Apply linear regression models to study associaiton between air pollution and age-adjusted-mortality using the SMSA dataset

Eliza Chai

02/15/2022

Responses

1. The age-adjusted-mortality rate is a weighted average of the age-specific mortality rates per 100,000 people, where the weights are the proportions of people in that age groups compared to a standard population. Age-standardized-mortality rates adjust for differences in age distribution of the population. It is calculated by multiplying the age-specific mortality rates (divide the number of deaths by the respective population) for each age group by 100,000.

2. In Figure 1 we observed an increasing relationship between log(NOx) and age-adjusted mortality rate: the mean age-adjusted mortality rate appears to increase as log(NOx) increases. However, we observe some evidence of non-linearity over the range of the full log(NOx) observed. There are also a number of outlying values for both age-adjusted mortality rate and log(NOx).

3. We fit a simple linear regression with log(NOx) as the predictor and age-adjusted mortality as the response using heteroskedasticity-robust standard errors and performed tests at a 5% significance level. We estimate that the difference in the mean age-adjusted mortality rate between two populations differing in log(NOx) by one log pollution potential unit is 15.099 deaths per 100,000 residents, with the higher nitrous oxide pollution having higher mortality (95% CI: -0.882, 31.080 deaths per 100,000 residents). The intercept is 905.613 deaths per 100,000 residents at zero log pollution potential units in log(NOx) (i.e. 1 pollution potential unit in NOx). We fail to reject the null hypothesis of no association between age-adjusted mortality and log(NOx) (p = 0.0636) and conclude that there is no first-order linear association between mortality and nitrous oxide pollution.

4. In Figure 2, we observed a difference in grouping between low rainfall cities and non-low rainfall cities, where the low rainfall groups have lower age-adjusted mortality and higher log(NOx). Therefore, rainfall could be an effect modifier as it differentiates the association between predictor (log(NOx)) and response (age-adjusted mortality) for the differing category of effect modifying variable (low rainfall vs. non-low rainfall).

5. The fitted model is as follows:

$\hat{Age - adjustedMortality_i}|(log(NOx)_i,\ rainfall_i)$ = 874.490 + 34.490 $*$ $log(NOx)_i$ - 56.546 $*$ $1_{rainfall_i}$ - 25.506 $*$ $1_{rainfall_i} * log(NOx)_i$.

$\hat{Age - adjustedMortality_i}$ is the fitted value of total age-adjusted all-cause mortality rate, measured in deaths per 100,000 residents. $log(NOx)_i$ is the log-transformed Nitrous Oxide pollution potential measured in log pollution potential units. $1_{rainfall_i} = 1$ for ith low rainfall city whose mean annual rainfall is under 20 inches of rain per year, $1_{rainfall_i} = 0$ otherwise. $1_{rainfall_i} * log(NOx)_i$ is the log-transformed Nitrous Oxide

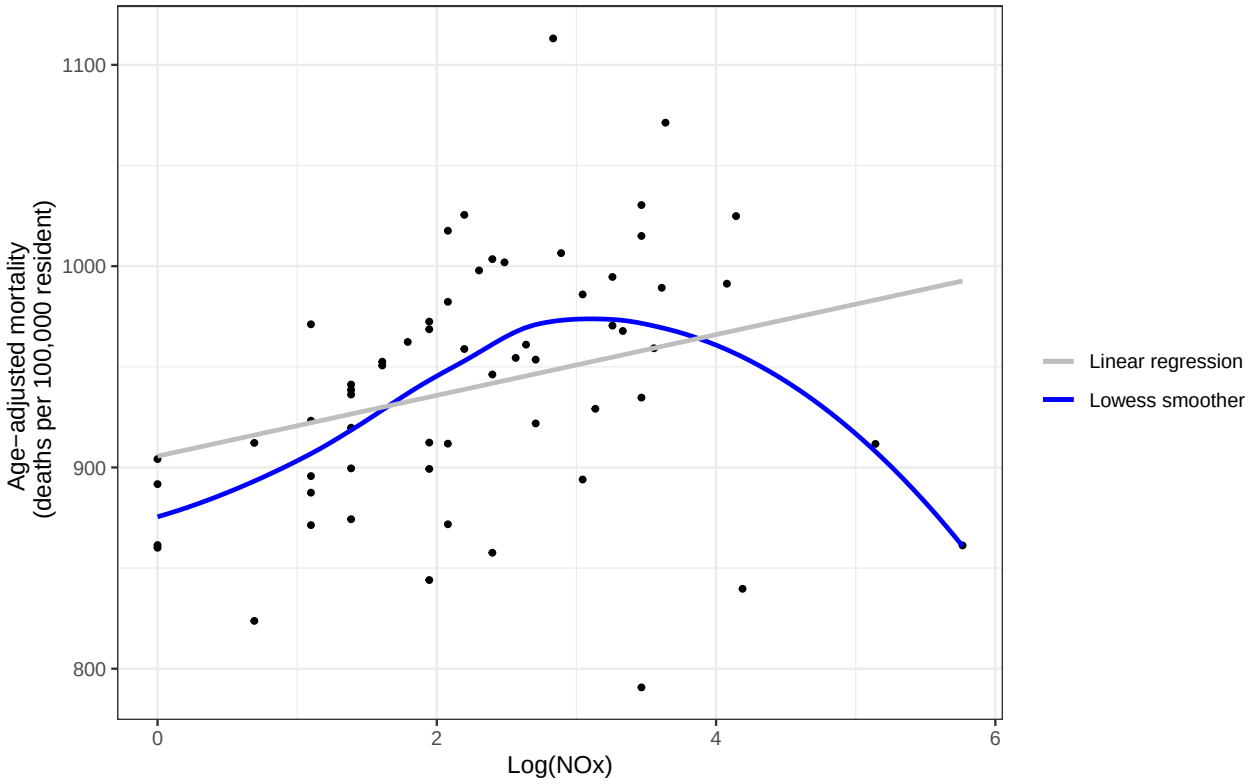

Figure 1: Scatterplot of age-adjusted mortality versus log(NOx) in the SMSA cohort

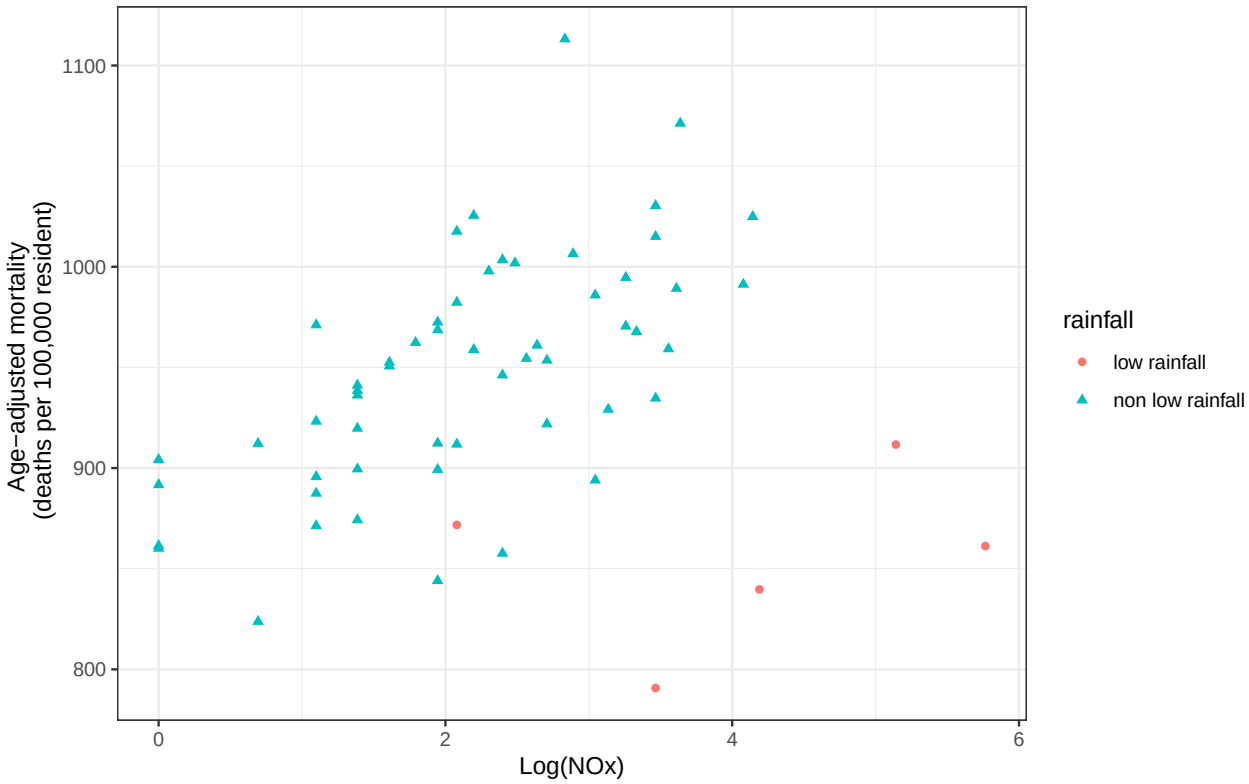

Figure 2: Scatterplot of age-adjusted mortality versus log(NOx) by low raninfall in the SMSA cohort

pollution potential (measured in log pollution potential units) for ith city with low rainfall (< 20 inches of rain per year).

We fit a multiple linear regression model for age-adjusted mortality, including log(NOx), rainfall, and an interaction between these two variables as predictors. We estimate that when comparing groups that differ in log(NOx) by one log pollution potential unit, the difference in mean age-adjusted mortality between two group of low rainfall cities would be 25.506 deaths per 100,000 residents lower than the difference in mean age-adjusted mortality between two group of non-low rainfalls cities (95% CI based on heteroskedasticity-robust standard errors: -51.150, 0.138).

The estimated difference in the expected age-adjusted mortality between two groups (one non-low rainfall group and one low rainfall group) of log(NOx) being zero is 56.546 deaths per 100,000 residents, with the zero-log(NOx) low rainfall group having lower estimated mean age-adjusted mortality than zero-log(NOx) non-low rainfall group.

The estimated difference in mean age-adjusted mortality between two groups of low rainfall cities who differ in log(NOx) by one log pollution potential unit would be 8.984 deaths per 100,000 residents, with the higher pollutant group having higher age-adjusted mortality (95% CI based on robust standard errors: -14.839, 32.807). The estimated difference in mean age-adjusted mortality between two groups of non-low rainfall cities who differ in log(NOx) by one log pollution potential unit would be 34.490 deaths per 100,000 residents, with the higher pollutant group having higher age-adjusted mortality (95% CI based on robust standard errors: 25.000, 43.980).

The estimated expected value of age-adjusted mortality for zero log(NOx) and non-low rainfall groups is 874.490 deaths per 100,000 residents. The estimated expected value of age-adjusted mortality for zero log(NOx) and low rainfall groups is 817.944 deaths per 100,000 residents.

6. Based on the fitted multiple linear regression model for age-adjusted mortality, including log(NOx), rainfall, and an interaction between these two variables as predictors, we estimate that when comparing groups that differ in log(NOx) by 1 log pollution potential unit, the difference in mean age-adjusted mortality between two group of low rainfall cities would be 25.506 deaths per 100,000 residents lower than the difference in mean age-adjusted mortality between two group of non-low rainfalls cities. We fail to reject the null hypothesis that the difference in mean age-adjusted mortality for groups who differ in log(NOx) by 1 log pollution potential unit is equal for low rainfall and non-low rainfall groups (95 CI for difference in differences based on robust standard errors: (-51.150, 0.138), p = 0.0512).

The estimated difference in mean age-adjusted mortality between two groups of low rainfall cities who differ in log(NOx) by one log pollution potential unit would be 8.984 deaths per 100,000 residents, with the higher pollutant group having higher age-adjusted mortality (95% CI based on robust standard error: -14.839, 32.807). Furthermore, the estimated difference in mean age-adjusted mortality between two groups of non-low rainfall cities who differ in log(NOx) by one log pollution potential unit would be 34.490 deaths per 100,000 residents, with the higher pollutant group having higher age-adjusted mortality (95% CI based on robust standard errors: 25.000, 43.980).

We find a statistically significant association between age-adjusted mortality and log(NOx) in the non-low rainfall group ($p < 0.00005$), but no statistically significant association between age-adjusted mortality and log(NOx) in the low rainfall group ($p = 0.453$). Therefore, we have a strong evidence for a first order linear trend of higher nitrous oxide pollution associated with higher mortality among those with non-low rainfall, but no linear trend in the relationship between mortality and nitrous oxide pollution in the low rainfall group.

7. I would not conclude that there is a causal effect of nitrous oxide pollution on mortality since multiple linear regression models only allow us to explore the association between variables. We would need to form a causal model to include potential confounders to study the causal effect of nitrous oxide pollution on mortality.

8. Since Los Angeles, CA, gets about 13 inches of rainfall per year, it would be considered as a low rainfall group. Based on the fitted multiple linear regression model for age-adjusted mortality, including log(NOx), rainfall, and an interaction between these two variables as predictors, the estimated difference in mean age-adjusted mortality between two groups of low rainfall cities who differ in log(NOx) by one log pollution potential unit would be 8.984 deaths per 100,000 residents, with the higher pollutant group having higher age-adjusted mortality. We find no statistically significant association between age-adjusted mortality and log(NOx) in the low rainfall group ($p = 0.453$). Therefore, we conclude that there is no linear trend in the relationship between mortality and nitrous oxide pollution in the low rainfall communities.

9. Since Seattle, WA, gets about 37 inches of rainfall per year, it would be considered as a non-low rainfall group. Based on the fitted multiple linear regression model for age-adjusted mortality, including log(NOx), rainfall, and an interaction between these two variables as predictors, the estimated difference in mean age-adjusted mortality between two groups of non-low rainfall cities who differ in log(NOx) by one log pollution potential unit would be 34.490 deaths per 100,000 residents, with the higher pollutant group having higher age-adjusted mortality. We find a statistically significant association between age-adjusted mortality and log(NOx) in the non-low rainfall group ($p < 0.00005$). Therefore, we have a strong evidence for a first order linear trend of higher nitrous oxide pollution associated with higher mortality among the non-low rainfall communities. However, we cannot conclude that reducing NOx levels in Seattle would lead to lower mortality. We would need further studies to study the causal relationship between NOx and Seattle's mortality rates.

Code Appendix

```
### Setting up the packages, options we'll need:
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(rigr)
### ------------------------------------------------------------
### Reading in the data.
smsa <- read_csv("smsa.csv")
### ------------------------------------------------------------
### Q2
### Plot a scatterplot of age-adjusted mortality on the vertical axis and log(NOx)
### on the horizontal axis
smsa %>%
  ggplot(aes(x = log(NOxPot), y = Mortality)) +
  geom_point(cex = 1) +
  xlab("Log(NOx)") +
  ylab("Age-adjusted mortality \n (deaths per 100,000 resident)") +
  theme_bw() +
  geom_smooth(aes(col = "Lowess smoother"), se=F, method = "loess", show.legend = T) +
  geom_smooth(aes(col = "Linear regression"), se=F, method = "lm", show.legend = T) +
  scale_color_manual(name = "",
                     values = c("Linear regression"="grey",
                                "Lowess smoother"="blue")) + NULL
### ------------------------------------------------------------
### Q3
### Here's a linear model for smsa dataset using rigr:regress
mod1 <- regress("mean", Mortality ~ log(NOxPot), data = smsa)
mod1 %>% coef() %>% round(4)
### ------------------------------------------------------------
### Q4
### Plot a scatterplot of age-adjusted mortality on the vertical axis and log(NOx)
### on the horizontal axis, using different colors for low rainfall or not
smsa %>%
  mutate(rainfall = ifelse(Rain < 20, "low rainfall", "non low rainfall")) %>%
  ggplot(aes(x = log(NOxPot), y = Mortality, col = rainfall, pch = rainfall)) +
  geom_point(cex = 1.5) +
  xlab("Log(NOx)") +
  ylab("Age-adjusted mortality \n (deaths per 100,000 resident)") +
  theme_bw()
### ------------------------------------------------------------
### Q5
### Here's a multiple linear regression model for smsa dataset using rigr:regress
### age-adjusted mortality as response, log(NOx) as predictor,
### and rainfall as effect modifier
mod2 <- smsa %>%
  mutate(rainfall = ifelse(Rain < 20, 1, 0)) %>%
  regress("mean", Mortality ~ log(NOxPot)*rainfall, data = .)
mod2 %>% coef() %>% round(4)
### ------------------------------------------------------------
### Q6
### Here's a multiple linear regression model output using rigr:regress
```

```
### age-adjusted mortality as response, log(NOx) as predictor,
### and rainfall as effect modifier
mod2 %>% coef() %>% round(4)
mod2 %>% lincom(c(0,1,0,1))
### ------------------------------------------------------------
```