor force, percent with management, business, science, and arts occupations
- Wealth: percent of occupied housing units, percent of housing units that are occupied out of total housing units, median value of occupied housing units, percent of housing units without vehicle

The above 12 variables are chosen from the ones used in Hajat et al. (2013). They are included in the principal component analysis to develop SES indices for our inference. As for demographic characteristics covariates, we use median age, and percent of persons self-identified as non-Hispanic white.

2018 U.S. EPA annual summary data provide county-level median AQI established for five major air pollutants; ground-level ozone, particulate matter, carbon monoxide, sulfur dioxide, and nitrogen dioxide. We chose the AQI for the air quality data since it reflects the health-based air quality standards. It is obtained daily based on the concentration of air pollutants at designated sites. Table 1 explains each level of air quality index. Annual summary AQI file by county shows the annual statistics for AQI. There are a total of 1056 observations in our data.

After we match two sets of data, we have 1045 county-level observations.

3 Methods

Our analysis includes two parts. In the first part, we do hypothesis testing to investigate the association between SES and AQI level, adjusting for age and race/ethnicity. The second part is to build a predictive model for the AQI level using demographic and SES variables by using the lasso. Analyses were perforemd using R, version 4.0.2.

In order to evaluate the association between SES and AQI level, we conducted principal component analysis (PCA) with orthogonal rotation to reduce a number of variables into a smaller number of dimensions. We used the first three components as a summary index to represent socioeconomic status in each county. They are uncorrelated with each other and explain the large amounts of variation in the original data. Twelve census variables were selected to be included in the PCA. All SES variables were scaled so that higher values indicate higher SES.

We used linear mixed effects models with a random intercept to account for the clustering of counties within states since our data are spatially correlated. From Figure 1 and Figure 2, we assume that counties located within the same state are more similar to each other than counties located in different states. But we do not make assumptions about

Levels of Concern	Values of Index	Description of Air Quality
Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Table 1: AQI Basics for Ozone and Particle Pollution

the relationship between counties. Median age and percentage of non-Hispanic white were included in the model as confounders. We adjusted for these variables because we expected age and race/ethnicity are associated with SES and the air quality level. The equation for our model is below:

$$Y_{ij} = \beta_{0j} + \beta_{1j}P_1 + \beta_{2j}P_2 + \beta_{3j}P_3 + \alpha_{1j}Age + \alpha_{2j}NHW + a_{ij} + \epsilon_i \quad (1)$$

where Y_{ij} is air quality index for i^{th} county within state j, P_{1j}, P_{2j}, P_{3j} are first three principal components of SES characteristics, a_{ij} is within state error term (random effect), and ϵ_i is between state error term.

We fit nested models and performed a likelihood ratio test to assess the strength of evidence against the null hypothesis of no association between SES and the air quality index. The alpha level used to define statistical significance was set to 0.05 and we tested the following:

- Full model: $Y_{ij} = \beta_{0j} + \beta_{1j}P_1 + \beta_{2j}P_2 + \beta_{3j}P_3 + \alpha_{1j}Age + \alpha_{2j}NHW + a_{ij} + \epsilon_i$
- Reduced model: $Y_{ij} = \beta_{0j} + \alpha_{1j}Age + \alpha_{2j}NHW + a_{ij} + \epsilon_i$
- Null hypothesis $H_0 : \beta_{1j} = \beta_{2j} = \beta_{3j} = 0$

To build a predictive model, we randomly split the sample into a training and testing set (1:1). We built predictive models for AQI level on the training test, using demographic and SES variables selected by lasso regression. The response variable in lasso regression was the 2018 median AQI obtained at the county level. The covariates were all the SES variables in the dataset. We first used ten-fold cross-validation to select the optimal lambda for the penalty term in lasso regression. Then we fitted lasso regression with the optimal lambda. We selected the variables with absolute coefficient estimates higher than 0.01 and built a predictive model on the training set as model 1. Then we only selected the variables with absolute coefficient estimates higher than 0.1 and built a predictive model on the

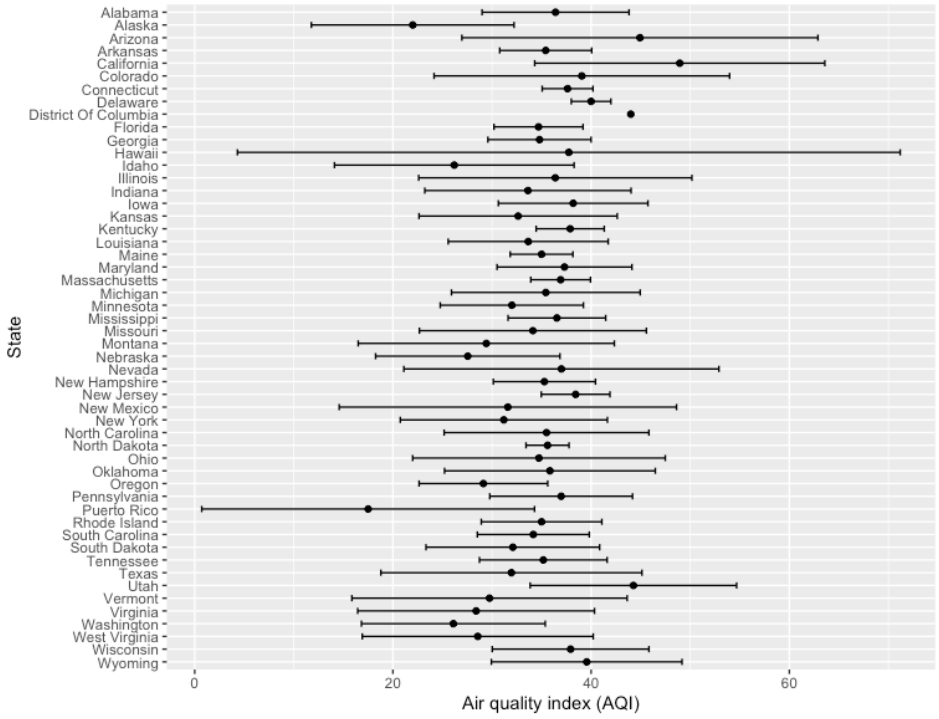

Figure 1: Error bar charts of mean AQI by each state presenting mean ± sd

training set as model 2. We tested model 1 and model 2 on the testing set and compared the mean square errors of the two predictive models.

4 Results

ACS data provided ones from 3220 counties and county-equivalents in 50 states, the District of Columbia, and Puerto Rico. On the other hand, we had AQI index from 1056 counties and county-equivalents because they were collected from designated monitors. After matching these two data, we had 1045 observations. We missed some of the samples because they stored the names of geographic areas in different ways. In addition, we had some missing values in a part of demographic variables, leaving observations in the fitted model 1038. We excluded these missing values because we have enough samples.

AQI	P_1	P_2	P_3	Age	NHW
Model A	-	-	-	-	-
Model B	-0.81 (-1.09, -0.52)	-1.38 (-1.81, -0.96)	1.71 (1.08, 2.34)	-	-
Model C	-	-	-	-0.08 (-0.13, -0.04)	-0.29 (-0.43, -0.14)
Model D	-0.85 (-1.15, -0.56)	-0.94 (-1.60, -0.29)	1.64 (0.92, 2.37)	-0.07 (-0.13, -0.02)	0.05 (-0.13, 0.23)

Table 2: Estimates with 95% CI from Model A through Model D, '-' variable not included in the model

PCA found that around 66% of variance in twelve factors is explained by the first two PC. Table 2 showed results from several models including Model D which we pre-specified. There were negative associations between the first

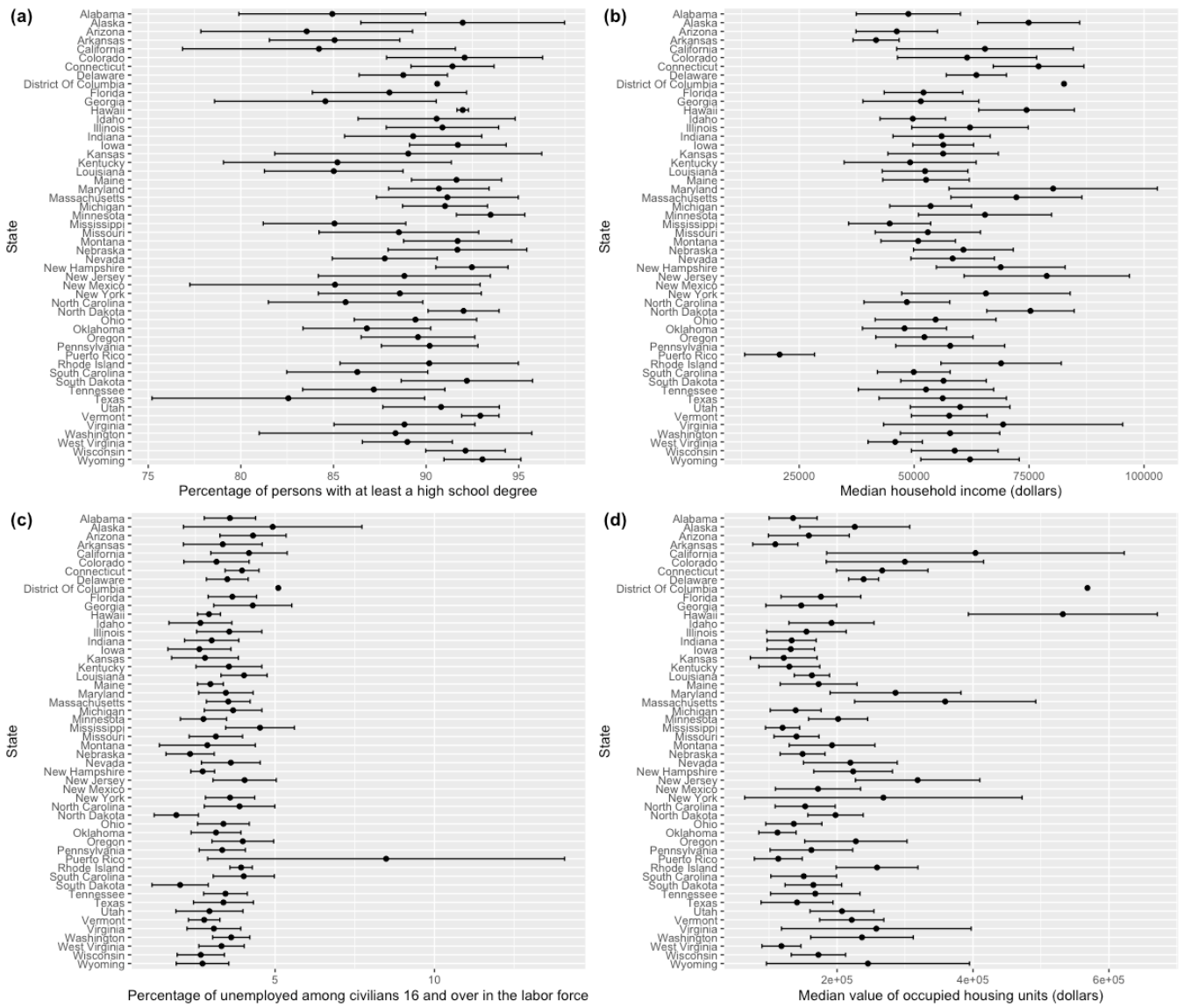

Figure 2: Error bar charts of mean socioeconomic characteristics by each state presenting mean ± sd

two PC and AQI, and it implies that the higher socioeconomic status population enjoys a better air quality index. For example, a 1-unit increase in the first PC was associated with 0.85 lower AQI (95% CI: -1.15, -0.56). Without adjustment (Model B), a 1-unit increase in the first PC was associated with 0.81 lower AQI (95% CI: -1.09, -0.52) and a 1-unit increase in the second PC was associated with 1.38 lower (95% CI: -1.81, -0.96). Even though the third PC showed positive associations in both Models B and D, the third PC explained only for 8.92% of variance.

To conduct the null hypothesis of no association between SES and AQI, we need different kinds of tests. The likelihood ratio test is one way to do it. The tests between Model A and Model B, and between Model C and Model D both provide statistically significant evidence that SES indices are associated with the air quality regardless of adjustment for race/ethnicity and age ($p < 0.001$).

From Table 5, we see that model 1 has a lower mean square error than model 2 when predicting the testing set. Therefore we concluded that model 1 has a better performance in predicting the air quality index than model 2.

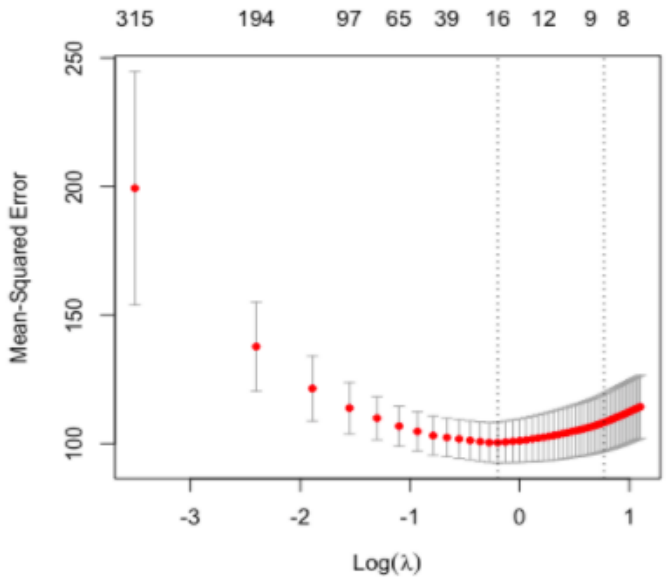

Figure 3: Lambda selection in lasso regression

Predictor Variable	Coefficient
Unemployed	1.11
Average household size of renter-occupied unit	3.29
House heating - Solar energy	0.0001
Housing units with a mortgage	0.0001

Table 3: Model 1

Predictor Variable	Coefficient
Unemployed	1.15
Average household size of renter-occupied unit	3.51
House heating - Solar energy	1.28
House heating - Utility gas	0.00001
Housing units with a mortgage	0.00003
Owner-occupied units value	0.19
Grandparents responsible for grandchildren 1-2 years	0.001

Table 4: Model 2

	Mean squared error
Model 1	23.48
Model 2	26.93

Table 5: Mean squared errors of AQI index on testing set from Model 1 and Model 2

5 Discussion

We examined the association of derived SES indices and the AQI obtained at the county level, with and without adjusting for race/ethnicity and age. Regardless of adjustment, we found strong evidence to reject the null hypothesis that there is no county-level association between SES and the air quality in the United States. Our findings are consistent with other findings in recent papers related to environmental justice. (Hajat et al. 2015)

We expected to conclude that areas with the higher socioeconomic status population experience a lower AQI as supported by previous studies (Bell and Ebisu 2012). However, it is hard to be supported by our findings. We observed lower AQI (better air quality) with the first and second PC but not with the third one. In addition, we do not see the differences between adjusted and unadjusted models with race/ethnicity and age demographic variables.

In terms of SES indices, we simply combined variables and summarized socioeconomic status variables with simple methods by selecting the first three principal components from PCA. Even though PCA has been validated as a method to describe SES differentiation within a population and construct SES indices, our way needs to be improved.

By fitting multilevel models, we could account for the closer relationship within states, but not explain the relationship between states. We might be able to use a spatial intrinsic conditional autoregressive (ICAR) model for our analysis, which also makes assumptions about the between states effect.

Regarding our data, since the AQI is estimated from designated monitor sites and only the median AQI was adopted for the analysis, we may not have been detected the significance in the relationship investigated ignoring the standard error of the estimates.

6 References

[1] Bell ML and Ebisu K: Environmental inequality in exposures to airborne particulate matter components in the United States. Environmental Health Perspectives 120(12):1699-1704, Dec, 2012.

[2] Hajat A, Diez-Roux AV, Adar SD, Auchincloss AH, Lovasi GS, O'Neill MS, Sheppard L, Kaufman JD: Air pollution and neighborhood socioeconomic status: evidence from the Multi-Ethnic Study of Atherosclerosis (MESA). Environmental Health Perspectives 121(11-12):1325-1333, Nov-Dec, 2013.

[3] Hajat A, Hsia C, O'Neill MS: Socioeconomic disparities and air pollution exposure: a global review. Curr Envir Health Rpt 2:440-450, 2015.

[4] AirNow. AQI Basics. URL: https://www.airnow.gov/aqi/aqi-basics/

Appendix

```
## inference

library(dplyr)
library(stringr)
library(ggplot2)
library(lme4)
library(scales)
library(ggpubr)

aqi <- read.csv('annual_aqi_by_county_2018.csv')

## We need median AQI for analysis and Geographic Area Name(State-County)
# for matching variables to demographic variables
aqi_data <- aqi[,c("State", "County", "Median.AQI")]
aqi_data <- aqi_data %>%
        mutate(state = tolower(aqi_data$State),
               county = tolower(aqi_data$County))
aqi_data$county <- unlist(sapply(aqi_data$county,
                                 function(x) if(word(x,1)=="saint"){
                                         x <- sub('.*saint','st.',x)
                                 } else if(word(x,1)=="sainte"){
                                         x <- sub('.*sainte','ste.',x)
                                 } else{x <- x}))

## 2014-2018 ACS 5-Year Data Social Characteristics
acs_soc <- read.csv('soc_char/ACSDP5Y2018.DP02_data_with_overlays_2020-12-03T145030.csv', header=TRUE)
names(acs_soc) <- acs_soc[1,]
acs_soc <- acs_soc[-1,]
acs_soc <- acs_soc %>%
        mutate(state = tolower(sub('.*, ','',acs_soc$`Geographic Area Name`)),
               county = tolower(sub(',.*','',acs_soc$`Geographic Area Name`)))

## We have the following patterns in then name of counties
unique(word(acs_soc$county,-1))

## We manipulate those patterns to match two datasets
acs_soc$county <- unlist(sapply(acs_soc$county,
                                function(x) if(word(x,-1)=="county"){
                                        x <- sub(' county.*','',x)
                                }else if(word(x,-1)=="borough"){
                                        if(word(x,-2,-1)=="and borough"){
                                                x <- sub('city and borough.*','',x)
                                        }else{
                                                x <- sub('borough.*','',x)
                                        }
                                }else if(word(x,-1)=="area"){
                                        x <- sub('census area.*','',x)
                                }else if(word(x,-1)=="municipality"){
                                        x <- sub('municipality.*','',x)
                                }else if(word(x,-1)=="columbia"){
                                        x <- x
```

```
                                }else if(word(x,-1)=="parish"){
                                        x <- sub(' parish.*','',x)
                                }else if(word(x,-1)=="city"){
                                        x <- x
                                }else{
                                        x <- sub(' municipio.*','',x)
                                }))

matched_soc <- inner_join(acs_soc, aqi_data, by=c("state", "county"))

matched_educ <- matched_soc[,c("Geographic Area Name", "State", "County", "Median.AQI",
                               "Percent Estimate!!EDUCATIONAL ATTAINMENT!!Population 25 years and over
                               !!High school graduate or higher",
                               "Percent Estimate!!EDUCATIONAL ATTAINMENT!!Population 25 years and over
                               !!Bachelor's degree or higher")]

names(matched_educ) <- unlist(sapply(names(matched_educ),
                                     function(x) if(word(x,1)=="Percent"){
                                             x <- sub('.*over!!','',x)
                                     }else{x <- x}))

matched_educ[,5:6] <- apply(matched_educ[,5:6], 2, as.numeric)

educ_summ <- matched_educ %>% group_by(State) %>%
        summarize(mean_high=mean(`High school graduate or higher`),
                  sd_high=sd(`High school graduate or higher`),
                  mean_bach=mean(`Bachelor's degree or higher`),
                  sd_bach=sd(`Bachelor's degree or higher`))

educ_summ$State <- factor(educ_summ$State, levels=educ_summ$State[rev(order(educ_summ$State))])

p1 <- ggplot(educ_summ, aes(x=mean_high, y=State)) +
        geom_point(position = position_dodge(width = .5)) +
        geom_errorbar(aes(xmin = mean_high - sd_high, xmax = mean_high + sd_high),
                      width = .5, position = "dodge") +
        labs(x = "Percentage of persons with at least a high school degree")

## 2014-2018 ACS 5-Year Data Economic Characteristics
acs_econ <- read.csv('econ_char/ACSDP5Y2018.DP03_data_with_overlays_2020-12-04T204808.csv')
names(acs_econ) <- acs_econ[1,]
acs_econ <- acs_econ[-1,]
acs_econ <- acs_econ %>%
        mutate(state = tolower(sub('.*, ','',acs_econ$`Geographic Area Name`)),
               county = tolower(sub(',.*','',acs_econ$`Geographic Area Name`)))

## manipulation
acs_econ$county <- unlist(sapply(acs_econ$county,
                                 function(x) if(word(x,-1)=="county"){
                                         x <- sub(' county.*','',x)
                                 }else if(word(x,-1)=="borough"){
                                         if(word(x,-2,-1)=="and borough"){
                                                 x <- sub('city and borough.*','',x)
                                         }else{
```

```
                                        x <- sub('borough.*','',x)
                                }
                        }else if(word(x,-1)=="area"){
                                x <- sub('census area.*','',x)
                        }else if(word(x,-1)=="municipality"){
                                x <- sub('municipality.*','',x)
                        }else if(word(x,-1)=="columbia"){
                                x <- x
                        }else if(word(x,-1)=="parish"){
                                x <- sub(' parish.*','',x)
                        }else if(word(x,-1)=="city"){
                                x <- x
                        }else{
                                x <- sub(' municipio.*','',x)
                        }))

matched_econ <- inner_join(acs_econ, aqi_data, by=c("state", "county"))

matched_inc <- matched_econ[,c("Geographic Area Name", "State", "County", "Median.AQI",
                               "Percent Estimate!!EMPLOYMENT STATUS!!Population 16 years and over
                               !!In labor force!!Civilian labor force!!Unemployed",
                               "Percent Estimate!!EMPLOYMENT STATUS!!Population 16 years and over
                               !!Not in labor force",
                               "Percent Estimate!!OCCUPATION!!Civilian employed
                               population 16 years and over
                               !!Management, business, science, and arts occupations",
                               "Estimate!!INCOME AND BENEFITS (IN 2018 INFLATION-
                               ADJUSTED DOLLARS)!!Total households!!Median household income (dollars)",
                               "Percent Estimate!!INCOME AND BENEFITS (IN 2018 INFLATION-
                               ADJUSTED DOLLARS)!!Total households!!$50,000 to $74,999",
                               "Percent Estimate!!INCOME AND BENEFITS (IN 2018 INFLATION-
                               ADJUSTED DOLLARS)!!Total households!!$75,000 to $99,999",
                               "Percent Estimate!!INCOME AND BENEFITS (IN 2018 INFLATION-
                               ADJUSTED DOLLARS)!!Total households!!$100,000 to $149,999",
                               "Percent Estimate!!INCOME AND BENEFITS (IN 2018 INFLATION-
                               ADJUSTED DOLLARS)!!Total households!!$150,000 to $199,999",
                               "Percent Estimate!!INCOME AND BENEFITS (IN 2018 INFLATION-
                               ADJUSTED DOLLARS)!!Total households!!$200,000 or more",
                               "Percent Estimate!!PERCENTAGE OF FAMILIES AND PEOPLE
                               WHOSE INCOME IN THE PAST 12 MONTHS IS BELOW THE POVERTY LEVEL!!
                               All people")]

names(matched_inc) <- c("Geographic Area Name", "State", "County", "Median.AQI",
                        "Unemployed", "Not in labor force", "Management", "Median household income",
                        "50,000-74,999", "75,000-99,999", "100,000-149,999",
                        "150,000-199,999", "200,000+", "Below the poverty level")

matched_inc[,5:14] <- apply(matched_inc[,5:14], 2, as.numeric)

matched_inc <- matched_inc %>%
        mutate("50,000+"=apply(matched_inc[,9:13], 1, sum))

inc_summ <- matched_inc %>% group_by(State) %>%
```

```
        summarize(mean_unemp=mean(`Unemployed`), sd_unemp=sd(`Unemployed`),
                  mean_nolab=mean(`Not in labor force`), sd_nolab=sd(`Not in labor force`),
                  mean_mngm=mean(`Management`), sd_mngm=sd(`Management`),
                  mean_hhi=mean(`Median household income`), sd_hhi=sd(`Median household income`),
                  mean_hhi50000=mean(`50,000+`), sd_hhi50000=sd(`50,000+`),
                  mean_pov=mean(`Below the poverty level`), sd_pov=sd(`Below the poverty level`))

inc_summ$State <- factor(inc_summ$State, levels=inc_summ$State[rev(order(inc_summ$State))])

p2 <- ggplot(inc_summ, aes(x=mean_hhi, y=State)) +
        geom_point(position = position_dodge(width = .5)) +
        geom_errorbar(aes(xmin = mean_hhi - sd_hhi, xmax = mean_hhi + sd_hhi),
                      width = .5, position = "dodge") +
        xlab("Median household income (dollars)")

p3 <- ggplot(inc_summ, aes(x=mean_unemp, y=State)) +
        geom_point(position = position_dodge(width = .5)) +
        geom_errorbar(aes(xmin = mean_unemp - sd_unemp, xmax = mean_unemp + sd_unemp),
                      width = .5, position = "dodge") +
        xlab("Percentage of unemployed among civilians 16 and over in the labor force")

## 2014-2018 ACS 5-Year Data Housing Characteristics
acs_hous <- read.csv('hous_char/ACSDP5Y2018.DP04_data_with_overlays_2020-12-07T194246.csv')
names(acs_hous) <- acs_hous[1,]
acs_hous <- acs_hous[-1,]
acs_hous <- acs_hous %>%
        mutate(state = tolower(sub('.*, ','',acs_hous$`Geographic Area Name`)),
               county = tolower(sub(',.*','',acs_hous$`Geographic Area Name`)))

## manipulation
acs_hous$county <- unlist(sapply(acs_hous$county,
                                 function(x) if(word(x,-1)=="county"){
                                         x <- sub(' county.*','',x)
                                 }else if(word(x,-1)=="borough"){
                                         if(word(x,-2,-1)=="and borough"){
                                                 x <- sub('city and borough.*','',x)
                                         }else{
                                                 x <- sub('borough.*','',x)
                                         }
                                 }else if(word(x,-1)=="area"){
                                         x <- sub('census area.*','',x)
                                 }else if(word(x,-1)=="municipality"){
                                         x <- sub('municipality.*','',x)
                                 }else if(word(x,-1)=="columbia"){
                                         x <- x
                                 }else if(word(x,-1)=="parish"){
                                         x <- sub(' parish.*','',x)
                                 }else if(word(x,-1)=="city"){
                                         x <- x
                                 }else{
                                         x <- sub(' municipio.*','',x)
                                 }))
```

```r
matched_hous <- inner_join(acs_hous, aqi_data, by=c("state", "county"))

matched_hs <- matched_hous[,c("Geographic Area Name", "State", "County", "Median.AQI",
                              "Percent Estimate!!HOUSING OCCUPANCY!!Total housing units!!
                              Occupied housing units",
                              "Percent Estimate!!HOUSING TENURE!!
                              Occupied housing units!!Owner-occupied",
                              "Percent Estimate!!VEHICLES AVAILABLE!!
                              Occupied housing units!!No vehicles available",
                              "Estimate!!VALUE!!Owner-occupied units!!Median (dollars)")]

matched_hs[,5:8] <- apply(matched_hs[,5:8], 2, as.numeric)

names(matched_hs) <- c("Geographic Area Name", "State", "County", "Median.AQI",
                       "Occupied housing units", "Owner-occupied", "No vehicles", "Median value")

hs_summ <- matched_hs %>% group_by(State) %>%
        summarize(mean_occ=mean(`Occupied housing units`), sd_occ=sd(`Occupied housing units`),
                  mean_ownocc=mean(`Owner-occupied`), sd_ownocc=sd(`Owner-occupied`),
                  mean_nv=mean(`No vehicles`), sd_nv=sd(`No vehicles`),
                  mean_mv=mean(`Median value`), sd_mv=sd(`Median value`))

hs_summ$State <- factor(hs_summ$State, levels=hs_summ$State[rev(order(hs_summ$State))])

p4 <- ggplot(hs_summ, aes(x=mean_mv, y=State)) +
        geom_point(position = position_dodge(width = .5)) +
        geom_errorbar(aes(xmin = mean_mv - sd_mv, xmax = mean_mv + sd_mv),
                      width = .5, position = "dodge") +
        xlab("Median value of occupied housing units (dollars)")

## 2014-2018 ACS 5-Year Data Demographic Characteristics
acs_demo <- read.csv('demo_char/ACSDP5Y2018.DP05_data_with_overlays_2020-12-03T134126.csv')
names(acs_demo) <- acs_demo[1,]
acs_demo <- acs_demo[-1,]
names(acs_demo) <- make.names(names(acs_demo), unique=TRUE)
acs_demo <- acs_demo %>%
        mutate(state = tolower(sub('.*, ','',acs_demo$`Geographic.Area.Name`)),
               county = tolower(sub(',.*','',acs_demo$`Geographic.Area.Name`)))

## manipulation
acs_demo$county <- unlist(sapply(acs_demo$county,
                                 function(x) if(word(x,-1)=="county"){
                                         x <- sub(' county.*','',x)
                                 }else if(word(x,-1)=="borough"){
                                         if(word(x,-2,-1)=="and borough"){
                                                 x <- sub('city and borough.*','',x)
                                         }else{
                                                 x <- sub('borough.*','',x)
                                         }
                                 }else if(word(x,-1)=="area"){
                                         x <- sub('census area.*','',x)
                                 }else if(word(x,-1)=="municipality"){
                                         x <- sub('municipality.*','',x)
```

```
                                }else if(word(x,-1)=="columbia"){
                                        x <- x
                                }else if(word(x,-1)=="parish"){
                                        x <- sub(' parish.*','',x)
                                }else if(word(x,-1)=="city"){
                                        x <- x
                                }else{
                                        x <- sub(' municipio.*','',x)
                                }))

matched_demo <- inner_join(acs_demo, aqi_data, by=c("state", "county"))

matched_cov <- matched_demo[,c("Geographic.Area.Name", "State", "County", "Median.AQI",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..Under.5.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..5.to.9.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..10.to.14.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..15.to.19.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..20.to.24.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..25.to.34.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..35.to.44.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..45.to.54.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..55.to.59.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..60.to.64.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..65.to.74.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..75.to.84.years",
                               "Percent.Estimate..SEX.AND.AGE..Total.population..85.years.and.over",
                               "Percent.Estimate..HISPANIC.OR.LATINO.AND.RACE..Total.population..
                               Hispanic.or.Latino..of.any.race.",
                               "Percent.Estimate..HISPANIC.OR.LATINO.AND.RACE..Total.population..
                               Not.Hispanic.or.Latino..White.alone",
                               "Percent.Estimate..HISPANIC.OR.LATINO.AND.RACE..Total.population..
                               Not.Hispanic.or.Latino..Black.or.African.American.alone",
                               "Percent.Estimate..HISPANIC.OR.LATINO.AND.RACE..Total.population..
                               Not.Hispanic.or.Latino..American.Indian.and.Alaska.Native.alone",
                               "Percent.Estimate..HISPANIC.OR.LATINO.AND.RACE..Total.population..
                               Not.Hispanic.or.Latino..Asian.alone",
                               "Percent.Estimate..HISPANIC.OR.LATINO.AND.RACE..Total.population..
                               Not.Hispanic.or.Latino..Native.Hawaiian.
                               and.Other.Pacific.Islander.alone",
                               "Percent.Estimate..HISPANIC.OR.LATINO.AND.RACE..Total.population..
                               Not.Hispanic.or.Latino..Some.other.race.alone",
                               "Percent.Estimate..HISPANIC.OR.LATINO.AND.RACE..Total.population..
                               Not.Hispanic.or.Latino..Two.or.more.races",
                               "Estimate..SEX.AND.AGE..Total.population..Median.age..years.")]

names(matched_cov) <- c("Geographic.Area.Name", "State", "County", "Median.AQI", "ageunder5",
                        "age5-9", "age10-14", "age15-19", "age20-24", "age25-34", "age35-44",
                        "age45-54", "age55-59", "age60-64", "age65-74", "age75-84", "age85+",
                        "hispanic", "nh_white", "nh_baa", "nh_ai", "nh_a", "nh_nh", "nh_otheralone",
                        "nh_two", "agemed")

matched_cov[,5:26] <- apply(matched_cov[,5:26], 2, as.numeric)
```

```r
cov_summ <- matched_cov %>% group_by(State) %>%
        summarize(mean_aqi=mean(`Median.AQI`), sd_aqi=sd(`Median.AQI`),
                  mean_age=mean(`agemed`), sd_age=sd(`agemed`),
                  mean_nhw=mean(`nh_white`), sd_nhw=sd(`nh_white`))

cov_summ$State <- factor(cov_summ$State, levels=cov_summ$State[rev(order(cov_summ$State))])

ggplot(cov_summ, aes(x=mean_aqi, y=State)) +
        geom_point(position = position_dodge(width = .5)) +
        geom_errorbar(aes(xmin = mean_aqi - sd_aqi, xmax = mean_aqi + sd_aqi),
                      width = .5, position = "dodge") +
        xlab("Air quality index (AQI)")

ggarrange(p1, p2, p3, p4,
          labels=c("(a)", "(b)", "(c)", "(d)"),
          ncol = 2, nrow = 2)

matched_cov <- matched_cov %>%
        mutate("nh_other"=apply(matched_cov[,c(21,23:25)], 1, sum),
               "age0-19"=apply(matched_cov[,5:8], 1, sum),
               "age20-44"=apply(matched_cov[,9:11], 1, sum),
               "age45-64"=apply(matched_cov[,12:14], 1, sum),
               "age65+"=apply(matched_cov[,15:17], 1, sum))

matched <- inner_join(matched_educ, matched_inc,
                      by=c('Geographic Area Name','State','County','Median.AQI'))
matched <- inner_join(matched, matched_hs,
                      by=c('Geographic Area Name','State','County','Median.AQI'))
matched <- inner_join(matched, matched_cov,
                      by=c('Geographic Area Name'='Geographic.Area.Name','State','County','Median.AQI'))

matched_use <- matched[,c(1:4, 5:10, 16:17, 18:21, 35:37, 39, 43:48)]

matched_use <- matched_use %>% na.omit()

matched_use.pca <- prcomp(matched_use[,5:16], center=TRUE, scale.=TRUE, retx=TRUE)
screeplot(matched_use.pca, type="l")
abline(h=1)

autoplot(matched_use.pca)

summary(matched_use.pca)

matched_use <- matched_use %>%
        mutate(ses_pc1=matched_use.pca$x[,1],
               ses_pc2=matched_use.pca$x[,2],
               ses_pc3=matched_use.pca$x[,3])

## without adjustment for any variables
fit <- lmer(Median.AQI ~ ses_pc1 + ses_pc2 + ses_pc3 + (1|State), data = matched_use)
summary(fit)

aux <- summary(fit)$coefficients
```

```r
aux <- round(aux, 3)
aux <- cbind(aux, round(aux[,1] + qnorm(0.025)*aux[,2], 3))
aux <- cbind(aux, round(aux[,1] + qnorm(0.975)*aux[,2], 3))
colnames(aux) <- c("Estimate", "robust SE", "t-value", "95% CI lower", "95% CI upper")

fit.null <- lmer(Median.AQI ~ (1|State), data = matched_use)

aux <- summary(fit.null)$coefficients
aux <- round(aux, 3)
aux <- cbind(aux, round(aux[,1] + qnorm(0.025)*aux[,2], 3))
aux <- cbind(aux, round(aux[,1] + qnorm(0.975)*aux[,2], 3))
colnames(aux) <- c("Estimate", "robust SE", "t-value", "95% CI lower", "95% CI upper")

## likelihood ratio test
anova(fit, fit.null, test="Chisq")

## with adjustment for race (nh_white) and age
fit_adj <- lmer(Median.AQI ~ ses_pc1 + ses_pc2 + ses_pc3
                + nh_white + agemed + (1|State), data = matched_use)
summary(fit_adj)
aux <- summary(fit_adj)$coefficients
aux <- round(aux, 3)
aux <- cbind(aux, round(aux[,1] + qnorm(0.025)*aux[,2], 3))
aux <- cbind(aux, round(aux[,1] + qnorm(0.975)*aux[,2], 3))
colnames(aux) <- c("Estimate", "robust SE", "t-value", "95% CI lower", "95% CI upper")

fit_null <- lmer(Median.AQI ~ nh_white + agemed + (1|State), data = matched_use)
aux <- summary(fit_null)$coefficients
aux <- round(aux, 3)
aux <- cbind(aux, round(aux[,1] + qnorm(0.025)*aux[,2], 3))
aux <- cbind(aux, round(aux[,1] + qnorm(0.975)*aux[,2], 3))
colnames(aux) <- c("Estimate", "robust SE", "t-value", "95% CI lower", "95% CI upper")
## likelihood ratio test
anova(fit_adj, fit_null, test="Chisq")
```

```r
## prediction model

library(dplyr)
library(stringr)
library(lme4)
library(scales)
library(readr)
library(glmnet)
aqi <- read.csv('annual_aqi_by_county_2018.csv')
aqi_data <- aqi[,c("State", "County", "Median.AQI")]

## 2014-2018 ACS 5-Year Data Social Characteristics
acs_soc <- read.csv('soc_char/ACSDP5Y2018.DP02_data_with_overlays_2020-12-03T145030.csv', header=TRUE)
names(acs_soc) <- acs_soc[1,]
acs_soc <- acs_soc[-1,]
acs_soc <- acs_soc %>%
  mutate(State = sub('.*, ','',acs_soc$`Geographic Area Name`)) %>%
  mutate(County = sub(',.*','',acs_soc$`Geographic Area Name`))

acs_soc <- acs_soc %>%
  mutate(County = sub(' County.*','',acs_soc$County))


## 2014-2018 ACS 5-Year Data Economic Characteristics
acs_econ <- read.csv('econ_char/ACSDP5Y2018.DP03_data_with_overlays_2020-12-04T204808.csv')
names(acs_econ) <- acs_econ[1,]
acs_econ <- acs_econ[-1,]

## 2014-2018 ACS 5-Year Data Housing Characteristics
acs_hous <- read.csv('hous_char/ACSDP5Y2018.DP04_data_with_overlays_2020-12-07T194246.csv')
names(acs_hous) <- acs_hous[1,]
acs_hous <- acs_hous[-1,]

## 2014-2018 ACS 5-Year Data Demographic Characteristics
acs_demo <- read.csv('demo_char/ACSDP5Y2018.DP05_data_with_overlays_2020-12-03T134126.csv')
names(acs_demo) <- acs_demo[1,]
acs_demo <- acs_demo[-1,]

matched <- inner_join(aqi_data, acs_soc, by=c("State", "County"))
matched <- inner_join(matched, acs_econ, by=c("Geographic Area Name","id"))
matched <- inner_join(matched, acs_hous, by=c("Geographic Area Name","id"))
matched <- inner_join(matched, acs_demo[,unique(colnames(acs_demo))],
                      by=c("Geographic Area Name","id"))

matched<-matched[,!str_detect(colnames(matched),"Margin of Error")]
matched[,-c(1:2,4:5)]<-macthed<-sapply(matched[,-c(1:2,4:5)],as.numeric)
not_all_na <- function(x) any(!is.na(x))
matched<-matched%>%select_if(not_all_na)

## need to manipulate better
unique(aqi_data$State)[!(unique(aqi_data$State) %in% unique(matched$State))]

##lasso
```

```
nlambda <- 50
maxlambda <- 3
my.lambda.seq <- seq(maxlambda, maxlambda*0.01, length.out=nlambda)
set.seed(2)

matched<-matched%>%na.omit()
matched.cv<-matched%>%mutate(fold=sample(1:2,nrow(matched),replace=TRUE))
##splitted into testing and training set
train<-matched.cv%>%filter(fold==1)%>%select(-c("fold"))
test<-matched.cv%>%filter(fold==2)%>%select(-c("fold"))
train.mat<-as.matrix(train[,-c(1:2,4:5)])
test.mat<-as.matrix(test[,-c(1:2,4:5)])

## lambda selection
fit.cv <- cv.glmnet(x=train.mat[,-1], y=train.mat[,1], alpha=1, lambda=my.lambda.seq)
plot(fit.cv)

## fitting lasso with optimal lambda on training set
model<-glmnet(x=train.mat[,-1], y=train.mat[,1], alpha=1, lambda=fit.cv$lambda.min)
model$df
model_coef<-coef(model)
rownames(model_coef)[model_coef[,1]!=0] ## selected variable

## fit predictive model on testing set
aqi_predict<-predict(model,test.mat[,-1])
mse<-mean((aqi_predict-test.mat[1])^2) ##mean square error
```