**Background:**

The United States Clean Air Act is a federal law that was first passed in 1970 and amended in 1977 and 1990(1). The Clean Air Act requires the United States Environmental Protection Agency (EPA) to set National Ambient Air Quality Standards (NAAQS) for six air pollutants termed “criteria air pollutants”:  ground-level ozone, particulate matter, carbon monoxide, lead, sulfur dioxide, and nitrogen dioxide(1).

Air pollution is recognized as being harmful to human health. The establishment of federal limits on ambient exposures, however, requires the formulation of quantifiable exposures with acceptable risk profiles. The Clean Air Act also calls upon the EPA to periodically review and revise air pollution standards, reducing permissible levels if further scientific advances support such revisions (2). A key component of these reviews is the examination of risk/exposure assessment studies which provide insights into relationships between air pollutant exposures and human health effects.

For this project, we propose to use air quality monitoring information on ambient air pollution levels from the EPA-maintained AirData system, which reports air quality monitor measurement for locations across the United States. We then plan to examine relationships between criteria air pollutant levels and regional differences in chronic respiratory disease mortality rates at the county level.

**Data Sets:**

***Air Quality***

AirData is a website maintained by the EPA that provides public access to air quality data collected at more than 4000 outdoor monitors across the United States, Puerto Rico, and the United States Virgin Islands (3). AirData has available for download annual and daily summary data tables for measurements of criteria gases, particulates, meteorological conditions (wind, temperature, pressure, barometric pressure, and RH/dewpoint), toxics, ozone precursors, and lead measurements (4). These measurements are indexed by monitor site (state, county, monitor identification number) for the years from 1980-2017.

***Respiratory Health***

We propose to merge information on air quality with data on the United States Chronic Respiratory Disease Mortality Rates by County for the period 1980-2014. This aggregated data set is available through the Global Health Data Exchange. Age-standardized mortality rate for both sexes combined, reporting deaths per 100,000 people in the population.

Institute for Health Metrics and Evaluation summary of data set [5]:

“ IHME research produced estimates for age-standardized mortality rates by county from chronic respiratory diseases. The estimates were generated using de-identified death records from the National Center for Health Statistics (NCHS); population counts from the U.S. Census Bureau, NCHS, and the Human Mortality Database; the cause list from the Global Burden of Disease Study (GBD); and the application of small area estimation models. This dataset provides estimates for age-standardized mortality rates by disease type and sex at the county level for each state, the District of Columbia, and the United States as a whole for 1980-2014, as well as the changes in rates for each location during this period. Also included are data on the 10 counties with the highest and lowest mortality rates for each disease type in 2014. Study results were published in JAMA in September 2017 in "Trends and patterns of differences in chronic respiratory disease mortality among US counties, 1980–2014.”

**Variables:**

> combined\_mort\_y = data[,5]

Age-standardized county-level annual mortality rate from chronic respiratory diseases given as the rate per 100,000 persons in the population per year. Estimates were generated using death records obtained from the National Center for Health Statistics, population counts from the U.S Census Bureau, the National Center for Health Statistics, and the Human Mortality Database. Continuous variable.

> male\_mort\_y = data[,10]

As for combined gender rate above, age-standardized county-level annual mortality rate from chronic respiratory diseases given as the rate per 100,000 men in the population per year. Continuous variable.

> female\_mort\_y = data[,8]

As for combined gender and male rates above, age-standardized county-level annual mortality rate from chronic respiratory diseases given as the rate per 100,000 women in the population per year. Continuous variable.

> location = data[,3]

Categorical label comprised of the state and county names combined. Each gives a unique location for which mortality rate data and air quality data are available during the period 1980 – 2014. County-level mortality data and EPA air quality data are matched by location and year for predictive analysis in this project.

Categorical variable, n = XX.

> x1\_year = data[,11]

Year of observation data collection.

> days\_with\_AQI = data[,12]

Annual air quality summary data for each location and year is used here. In each annual summary file, the number of days on which data was recorded for each location is given – not every day of every year has a measurement for each location.

The Air Quality Index (AQI) is a summary metric used to report air quality. The EPA establishes and monitors compliance with national air quality standards. An AQI value below 100 corresponds to pollutant levels below the regulatory standard, with values below 50 being optimal.

AQI values above 100 are considered to be of increasing concern, with values in the range of 300-500 considered hazardous for the entire population.

Daily Air Quality Index (AQI) measurements may be viewed online: <https://airnow.gov/index.cfm?action=airnow.mapcenter&mapcenter=1>. AQI is reported for both combined and individual criteria pollutants.

For this analysis, we used the yearly AQI summary data files, which include counts of the number of recorded days on which measurements fell into each of the AQI categories, as well as counts of days on which the AQI was attributed to each criteria pollutant. These records also include the maximum, 90th percentile, and median AQI for that location and year.

> x2\_good\_days\_ratio = data[,13]

The proportion of days recorded where combined AQI was between 0-50.

> x3\_mod\_days\_ratio = data[,14]

The proportion of days recorded where combined AQI was between 51-100.

> x4\_unhealth\_sens\_ratio = data[,15]

The proportion of days recorded where combined AQI was between 101-200.

> x5\_unhealth\_ratio = data[,16]

The proportion of days recorded where combined AQI was between 201-300.

> x6\_very\_unhealth\_ratio = data[,17]

The proportion of days recorded where combined AQI was between 301-400.

> x7\_hazardous\_ratio = data[,18]

The proportion of days recorded where combined AQI was between 401-500.

> x8\_maxAQI = data[,19]

The maximum recorded overall AQI for that location and year.

> x9\_90percentileAQUI = data[,20]

The 90th percentile overall AQI for that location and year.

> x10\_median\_AQI = data[,21]

The median overall AQI for that location and year.

> x11\_CO\_ratio = data[,22]

Count of days on which the AQI was attributed to CO pollution.

Carbon monoxide (CO) is generated by combustion. Inhaling CO in high concentrations impairs the oxygen-carrying ability of the blood. This reduces the amount of oxygen being transported to the tissues of the body and places stress on the cardiovascular system.

> x12\_NO2\_ratio = data[,23]

Count of days on which the AQI was attributed to NO2 pollution.

Nitrogen dioxide is a reactive gas generated mainly from fuel combustion. Inhalation of high concentrations can cause respiratory irritation and respiratory symptoms such as coughing, wheezing, or shortness of breath. These effects may aggravate symptoms from pre-existing respiratory conditions. Prolonged exposure to nitrogen dioxide may contribute to the new development asthma and may increase risk of respiratory infection. Nitrogen oxide reactions also contribute to ground-level ozone levels.

> x13\_ozone\_ratio = data[,24]

Count of days on which the AQI was attributed to ozone pollution.

Ground-level ozone is generated as result of chemical reactions between nitrogen oxides and volatile organic compounds present in the air. Industrial emissions, fuel vapors, and vehicle exhaust are among potential sources of nitrogen oxides and volatile organic compounds. Ozone reactions are catalyzed by sunlight, so longer, brighter days may have higher ozone levels. Ozone inhalation may provoke respiratory symptoms and worsen symptoms due to pre-existing respiratory disease.

> x14\_SO2\_ratio = data[,25]

Count of days on which the AQI was attributed to SO2 pollution.

Sulfur dioxide is one of a number of sulfur oxide gases and is monitored and regulated as an indicator of sulfur oxides as a group. Sulfur oxide are released as a result of fuel combustion. Inhalation of sulfur oxides can result in respiratory system irritation and respiratory symptoms. Reactions between sulfur oxide and other compunds in air can generate particulate air pollution, which is further irritating to the lungs and may contribute to a number of health effects.

> x15\_PM2.5\_ratio = data[,26]

Count of days on which the AQI was attributed to PM 2.5 pollution.

Particulate air pollution includes a mixture of compounds present in air as small particle or liquid droplets. Fine particles (PM 2.5) are defined as those particles of diameter less than 2.5 micrometers. These very small particles are particularly hazardous, as their small size allows them to penetrate deeply into the progressively smaller air conduits in the lungs and potentially pass into the bloodstream. Previous research has linked particle air pollution exposure with adverse cardiovascular and respiratory effects, including increased symptoms from pre-exisiting conditions and increased risk of death.

> x16\_PM10\_ratio = data[,27]

Count of days on which the AQI was attributed to PM 10 pollution.

Particulate air pollution includes a mixture of compounds present in air as small particle or liquid droplets. Small particles (PM 10) are defined as those particles of diameter less than 10 micrometers. Particles of such small size may penetrate deeply into the lungs. As a general rule, the smaller the particle, the more potential it has for contributing to adverse health effects.

**Problem Statement:**

We are trying to predict annual respiratory mortality using air quality data.

This information is useful for such applications as

* Predicting the health impacts of new proposed industrial developments
* Selecting facility sites serving sensitive populations, such as schools for children or residential care facilities for older adults
* Predicting the benefits of more stringent air quality regulations
* Predicting health impacts of relocation

We can also use this data set to investigate historical impacts of air quality improvements following the establishment of regulations under the clean air act.

**Methods**

**Regression**

**Variable Summaries:**

> summary(data)

X index Unnamed..0

Min. : 0 Min. : 0.0 new york\_suffolk : 33

1st Qu.: 7850 1st Qu.: 222.0 colorado\_arapahoe : 32

Median :15530 Median : 443.0 montana\_rosebud : 32

Mean :15561 Mean : 461.4 ohio\_cuyahoga : 32

3rd Qu.:23354 3rd Qu.: 686.0 virginia\_fauquier : 32

Max. :31017 Max. :1077.0 california\_alameda: 31

(Other) :23072

FIPS both\_gender\_resp\_mort cause\_id

Min. : 1001 Min. : 15.18 Min. :508

1st Qu.:17179 1st Qu.: 45.81 1st Qu.:508

Median :30031 Median : 54.13 Median :508

Mean :29542 Mean : 55.31 Mean :508

3rd Qu.:42013 3rd Qu.: 63.36 3rd Qu.:508

Max. :56045 Max. :133.79 Max. :508

cause\_name female\_resp\_mort

Chronic respiratory diseases:23264 Min. : 7.865

1st Qu.: 32.928

Median : 42.167

Mean : 42.931

3rd Qu.: 51.743

Max. :114.820

location\_id male\_resp\_mort Year Days.with.AQI

Min. : 574 Min. : 17.67 Min. :1980 Min. : 1.0

1st Qu.:1260 1st Qu.: 63.79 1st Qu.:1991 1st Qu.:209.0

Median :2184 Median : 74.88 Median :2000 Median :355.0

Mean :2095 Mean : 75.47 Mean :1999 Mean :280.4

3rd Qu.:2833 3rd Qu.: 85.64 3rd Qu.:2007 3rd Qu.:365.0

Max. :3714 Max. :227.42 Max. :2014 Max. :366.0

Good.Days Moderate.Days

Min. :0.0000 Min. :0.0000

1st Qu.:0.5827 1st Qu.:0.1068

Median :0.7348 Median :0.2082

Mean :0.7103 Mean :0.2224

3rd Qu.:0.8767 3rd Qu.:0.3170

Max. :1.0000 Max. :0.8607

Unhealthy.for.Sensitive.Groups.Days Unhealthy.Days

Min. :0.00000 Min. :0.00000

1st Qu.:0.00000 1st Qu.:0.00000

Median :0.02459 Median :0.00000

Mean :0.05150 Mean :0.01353

3rd Qu.:0.07377 3rd Qu.:0.01370

Max. :0.66120 Max. :1.00000

Very.Unhealthy.Days Hazardous.Days Max.AQI

Min. :0.000000 Min. :0.000e+00 Min. : 0.0

1st Qu.:0.000000 1st Qu.:0.000e+00 1st Qu.: 100.0

Median :0.000000 Median :0.000e+00 Median : 147.0

Mean :0.002108 Mean :3.784e-05 Mean : 148.9

3rd Qu.:0.000000 3rd Qu.:0.000e+00 3rd Qu.: 185.0

Max. :0.438356 Max. :7.812e-02 Max. :20646.0

X90th.Percentile.AQI Median.AQI Days.CO

Min. : 0.00 Min. : 0.00 Min. :0.00000

1st Qu.: 54.00 1st Qu.: 30.00 1st Qu.:0.00000

Median : 73.00 Median : 40.00 Median :0.00000

Mean : 77.42 Mean : 39.14 Mean :0.03228

3rd Qu.: 98.00 3rd Qu.: 46.00 3rd Qu.:0.00000

Max. :347.00 Max. :200.00 Max. :1.00000

Days.NO2 Days.Ozone Days.SO2

Min. :0.00000 Min. :0.0000 Min. :0.0000

1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.0000

Median :0.00000 Median :0.4294 Median :0.0000

Mean :0.04059 Mean :0.4421 Mean :0.1515

3rd Qu.:0.00000 3rd Qu.:0.8221 3rd Qu.:0.1328

Max. :1.00000 Max. :1.0000 Max. :1.0000

Days.PM2.5 Days.PM10

Min. :0.0000 Min. :0.00000

1st Qu.:0.0000 1st Qu.:0.00000

Median :0.0000 Median :0.00000

Mean :0.1955 Mean :0.13803

3rd Qu.:0.2816 3rd Qu.:0.04098

Max. :1.0000 Max. :1.00000

**Full Model Regression:**

Combined Mortality Rate: R2 = 0.8876; Adjusted R2 = 0.88

Male Mortality Rate: R2 = 0.9093; Adjusted R2 = 0.9032

Female Mortality Rate: R2 = 0.8826; Adjusted R2 = 0.8747

This model includes location, but we have a high number of unique location categories.

**Best Subset Selection:**

Best subset selection using regsubsets with all variables generates an error.

Regsubsets Error:

in leaps.exhaustive(a, really.big) :

Exhaustive search will be S L O W, must specify really.big=T

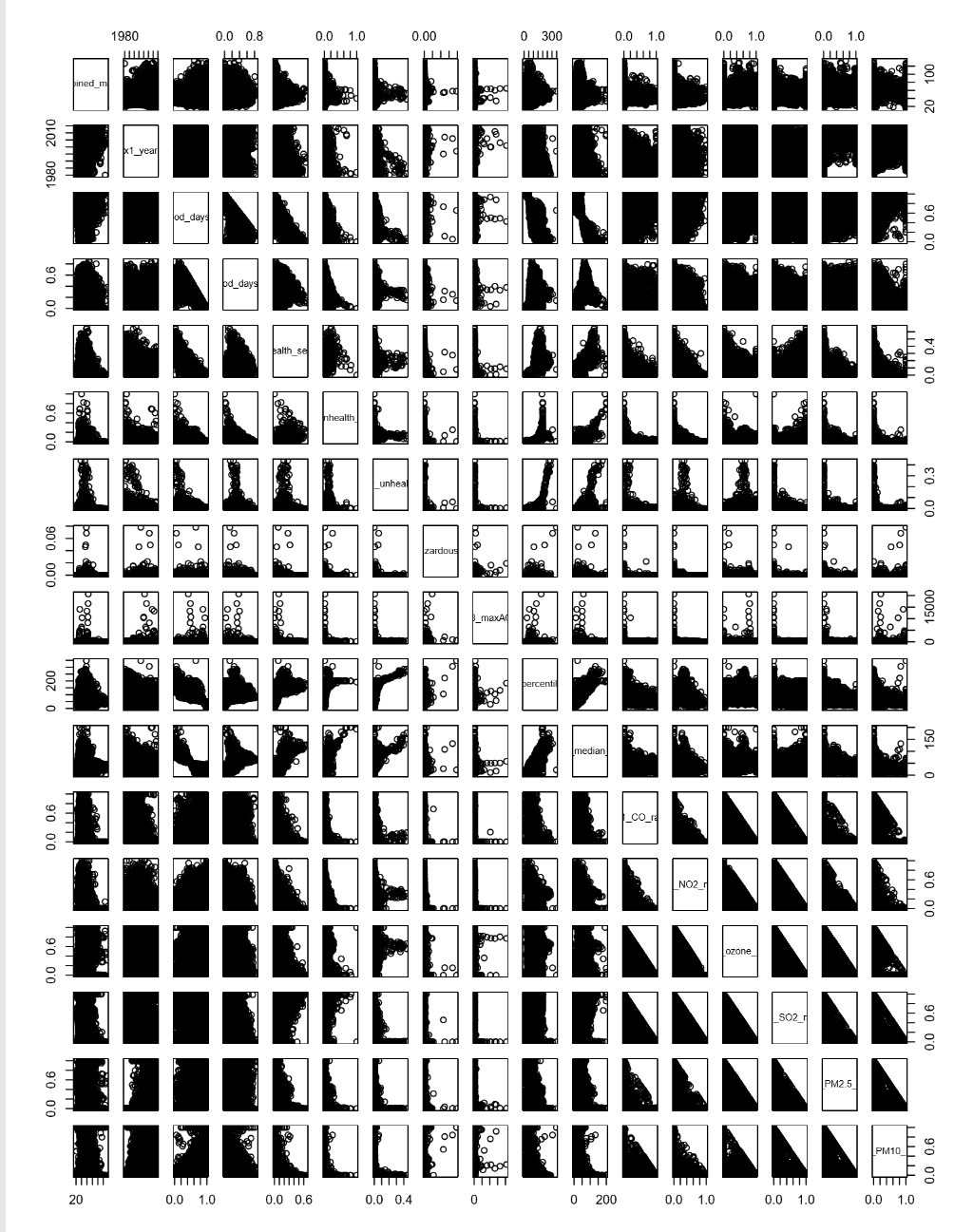
In addition: Warning message:

In leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in = force.in, :

1 linear dependencies found

For regression, then, we will leave out location as a variable.

**Data Exploration (location left out as factor)**



|  |
| --- |
| **Correlations**  combined\_mort\_y male\_mort\_y female\_mort\_y x1\_year  combined\_mort\_y 1.000000000 0.830652712 0.906617398 0.407998564  male\_mort\_y 0.830652712 1.000000000 0.528057746 -0.048225326  female\_mort\_y 0.906617398 0.528057746 1.000000000 0.652105303  x1\_year 0.407998564 -0.048225326 0.652105303 1.000000000  x2\_good\_days\_ratio 0.078452467 0.041302393 0.066004030 0.132646776  x3\_mod\_days\_ratio 0.014263054 -0.035095992 0.072342165 0.065360143  x4\_unhealth\_sens\_ratio -0.174396615 -0.020327314 -0.236116169 -0.362482242  x5\_unhealth\_ratio -0.146302990 -0.035678167 -0.196578943 -0.280955885  x6\_very\_unhealth\_ratio -0.072992464 -0.044394862 -0.074784933 -0.140188474  x7\_hazardous\_ratio 0.002308101 -0.006947582 0.005659898 0.008548379  x8\_maxAQI -0.048833338 -0.047691197 -0.033706092 -0.041687125  x9\_90percentileAQUI -0.160439335 -0.049160399 -0.186548988 -0.297990753  x10\_median\_AQI -0.075247485 -0.060812452 -0.052706273 -0.090577366  x11\_CO\_ratio -0.142913898 -0.068919585 -0.160980540 -0.243917041  x12\_NO2\_ratio -0.170521623 -0.137321015 -0.145033097 -0.155995469  x13\_ozone\_ratio -0.053922525 -0.129042483 0.019325506 0.090237317  x14\_SO2\_ratio -0.131930168 0.075226010 -0.255410108 -0.366563808  x15\_PM2.5\_ratio 0.224932595 0.003035080 0.334602473 0.481726839  x16\_PM10\_ratio 0.075263053 0.161651795 -0.014430715 -0.112596991  days\_with\_AQI -0.119773280 -0.195356641 -0.026148277 0.043649240  x2\_good\_days\_ratio x3\_mod\_days\_ratio  combined\_mort\_y 0.07845247 0.014263054  male\_mort\_y 0.04130239 -0.035095992  female\_mort\_y 0.06600403 0.072342165  x1\_year 0.13264678 0.065360143  x2\_good\_days\_ratio 1.00000000 -0.888459503  x3\_mod\_days\_ratio -0.88845950 1.000000000  x4\_unhealth\_sens\_ratio -0.75676330 0.407037139  x5\_unhealth\_ratio -0.51141177 0.143784884  x6\_very\_unhealth\_ratio -0.24409313 0.072898445  x7\_hazardous\_ratio -0.03887775 0.014809544  x8\_maxAQI -0.15289233 0.110736279  x9\_90percentileAQUI -0.84250994 0.580485918  x10\_median\_AQI -0.89102873 0.719713598  x11\_CO\_ratio -0.02286350 0.033456762  x12\_NO2\_ratio -0.15404206 0.129996822  x13\_ozone\_ratio -0.06686925 0.054125210  x14\_SO2\_ratio -0.19019573 -0.006490388  x15\_PM2.5\_ratio -0.06779723 0.258146914  x16\_PM10\_ratio 0.39256423 -0.384815254  days\_with\_AQI -0.33702875 0.288951374  x4\_unhealth\_sens\_ratio x5\_unhealth\_ratio  combined\_mort\_y -0.174396615 -0.146302990  male\_mort\_y -0.020327314 -0.035678167  female\_mort\_y -0.236116169 -0.196578943  x1\_year -0.362482242 -0.280955885  x2\_good\_days\_ratio -0.756763303 -0.511411769  x3\_mod\_days\_ratio 0.407037139 0.143784884  x4\_unhealth\_sens\_ratio 1.000000000 0.651815860  x5\_unhealth\_ratio 0.651815860 1.000000000  x6\_very\_unhealth\_ratio 0.206325857 0.308255870  x7\_hazardous\_ratio 0.031726829 0.045777207  x8\_maxAQI 0.138126416 0.099937233  x9\_90percentileAQUI 0.825674848 0.697480702  x10\_median\_AQI 0.730860115 0.609515665  x11\_CO\_ratio -0.007747638 -0.003640731  x12\_NO2\_ratio 0.105393515 0.080123302  x13\_ozone\_ratio 0.055632889 0.030602406  x14\_SO2\_ratio 0.442184101 0.272400384  x15\_PM2.5\_ratio -0.251118637 -0.165285698  x16\_PM10\_ratio -0.256658948 -0.148550709  days\_with\_AQI 0.282703979 0.164735178  x6\_very\_unhealth\_ratio x7\_hazardous\_ratio x8\_maxAQI  combined\_mort\_y -0.07299246 0.0023081008 -0.048833338  male\_mort\_y -0.04439486 -0.0069475819 -0.047691197  female\_mort\_y -0.07478493 0.0056598984 -0.033706092  x1\_year -0.14018847 0.0085483791 -0.041687125  x2\_good\_days\_ratio -0.24409313 -0.0388777508 -0.152892329  x3\_mod\_days\_ratio 0.07289844 0.0148095440 0.110736279  x4\_unhealth\_sens\_ratio 0.20632586 0.0317268287 0.138126416  x5\_unhealth\_ratio 0.30825587 0.0457772066 0.099937233  x6\_very\_unhealth\_ratio 1.00000000 0.0372569203 0.065159947  x7\_hazardous\_ratio 0.03725692 1.0000000000 0.267848842  x8\_maxAQI 0.06515995 0.2678488423 1.000000000  x9\_90percentileAQUI 0.38846272 0.0797407524 0.189093877  x10\_median\_AQI 0.32343525 0.0385251289 0.129456957  x11\_CO\_ratio 0.02917345 -0.0005531548 0.005023113  x12\_NO2\_ratio 0.16102289 -0.0077553366 0.031835165  x13\_ozone\_ratio 0.05603399 -0.0184283109 0.071499075  x14\_SO2\_ratio -0.01247065 -0.0131565355 0.037343746  x15\_PM2.5\_ratio -0.07069531 -0.0156222995 -0.064747598  x16\_PM10\_ratio -0.05337440 0.0539418753 -0.069622960  days\_with\_AQI 0.07304694 0.0029415329 0.128017055    x9\_90percentileAQUI x10\_median\_AQI x11\_CO\_ratio  combined\_mort\_y -0.160439335 -0.075247485 -0.1429138980  male\_mort\_y -0.049160399 -0.060812452 -0.0689195845  female\_mort\_y -0.186548988 -0.052706273 -0.1609805402  x1\_year -0.297990753 -0.090577366 -0.2439170410  x2\_good\_days\_ratio -0.842509935 -0.891028733 -0.0228635024  x3\_mod\_days\_ratio 0.580485918 0.719713598 0.0334567622  x4\_unhealth\_sens\_ratio 0.825674848 0.730860115 -0.0077476376  x5\_unhealth\_ratio 0.697480702 0.609515665 -0.0036407310  x6\_very\_unhealth\_ratio 0.388462724 0.323435250 0.0291734463  x7\_hazardous\_ratio 0.079740752 0.038525129 -0.0005531548  x8\_maxAQI 0.189093877 0.129456957 0.0050231131  x9\_90percentileAQUI 1.000000000 0.796693834 -0.0044619778  x10\_median\_AQI 0.796693834 1.000000000 0.0074898973  x11\_CO\_ratio -0.004461978 0.007489897 1.0000000000  x12\_NO2\_ratio 0.151651814 0.149555108 0.0222259820  x13\_ozone\_ratio 0.241733620 0.253802171 -0.1225320910  x14\_SO2\_ratio 0.261628306 0.050846873 -0.0523775156  x15\_PM2.5\_ratio -0.186913003 -0.026407818 -0.1334417386  x16\_PM10\_ratio -0.401344587 -0.388707274 -0.0668960343  days\_with\_AQI 0.365195233 0.331837777 0.1367568701    x12\_NO2\_ratio x13\_ozone\_ratio x14\_SO2\_ratio  combined\_mort\_y -0.170521623 -0.05392253 -0.131930168  male\_mort\_y -0.137321015 -0.12904248 0.075226010  female\_mort\_y -0.145033097 0.01932551 -0.255410108  x1\_year -0.155995469 0.09023732 -0.366563808  x2\_good\_days\_ratio -0.154042056 -0.06686925 -0.190195732  x3\_mod\_days\_ratio 0.129996822 0.05412521 -0.006490388  x4\_unhealth\_sens\_ratio 0.105393515 0.05563289 0.442184101  x5\_unhealth\_ratio 0.080123302 0.03060241 0.272400384  x6\_very\_unhealth\_ratio 0.161022888 0.05603399 -0.012470652  x7\_hazardous\_ratio -0.007755337 -0.01842831 -0.013156535  x8\_maxAQI 0.031835165 0.07149908 0.037343746  x9\_90percentileAQUI 0.151651814 0.24173362 0.261628306  x10\_median\_AQI 0.149555108 0.25380217 0.050846873  x11\_CO\_ratio 0.022225982 -0.12253209 -0.052377516  x12\_NO2\_ratio 1.000000000 -0.06407119 -0.014715465  x13\_ozone\_ratio -0.064071190 1.00000000 -0.363048569  x14\_SO2\_ratio -0.014715465 -0.36304857 1.000000000  x15\_PM2.5\_ratio -0.137217225 -0.37536865 -0.267835841  x16\_PM10\_ratio -0.129264837 -0.43720622 -0.182976432  days\_with\_AQI 0.239040479 0.24206144 0.271393292  x15\_PM2.5\_ratio x16\_PM10\_ratio days\_with\_AQI  combined\_mort\_y 0.22493259 0.07526305 -0.119773280  male\_mort\_y 0.00303508 0.16165179 -0.195356641  female\_mort\_y 0.33460247 -0.01443071 -0.026148277  x1\_year 0.48172684 -0.11259699 0.043649240  x2\_good\_days\_ratio -0.06779723 0.39256423 -0.337028753  x3\_mod\_days\_ratio 0.25814691 -0.38481525 0.288951374  x4\_unhealth\_sens\_ratio -0.25111864 -0.25665895 0.282703979  x5\_unhealth\_ratio -0.16528570 -0.14855071 0.164735178  x6\_very\_unhealth\_ratio -0.07069531 -0.05337440 0.073046937  x7\_hazardous\_ratio -0.01562230 0.05394188 0.002941533  x8\_maxAQI -0.06474760 -0.06962296 0.128017055  x9\_90percentileAQUI -0.18691300 -0.40134459 0.365195233  x10\_median\_AQI -0.02640782 -0.38870727 0.331837777  x11\_CO\_ratio -0.13344174 -0.06689603 0.136756870  x12\_NO2\_ratio -0.13721723 -0.12926484 0.239040479  x13\_ozone\_ratio -0.37536865 -0.43720622 0.242061442  x14\_SO2\_ratio -0.26783584 -0.18297643 0.271393292  x15\_PM2.5\_ratio 1.00000000 -0.21335672 -0.145676387  x16\_PM10\_ratio -0.21335672 1.00000000 -0.539965205  days\_with\_AQI -0.14567639 -0.53996521 1.000000000 |
|  |
| |  | | --- | | **Initial Regression (location variable left out)** | |

Linear regression was performed using the combined male and female age-adjusted respiratory mortality rate, the male age-adjusted mortality rate, and the female age-adjusted mortality rate as dependent variables in three separate regressions. Summary outputs are presented below. Statistically significant coefficients are highlighted and R2 values are highlighted.

Observations:

R2 and adjusted R2 are quite different when comparing regression that use male versus female respiratory mortality rates. The overall AQI day ratio coefficients do not have strong t-statistics in comparison with the counts of days on which AQI is attributable to specific criteria pollutants. We know that we have a fair amount of correlation between our predictor variables, so this is not surprising.

> summary(full\_comb)

Call:

lm(formula = combined\_mort\_y ~ x1\_year + x2\_good\_days\_ratio +

x3\_mod\_days\_ratio + x4\_unhealth\_sens\_ratio + x5\_unhealth\_ratio +

x6\_very\_unhealth\_ratio + x7\_hazardous\_ratio + x8\_maxAQI +

x9\_90percentileAQUI + x10\_median\_AQI + x11\_CO\_ratio + x12\_NO2\_ratio +

x13\_ozone\_ratio + x14\_SO2\_ratio + x15\_PM2.5\_ratio + x16\_PM10\_ratio +

days\_with\_AQI, data = dataf)

Residuals:

Min 1Q Median 3Q Max

-50.957 -7.325 -0.565 7.018 84.050

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.168e+03 1.338e+02 -8.735 < 2e-16 \*\*\*

x1\_year 5.881e-01 1.124e-02 52.323 < 2e-16 \*\*\*

x2\_good\_days\_ratio 4.977e+01 1.312e+02 0.379 0.705

x3\_mod\_days\_ratio 4.517e+01 1.312e+02 0.344 0.731

x4\_unhealth\_sens\_ratio 2.841e+01 1.312e+02 0.217 0.829

x5\_unhealth\_ratio 9.443e+00 1.312e+02 0.072 0.943

x6\_very\_unhealth\_ratio 2.798e+01 1.313e+02 0.213 0.831

x7\_hazardous\_ratio -5.976e+01 2.051e+02 -0.291 0.771

x8\_maxAQI -7.266e-04 4.611e-04 -1.576 0.115

x9\_90percentileAQUI 6.184e-02 7.647e-03 8.087 6.44e-16 \*\*\*

x10\_median\_AQI 9.659e-02 1.441e-02 6.703 2.09e-11 \*\*\*

x11\_CO\_ratio -5.812e+00 7.299e-01 -7.963 1.76e-15 \*\*\*

x12\_NO2\_ratio -1.336e+01 8.060e-01 -16.577 < 2e-16 \*\*\*

x13\_ozone\_ratio -5.824e+00 3.792e-01 -15.360 < 2e-16 \*\*\*

x14\_SO2\_ratio -3.127e-01 4.518e-01 -0.692 0.489

x15\_PM2.5\_ratio -3.479e+00 3.821e-01 -9.104 < 2e-16 \*\*\*

x16\_PM10\_ratio NA NA NA NA

days\_with\_AQI -1.238e-02 9.564e-04 -12.942 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.84 on 23247 degrees of freedom

Multiple R-squared: 0.205, Adjusted R-squared: 0.2045

F-statistic: 374.7 on 16 and 23247 DF, p-value: < 2.2e-16

> summary(male\_comb)

Call:

lm(formula = male\_mort\_y ~ x1\_year + x2\_good\_days\_ratio + x3\_mod\_days\_ratio +

x4\_unhealth\_sens\_ratio + x5\_unhealth\_ratio + x6\_very\_unhealth\_ratio +

x7\_hazardous\_ratio + x8\_maxAQI + x9\_90percentileAQUI + x10\_median\_AQI +

x11\_CO\_ratio + x12\_NO2\_ratio + x13\_ozone\_ratio + x14\_SO2\_ratio +

x15\_PM2.5\_ratio + x16\_PM10\_ratio + days\_with\_AQI, data = dataf)

Residuals:

Min 1Q Median 3Q Max

-62.213 -11.027 -0.757 9.582 154.860

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.305e+01 1.898e+02 -0.121 0.9034

x1\_year -2.898e-02 1.595e-02 -1.816 0.0693 .

x2\_good\_days\_ratio 1.584e+02 1.863e+02 0.851 0.3950

x3\_mod\_days\_ratio 1.452e+02 1.862e+02 0.780 0.4355

x4\_unhealth\_sens\_ratio 1.097e+02 1.862e+02 0.589 0.5557

x5\_unhealth\_ratio 7.306e+01 1.862e+02 0.392 0.6948

x6\_very\_unhealth\_ratio 7.518e+01 1.863e+02 0.404 0.6865

x7\_hazardous\_ratio -5.975e+01 2.912e+02 -0.205 0.8374

x8\_maxAQI -1.289e-03 6.544e-04 -1.970 0.0489 \*

x9\_90percentileAQUI 1.262e-01 1.085e-02 11.626 < 2e-16 \*\*\*

x10\_median\_AQI 2.273e-01 2.045e-02 11.111 < 2e-16 \*\*\*

x11\_CO\_ratio -1.138e+01 1.036e+00 -10.987 < 2e-16 \*\*\*

x12\_NO2\_ratio -2.116e+01 1.144e+00 -18.499 < 2e-16 \*\*\*

x13\_ozone\_ratio -9.618e+00 5.382e-01 -17.873 < 2e-16 \*\*\*

x14\_SO2\_ratio 3.524e+00 6.413e-01 5.495 3.95e-08 \*\*\*

x15\_PM2.5\_ratio -5.994e+00 5.423e-01 -11.053 < 2e-16 \*\*\*

x16\_PM10\_ratio NA NA NA NA

days\_with\_AQI -2.689e-02 1.357e-03 -19.812 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.81 on 23247 degrees of freedom

Multiple R-squared: 0.08257, Adjusted R-squared: 0.08194

F-statistic: 130.8 on 16 and 23247 DF, p-value: < 2.2e-16

> summary(female\_comb)

Call:

lm(formula = female\_mort\_y ~ x1\_year + x2\_good\_days\_ratio + x3\_mod\_days\_ratio +

x4\_unhealth\_sens\_ratio + x5\_unhealth\_ratio + x6\_very\_unhealth\_ratio +

x7\_hazardous\_ratio + x8\_maxAQI + x9\_90percentileAQUI + x10\_median\_AQI +

x11\_CO\_ratio + x12\_NO2\_ratio + x13\_ozone\_ratio + x14\_SO2\_ratio +

x15\_PM2.5\_ratio + x16\_PM10\_ratio + days\_with\_AQI, data = dataf)

Residuals:

Min 1Q Median 3Q Max

-44.798 -6.136 -0.403 6.001 59.518

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.846e+03 1.151e+02 -16.046 < 2e-16 \*\*\*

x1\_year 9.373e-01 9.670e-03 96.928 < 2e-16 \*\*\*

x2\_good\_days\_ratio 1.625e+01 1.129e+02 0.144 0.885567

x3\_mod\_days\_ratio 1.727e+01 1.129e+02 0.153 0.878429

x4\_unhealth\_sens\_ratio 9.407e+00 1.129e+02 0.083 0.933578

x5\_unhealth\_ratio -8.110e+00 1.129e+02 -0.072 0.942718

x6\_very\_unhealth\_ratio 2.416e+01 1.129e+02 0.214 0.830561

x7\_hazardous\_ratio -1.024e+02 1.765e+02 -0.580 0.561918

x8\_maxAQI -2.295e-04 3.967e-04 -0.579 0.562853

x9\_90percentileAQUI 4.215e-02 6.579e-03 6.406 1.52e-10 \*\*\*

x10\_median\_AQI 2.111e-02 1.240e-02 1.702 0.088702 .

x11\_CO\_ratio -2.127e+00 6.280e-01 -3.387 0.000708 \*\*\*

x12\_NO2\_ratio -7.771e+00 6.934e-01 -11.207 < 2e-16 \*\*\*

x13\_ozone\_ratio -3.888e+00 3.262e-01 -11.918 < 2e-16 \*\*\*

x14\_SO2\_ratio -2.725e+00 3.887e-01 -7.009 2.46e-12 \*\*\*

x15\_PM2.5\_ratio -2.440e+00 3.287e-01 -7.424 1.17e-13 \*\*\*

x16\_PM10\_ratio NA NA NA NA

days\_with\_AQI -4.012e-03 8.228e-04 -4.876 1.09e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.19 on 23247 degrees of freedom

Multiple R-squared: 0.4363, Adjusted R-squared: 0.4359

F-statistic: 1125 on 16 and 23247 DF, p-value: < 2.2e-16

Before we move on to model selection, we can examine some hypotheses about relationships between our predictors, too. This gives us an idea of how predictor variables may be changing over the period of our study.

**Criteria Pollutants versus Year**

R2 values for these regressions are not particularly high (highlighted in outputs). Not all criteria air pollutants demonstrate a negative linear association with increasing year. 95% confidence intervals for the coefficients of year for the outcome variables of interest in these simple linear regressions are presented. Note that year does not have a negative coefficient in regressions for ozone, PM2.5, and female mortality.

> confint(lm.CO)

2.5 % 97.5 %

x1\_year -0.003355114 -0.003028917

> confint(lm.NO2)

2.5 % 97.5 %

x1\_year -0.001974636 -0.001677451

> confint(lm.ozone)

2.5 % 97.5 %

x1\_year 0.003154478 0.00419721

> confint(lm.SO2)

2.5 % 97.5 %

x1\_year -0.01165493 -0.01091862

> confint(lm.PM2.5)

2.5 % 97.5 %

x1\_year 0.01597084 0.01673545

> confint(lm.PM10)

2.5 % 97.5 %

x1\_year -0.00414651 -0.0033018

> confint(lm.male\_mort)

2.5 % 97.5 %

x1\_year -0.1122382 -0.06504892

> confint(lm.female\_mort)

2.5 % 97.5 %

x1\_year 0.913105 0.9408038

> confint(lm.medAQI)

2.5 % 97.5 %

x1\_year -0.1854171 -0.1395058

> summary(lm.CO)

Call:

lm(formula = x11\_CO\_ratio ~ x1\_year, data = dataf)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.413e+00 1.663e-01 38.55 <2e-16 \*\*\*

x1\_year -3.192e-03 8.321e-05 -38.36 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1211 on 23262 degrees of freedom

Multiple R-squared: 0.0595, Adjusted R-squared: 0.05946

F-statistic: 1472 on 1 and 23262 DF, p-value: < 2.2e-16

> summary(lm.NO2)

Call:

lm(formula = x12\_NO2\_ratio ~ x1\_year, data = dataf)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.691e+00 1.515e-01 24.36 <2e-16 \*\*\*

x1\_year -1.826e-03 7.581e-05 -24.09 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1104 on 23262 degrees of freedom

Multiple R-squared: 0.02433, Adjusted R-squared: 0.02429

F-statistic: 580.2 on 1 and 23262 DF, p-value: < 2.2e-16

> summary(lm.ozone)

Call:

lm(formula = x13\_ozone\_ratio ~ x1\_year, data = dataf)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.905486 0.531700 -12.99 <2e-16 \*\*\*

x1\_year 0.003676 0.000266 13.82 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3872 on 23262 degrees of freedom

Multiple R-squared: 0.008143, Adjusted R-squared: 0.0081

F-statistic: 191 on 1 and 23262 DF, p-value: < 2.2e-16

> summary(lm.SO2)

Call:

lm(formula = x14\_SO2\_ratio ~ x1\_year, data = dataf)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 22.7125741 0.3754552 60.49 <2e-16 \*\*\*

x1\_year -0.0112868 0.0001878 -60.09 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2734 on 23262 degrees of freedom

Multiple R-squared: 0.1344, Adjusted R-squared: 0.1343

F-statistic: 3611 on 1 and 23262 DF, p-value: < 2.2e-16

> summary(lm.PM2.5)

Call:

lm(formula = x15\_PM2.5\_ratio ~ x1\_year, data = dataf)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -32.492760 0.389884 -83.34 <2e-16 \*\*\*

x1\_year 0.016353 0.000195 83.84 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2839 on 23262 degrees of freedom

Multiple R-squared: 0.2321, Adjusted R-squared: 0.232

F-statistic: 7029 on 1 and 23262 DF, p-value: < 2.2e-16

> summary(lm.PM10)

Call:

lm(formula = x16\_PM10\_ratio ~ x1\_year, data = dataf)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.5822229 0.4307267 17.60 <2e-16 \*\*\*

x1\_year -0.0037242 0.0002155 -17.28 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3137 on 23262 degrees of freedom

Multiple R-squared: 0.01268, Adjusted R-squared: 0.01264

F-statistic: 298.7 on 1 and 23262 DF, p-value: < 2.2e-16

**Median AQI versus Year**

> summary(lm.medAQI)

Call:

lm(formula = x10\_median\_AQI ~ x1\_year, data = dataf)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 363.88064 23.41063 15.54 <2e-16 \*\*\*

x1\_year -0.16246 0.01171 -13.87 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.05 on 23262 degrees of freedom

Multiple R-squared: 0.008204, Adjusted R-squared: 0.008162

F-statistic: 192.4 on 1 and 23262 DF, p-value: < 2.2e-16

**Respiratory Mortality by Gender versus Year**

> lm.male\_mort = lm(male\_mort\_y ~

+ x1\_year, data = dataf)

> summary(lm.male\_mort)

Call:

lm(formula = male\_mort\_y ~ x1\_year, data = dataf)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 252.65513 24.06234 10.500 < 2e-16 \*\*\*

x1\_year -0.08864 0.01204 -7.364 1.85e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.52 on 23262 degrees of freedom

Multiple R-squared: 0.002326, Adjusted R-squared: 0.002283

F-statistic: 54.23 on 1 and 23262 DF, p-value: 1.846e-13

> lm.female\_mort =lm(female\_mort\_y ~

+ x1\_year, data = dataf)

> summary(lm.female\_mort)

Call:

lm(formula = female\_mort\_y ~ x1\_year, data = dataf)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.810e+03 1.412e+01 -128.1 <2e-16 \*\*\*

x1\_year 9.270e-01 7.066e-03 131.2 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.29 on 23262 degrees of freedom

Multiple R-squared: 0.4252, Adjusted R-squared: 0.4252

F-statistic: 1.721e+04 on 1 and 23262 DF, p-value: < 2.2e-16

**Can a linear model predict respiratory mortality from yearly air quality?**

**Step 1 – Select a model using forward and backward stepwise subset selection and Lasso.**

**Forward Stepwise Subset Selection:**

> step(null\_comb, scope=list(lower=null\_comb, upper=full\_comb), direction="forward",level="0.01")

Call:

lm(formula = combined\_mort\_y ~ x1\_year + days\_with\_AQI + x12\_NO2\_ratio +

x13\_ozone\_ratio + x11\_CO\_ratio + x15\_PM2.5\_ratio + x3\_mod\_days\_ratio +

x5\_unhealth\_ratio + x9\_90percentileAQUI + x4\_unhealth\_sens\_ratio +

x10\_median\_AQI + x2\_good\_days\_ratio + x8\_maxAQI, data = data)

Coefficients:

(Intercept) x1\_year days\_with\_AQI

-1.141e+03 5.885e-01 -1.267e-02

x12\_NO2\_ratio x13\_ozone\_ratio x11\_CO\_ratio

-1.315e+01 -5.653e+00 -5.628e+00

x15\_PM2.5\_ratio x3\_mod\_days\_ratio x5\_unhealth\_ratio

-3.361e+00 1.748e+01 -1.817e+01

x9\_90percentileAQUI x4\_unhealth\_sens\_ratio x10\_median\_AQI

6.103e-02 4.222e-01 9.785e-02

x2\_good\_days\_ratio x8\_maxAQI

2.206e+01 -8.814e-04

> step(null\_male, scope=list(lower=null\_comb, upper=male\_comb), direction="forward",level="0.01")

Call:

lm(formula = male\_mort\_y ~ days\_with\_AQI + x14\_SO2\_ratio + x12\_NO2\_ratio +

x16\_PM10\_ratio + x3\_mod\_days\_ratio + x2\_good\_days\_ratio +

x9\_90percentileAQUI + x10\_median\_AQI + x15\_PM2.5\_ratio +

x4\_unhealth\_sens\_ratio + x8\_maxAQI + x7\_hazardous\_ratio,

data = data)

Coefficients:

(Intercept) days\_with\_AQI x14\_SO2\_ratio

-1.923e+01 -2.775e-02 1.355e+01

x12\_NO2\_ratio x16\_PM10\_ratio x3\_mod\_days\_ratio

-1.101e+01 9.810e+00 7.284e+01

x2\_good\_days\_ratio x9\_90percentileAQUI x10\_median\_AQI

8.676e+01 1.327e-01 2.303e-01

x15\_PM2.5\_ratio x4\_unhealth\_sens\_ratio x8\_maxAQI

3.562e+00 3.739e+01 -1.464e-03

x7\_hazardous\_ratio

-1.789e+02

> step(null\_female, scope=list(lower=null\_comb, upper=female\_comb), direction="forward",level="0.01")

Call:

lm(formula = female\_mort\_y ~ x1\_year + x16\_PM10\_ratio + x3\_mod\_days\_ratio +

x12\_NO2\_ratio + days\_with\_AQI + x6\_very\_unhealth\_ratio +

x5\_unhealth\_ratio + x9\_90percentileAQUI + x13\_ozone\_ratio +

x2\_good\_days\_ratio + x4\_unhealth\_sens\_ratio + x10\_median\_AQI,

data = data)

Coefficients:

(Intercept) x1\_year x16\_PM10\_ratio

-1.917e+03 9.378e-01 2.516e+00

x3\_mod\_days\_ratio x12\_NO2\_ratio days\_with\_AQI

8.408e+01 -5.241e+00 -4.160e-03

x6\_very\_unhealth\_ratio x5\_unhealth\_ratio x9\_90percentileAQUI

9.108e+01 5.848e+01 4.173e-02

x13\_ozone\_ratio x2\_good\_days\_ratio x4\_unhealth\_sens\_ratio

-1.368e+00 8.303e+01 7.541e+01

x10\_median\_AQI

2.374e-02

**Backward Stepwise Subset Selection:**

> step(full\_comb, data = dataf, direction="backward",level="0.05")

Call:

lm(formula = combined\_mort\_y ~ x1\_year + x2\_good\_days\_ratio +

x3\_mod\_days\_ratio + x4\_unhealth\_sens\_ratio + x6\_very\_unhealth\_ratio +

x8\_maxAQI + x9\_90percentileAQUI + x10\_median\_AQI + x11\_CO\_ratio +

x12\_NO2\_ratio + x13\_ozone\_ratio + x15\_PM2.5\_ratio + days\_with\_AQI,

data = dataf)

Coefficients:

(Intercept) x1\_year x2\_good\_days\_ratio

-1.160e+03 5.888e-01 4.049e+01

x3\_mod\_days\_ratio x4\_unhealth\_sens\_ratio x6\_very\_unhealth\_ratio

3.589e+01 1.885e+01 1.907e+01

x8\_maxAQI x9\_90percentileAQUI x10\_median\_AQI

-8.093e-04 6.114e-02 9.802e-02

x11\_CO\_ratio x12\_NO2\_ratio x13\_ozone\_ratio

-5.631e+00 -1.316e+01 -5.664e+00

x15\_PM2.5\_ratio days\_with\_AQI

-3.366e+00 -1.268e-02

> step(male\_comb, data = dataf, direction="backward",level="0.05")

Call:

lm(formula = male\_mort\_y ~ x1\_year + x2\_good\_days\_ratio + x3\_mod\_days\_ratio +

x4\_unhealth\_sens\_ratio + x8\_maxAQI + x9\_90percentileAQUI +

x10\_median\_AQI + x11\_CO\_ratio + x12\_NO2\_ratio + x13\_ozone\_ratio +

x14\_SO2\_ratio + x15\_PM2.5\_ratio + days\_with\_AQI, data = dataf)

Coefficients:

(Intercept) x1\_year x2\_good\_days\_ratio

53.365299 -0.030277 84.636132

x3\_mod\_days\_ratio x4\_unhealth\_sens\_ratio x8\_maxAQI

71.570949 36.156277 -0.001632

x9\_90percentileAQUI x10\_median\_AQI x11\_CO\_ratio

0.125156 0.226474 -11.348707

x12\_NO2\_ratio x13\_ozone\_ratio x14\_SO2\_ratio

-21.066126 -9.536502 3.569725

x15\_PM2.5\_ratio days\_with\_AQI

-5.946726 -0.026866

> step(female\_comb, data = dataf, direction="backward",level="0.05")

Call:

lm(formula = female\_mort\_y ~ x1\_year + x2\_good\_days\_ratio + x3\_mod\_days\_ratio +

x4\_unhealth\_sens\_ratio + x6\_very\_unhealth\_ratio + x7\_hazardous\_ratio +

x9\_90percentileAQUI + x10\_median\_AQI + x11\_CO\_ratio + x12\_NO2\_ratio +

x13\_ozone\_ratio + x14\_SO2\_ratio + x15\_PM2.5\_ratio + days\_with\_AQI,

data = dataf)

Coefficients:

(Intercept) x1\_year x2\_good\_days\_ratio

-1.855e+03 9.374e-01 2.451e+01

x3\_mod\_days\_ratio x4\_unhealth\_sens\_ratio x6\_very\_unhealth\_ratio

2.547e+01 1.760e+01 3.231e+01

x7\_hazardous\_ratio x9\_90percentileAQUI x10\_median\_AQI

-1.059e+02 4.204e-02 2.176e-02

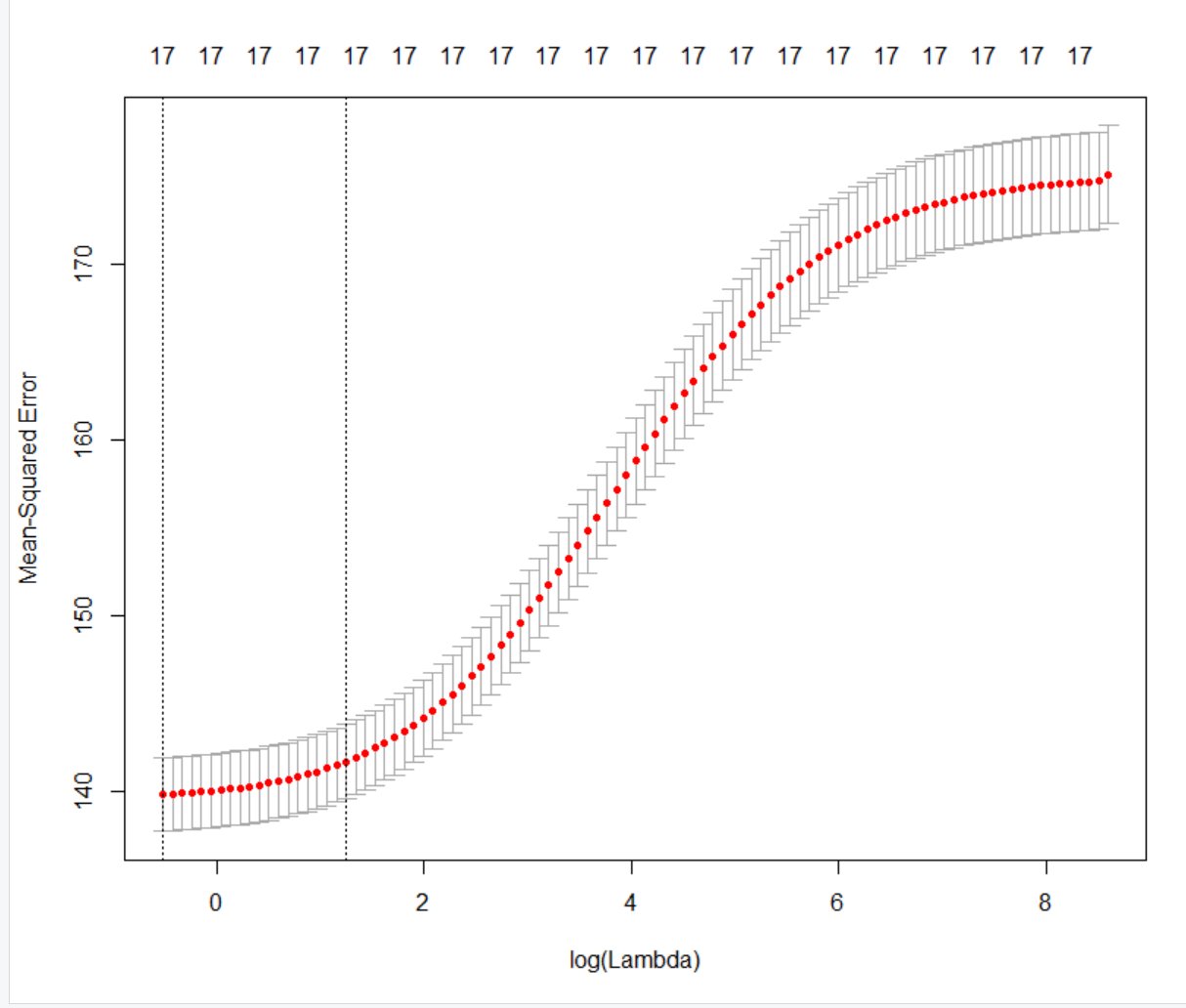
x11\_CO\_ratio x12\_NO2\_ratio x13\_ozone\_ratio

-2.121e+00 -7.766e+00 -3.894e+00

x14\_SO2\_ratio x15\_PM2.5\_ratio days\_with\_AQI

-2.716e+00 -2.433e+00 -4.060e-03

**Ridge Regression**



**Combined respiratory mortality**

> bestlam=cv.out$lambda.min

> bestlam

[1] 0.5971866

> ridge.pred=predict(ridge.mod, s=bestlam, newx=x[test,])

> mean((ridge.pred-y.test)^2)

[1] 141.577

> out=glmnet(x,y,alpha=0)

> predict(out, type="coefficients", s=bestlam)[1:19,]

(Intercept) (Intercept) x1\_year

-1.021282e+03 0.000000e+00 5.383199e-01

x2\_good\_days\_ratio x3\_mod\_days\_ratio x4\_unhealth\_sens\_ratio

1.107954e+00 1.693433e+00 -9.473412e+00

x5\_unhealth\_ratio x6\_very\_unhealth\_ratio x7\_hazardous\_ratio

-2.105888e+01 -3.601318e+00 -5.486378e+01

x8\_maxAQI x9\_90percentileAQUI x10\_median\_AQI

-8.368965e-04 2.666447e-02 4.535376e-02

x11\_CO\_ratio x12\_NO2\_ratio x13\_ozone\_ratio

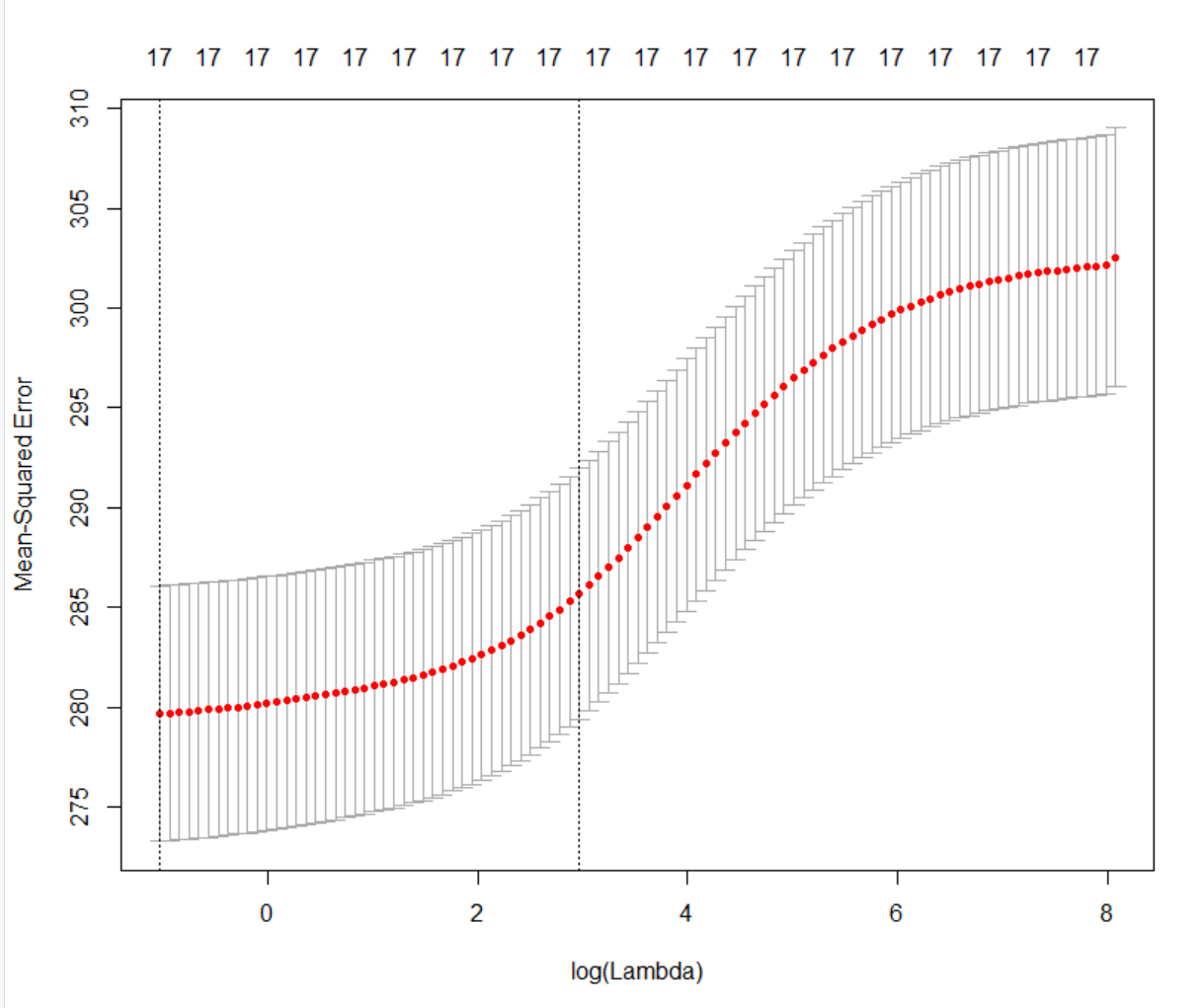
-3.568211e+00 -1.028769e+01 -1.856889e+00

x14\_SO2\_ratio x15\_PM2.5\_ratio x16\_PM10\_ratio

1.911653e+00 9.467993e-02 2.874644e+00

days\_with\_AQI

-1.033569e-02



**Male respiratory mortality**

> bestlam=cv.out$lambda.min

> bestlam

[1] 0.3552032

> ridge.pred=predict(ridge.mod, s=bestlam, newx=x[test,])

> mean((ridge.pred-y.test)^2)

[1] 286.8669

> out=glmnet(x,y,alpha=0)

> predict(out, type="coefficients", s=bestlam)[1:19,]

(Intercept) (Intercept) x1\_year

1.388993e+02 0.000000e+00 -3.451204e-02

x2\_good\_days\_ratio x3\_mod\_days\_ratio x4\_unhealth\_sens\_ratio

5.851006e+00 9.906756e-02 -2.677949e+01

x5\_unhealth\_ratio x6\_very\_unhealth\_ratio x7\_hazardous\_ratio

-5.406903e+01 -5.112200e+01 -1.982218e+02

x8\_maxAQI x9\_90percentileAQUI x10\_median\_AQI

-1.681878e-03 8.354906e-02 1.475071e-01

x11\_CO\_ratio x12\_NO2\_ratio x13\_ozone\_ratio

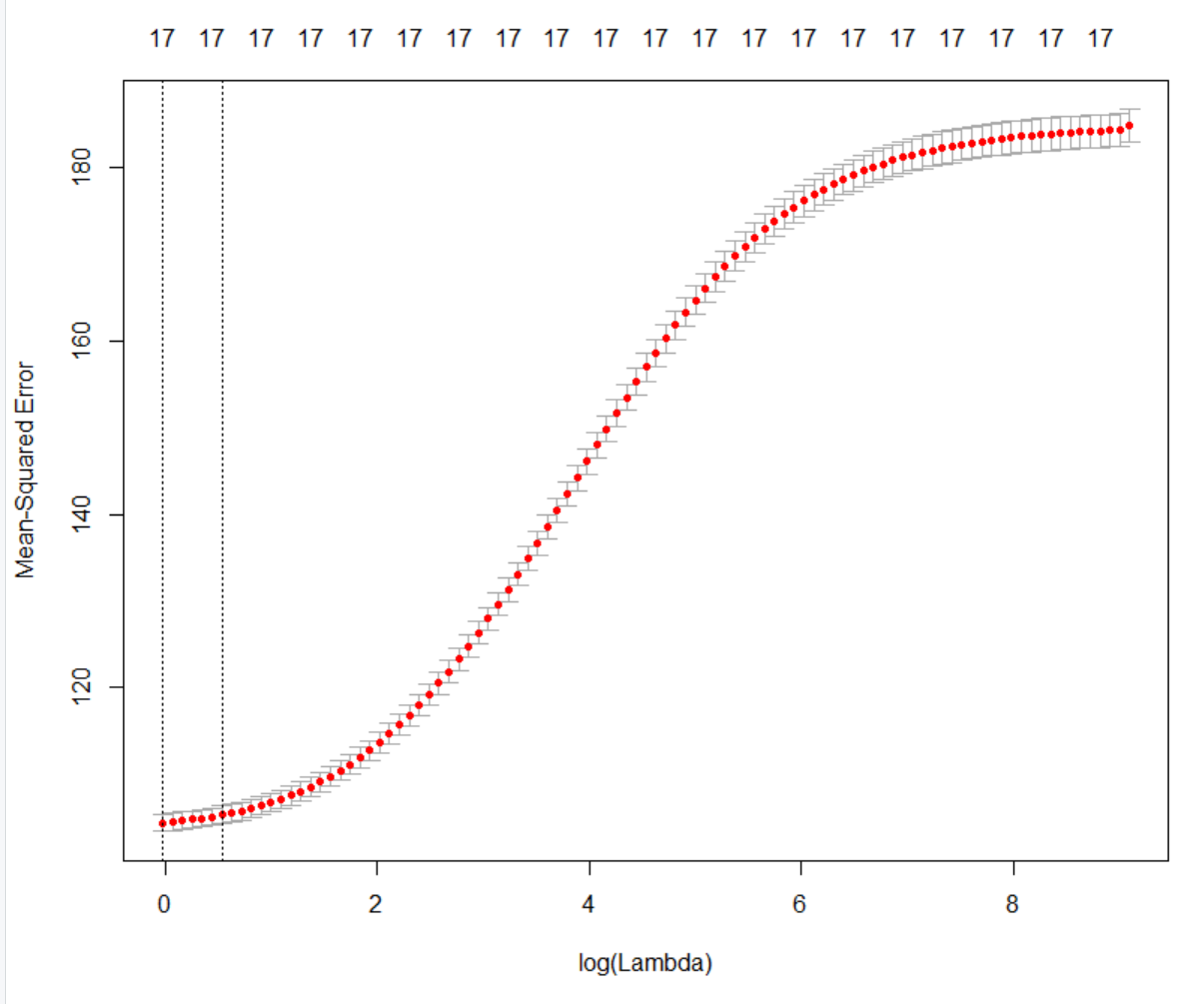
-6.972678e+00 -1.630263e+01 -4.015698e+00

x14\_SO2\_ratio x15\_PM2.5\_ratio x16\_PM10\_ratio

7.230254e+00 -1.468097e+00 4.344960e+00

days\_with\_AQI

-2.545165e-02



**Female respiratory mortality**

> bestlam=cv.out$lambda.min

> bestlam

[1] 0.9765151

> ridge.pred=predict(ridge.mod, s=bestlam, newx=x[test,])

> mean((ridge.pred-y.test)^2)

[1] 104.5619

> out=glmnet(x,y,alpha=0)

> predict(out, type="coefficients", s=bestlam)[1:19,]

(Intercept) (Intercept) x1\_year

-1.608256e+03 0.000000e+00 8.258597e-01

x2\_good\_days\_ratio x3\_mod\_days\_ratio x4\_unhealth\_sens\_ratio

-6.786916e-01 2.912720e+00 -3.676532e+00

x5\_unhealth\_ratio x6\_very\_unhealth\_ratio x7\_hazardous\_ratio

-1.485452e+01 1.369839e+01 -4.680835e+01

x8\_maxAQI x9\_90percentileAQUI x10\_median\_AQI

-2.487873e-04 1.317417e-02 8.824142e-03

x11\_CO\_ratio x12\_NO2\_ratio x13\_ozone\_ratio

-1.615451e+00 -6.004540e+00 -8.067617e-01

x14\_SO2\_ratio x15\_PM2.5\_ratio x16\_PM10\_ratio

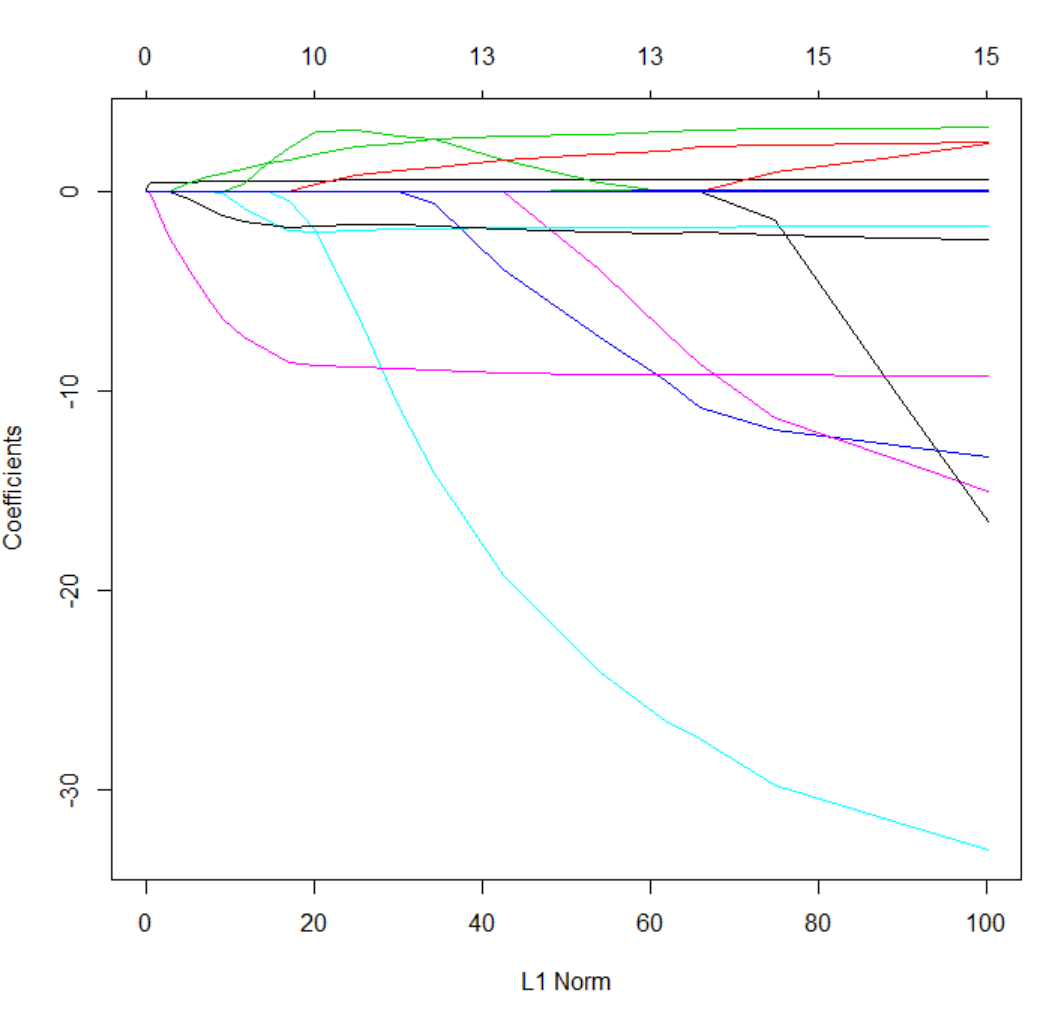
-1.094516e+00 1.037653e+00 2.086712e+00

days\_with\_AQI

-1.507977e-03

**Lasso**

We saw that Ridge can outperform least squares for predicting mortality rates. We now apply Lasso to determine if we can identify a more accurate or interpretable model.

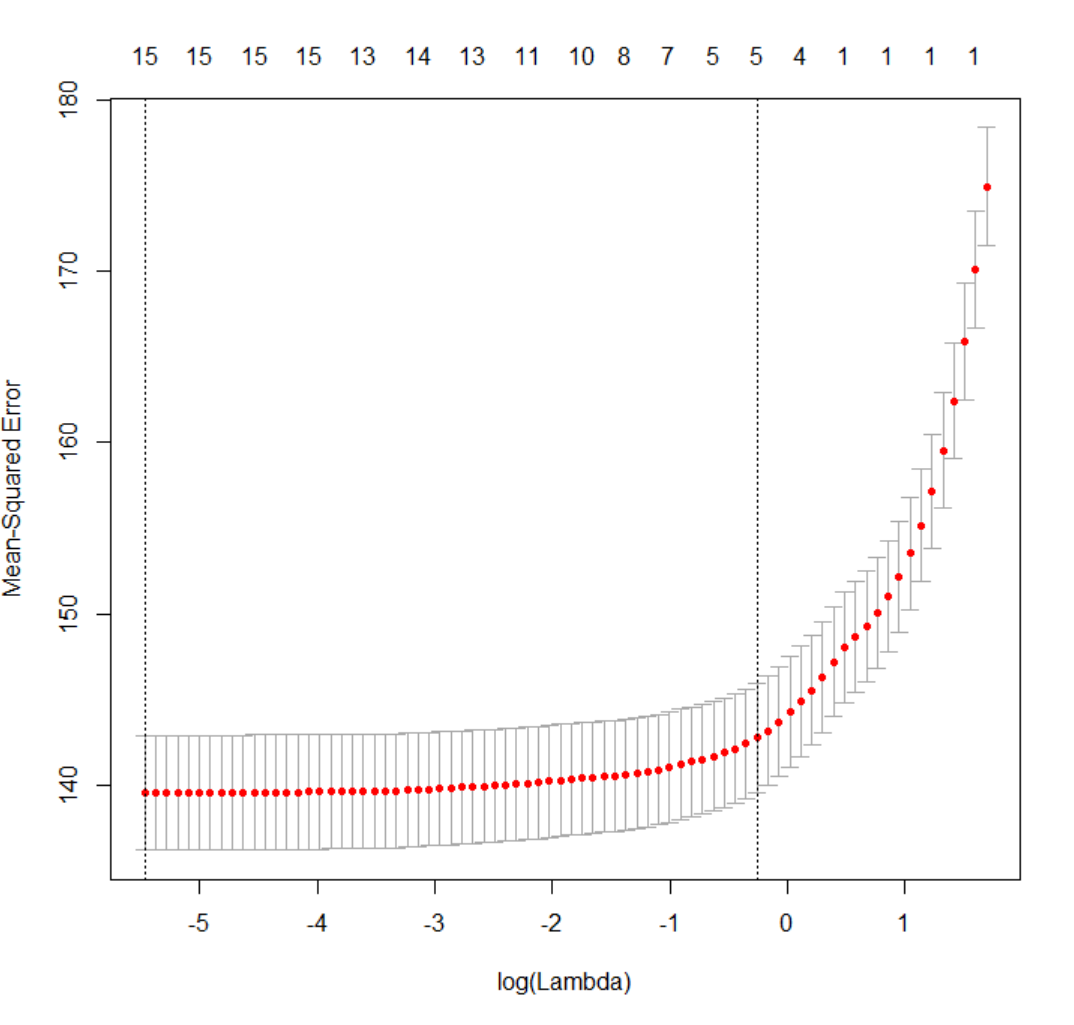


> lasso.mod = glmnet(x[train,],y[train], alpha=1, lambda=grid)

> plot(lasso.mod)

(combined mortality rate)

**Combined Respiratory Mortality**



> mean((lasso.pred-y.test)^2)

[1] 268.9173

> lasso.coef

(Intercept) (Intercept) x1\_year

-1.117770e+03 0.000000e+00 5.851923e-01

x2\_good\_days\_ratio x3\_mod\_days\_ratio x4\_unhealth\_sens\_ratio

2.423071e+00 0.000000e+00 -1.472461e+01

x5\_unhealth\_ratio x6\_very\_unhealth\_ratio x7\_hazardous\_ratio

-3.079316e+01 -1.166931e+01 -9.654342e+01

x8\_maxAQI x9\_90percentileAQUI x10\_median\_AQI

-8.289708e-04 5.010716e-02 7.430389e-02

x11\_CO\_ratio x12\_NO2\_ratio x13\_ozone\_ratio

-2.329848e+00 -9.792814e+00 -2.025703e+00

x14\_SO2\_ratio x15\_PM2.5\_ratio x16\_PM10\_ratio

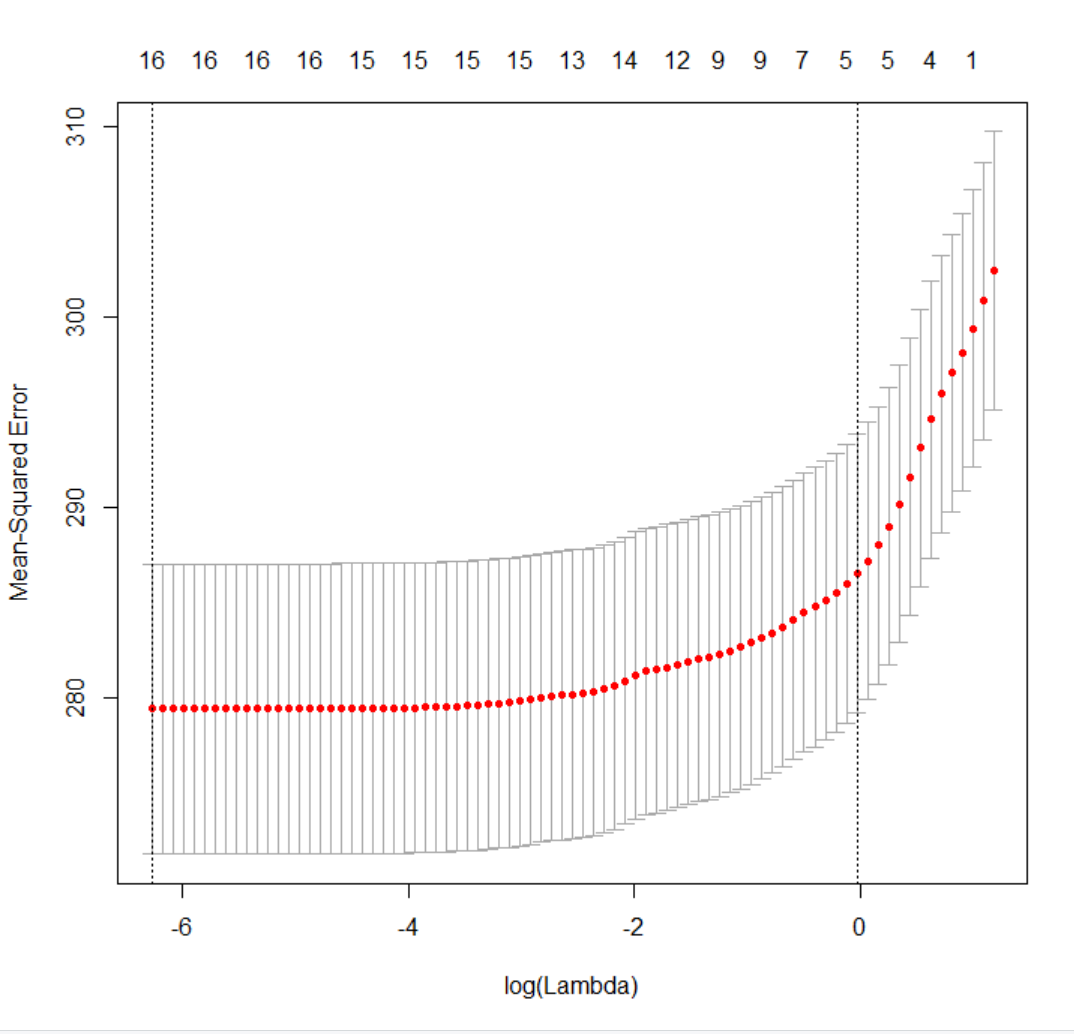
2.972398e+00 0.000000e+00 3.418261e+00

days\_with\_AQI

-1.209842e-02

Note that the moderate days ratio and PM2.5 ratio are assigned zero coefficients in this model.

**Male Respiratory Mortality**



> mean((lasso.pred-y.test)^2)

[1] 1265.471

> lasso.coef

(Intercept) (Intercept) x1\_year

1.176809e+02 0.000000e+00 -2.823352e-02

x2\_good\_days\_ratio x3\_mod\_days\_ratio x4\_unhealth\_sens\_ratio

1.097300e+01 0.000000e+00 -3.354436e+01

x5\_unhealth\_ratio x6\_very\_unhealth\_ratio x7\_hazardous\_ratio

-6.710369e+01 -6.447628e+01 -2.401990e+02

x8\_maxAQI x9\_90percentileAQUI x10\_median\_AQI

-1.654329e-03 1.150446e-01 2.042818e-01

x11\_CO\_ratio x12\_NO2\_ratio x13\_ozone\_ratio

-5.264371e+00 -1.498156e+01 -3.255491e+00

x14\_SO2\_ratio x15\_PM2.5\_ratio x16\_PM10\_ratio

9.408075e+00 0.000000e+00 5.979231e+00

days\_with\_AQI

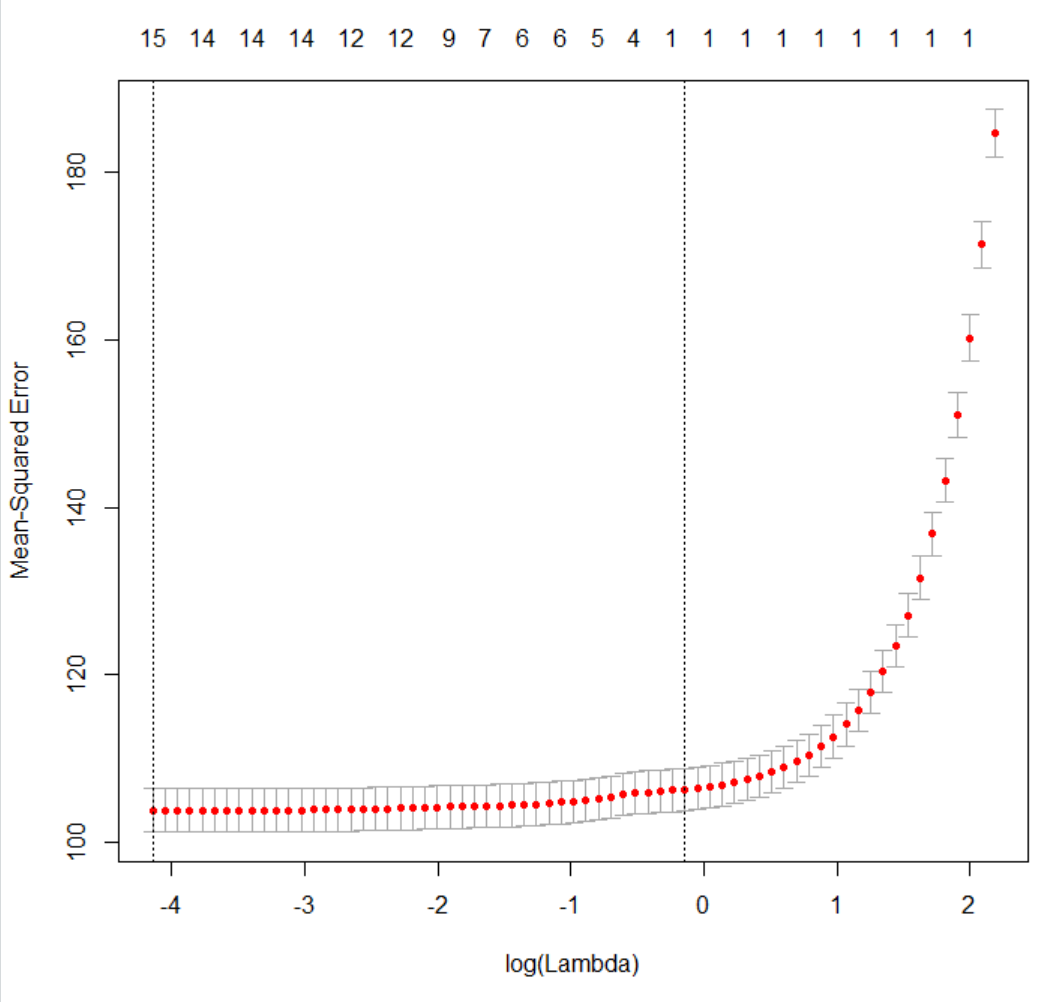
-2.668347e-02

Again moderate days ratio and PM2.5 are assigned zero coefficients in this model.

**Female Respiratory Mortality**

> mean((lasso.pred-y.test)^2)

[1] 104.0363



> mean((lasso.pred-y.test)^2)

[1] 104.0363

> lasso.coef

(Intercept) (Intercept) x1\_year

-1.824775e+03 0.000000e+00 9.335795e-01

x2\_good\_days\_ratio x3\_mod\_days\_ratio x4\_unhealth\_sens\_ratio

0.000000e+00 2.848291e+00 -2.800187e+00

x5\_unhealth\_ratio x6\_very\_unhealth\_ratio x7\_hazardous\_ratio

-1.746902e+01 1.318261e+01 -8.841096e+01

x8\_maxAQI x9\_90percentileAQUI x10\_median\_AQI

-2.262383e-04 3.031276e-02 1.698825e-03

x11\_CO\_ratio x12\_NO2\_ratio x13\_ozone\_ratio

5.182185e-01 -4.737965e+00 -7.206051e-01

x14\_SO2\_ratio x15\_PM2.5\_ratio x16\_PM10\_ratio

0.000000e+00 4.041797e-01 2.760086e+00

days\_with\_AQI

-3.725950e-03

Lasso identifies different coefficients as predictors of female respiratory mortality. Here the only predictors assigned zero coefficients are the ratio of good days and the ratio of days elevated sulfur dioxide contributes to elevated AQI.

Of particular note, air pollution data achieves lower test MSE when used to predict female respiratory mortality in comparison with predicting male respiratory mortality.

**Regression Summary**

Linear regression was performed using the combined male and female age-adjusted respiratory mortality rate, the male age-adjusted mortality rate, and the female age-adjusted mortality rate as dependent variables in three separate regressions. In the initial regression analysese, R2 and adjusted R2 were observed to be quite different. The summary AQI day ratio coefficients do not have strong t-statistics in comparison with the counts of days on which AQI is attributable to specific criteria pollutants. Forward and backward best subset selection were then tried. Ridge and Lasso regression were tried to see if least squares regression could be outperformed for prediction. In these models, we found that there were slight differences in the predictors selected for male and female respiratory mortality rates and that female respiratory mortality could be more accurately predicted from yearly air pollution data, as reflected by test MSE.