BIOST 515/518 Homework 4

Latera Tesfaye Olana

17 February, 2023

Responses:

Question 1: Age-adjusted mortality is a method used to compare mortality rates between populations by accounting for differences in age distributions. Essentially, it allows for a fairer comparison of mortality rates over time or across population with different age structures [1].

Question 2: The log transformation of the pollution variable is helpful for changing skewed distribution of this variable to nearly (almost) symmetric. Figure 1 shows, a first order increasing trend for the relationship between age adjusted mortality and nitrous oxide pollution, for smaller nitrous oxide pollution values (approximately below 2.4 pollution potential). However, due to the last four points (marked in red on figure 1) this observed trend changes.
Since our predictor of interest i.e, nitrous oxide pollution is already transformed, no point were above the computed studentized residuals (fitted with linear regression - mortality per 100,000 residents as an outcome and pollution potential as predictor of interest with 99.99% as *as or less or more extreme* threshold). To test for more unusual observations (specially, influential observation), a leave-one-out approach was implemented. This method estimates varying slope and intercept after removing each subject one-by-one. As it is clearly implied on figure 2 and 3, we have some observation highly influencing our estimated parameters. Keep in mind, the magnitude of the presumed influence depends on how these potentially unusual observations are compared to the standard error. For instance from our robust fit the standard error (the same model as a model used for the studentized residuals) was estimated to be 7.983. For instance, as shown on figure 3 removing subject (observation) number 29 results in a decrease in the estimated mean difference in mortality per 100,000 residents for cities varying by 1 log of nitrous oxide pollution level is 15 deaths per 100,000 residents. Compared to the standard error value (as shown in table 1) this seems significant. On the other hand, removing this observation increases the estimated age adjusted mortality for cities with zero log of nitrous oxide pollution level. However this change (approximately 12) compared to the standard error of the intercept (robust regression), which was estimated to be 16. As these two values are close, it safe to say removing subject number 29 will not impact the intercept. Figure 4 and 5 (p-values after removing each observations one by one) also shows whether the influence of removing subjects (observations) is significant or not.

Table 1: A summary of a linear fit of mortality rate per 100,000 residents considering varying nitrous oxide pollution level as a predictor of interest

	Estimate	Naive SE	Robust SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	905.6132	16.6722	16.0668	873.4520	937.7743	56.3656	0.0000
log_NOX	15.0990	6.4187	7.9835	-0.8817	31.0796	1.8913	0.0636

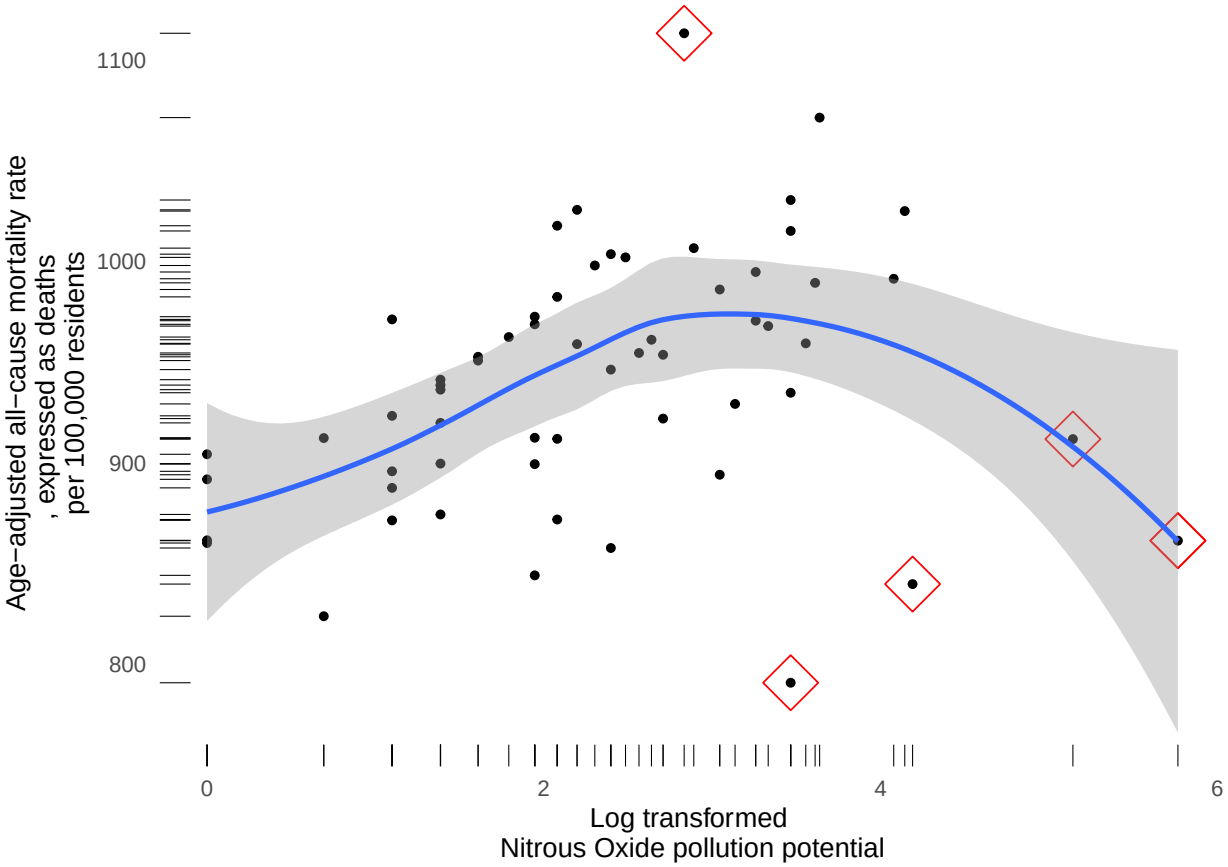


Figure 1: age-adjusted mortality against log(NOx)

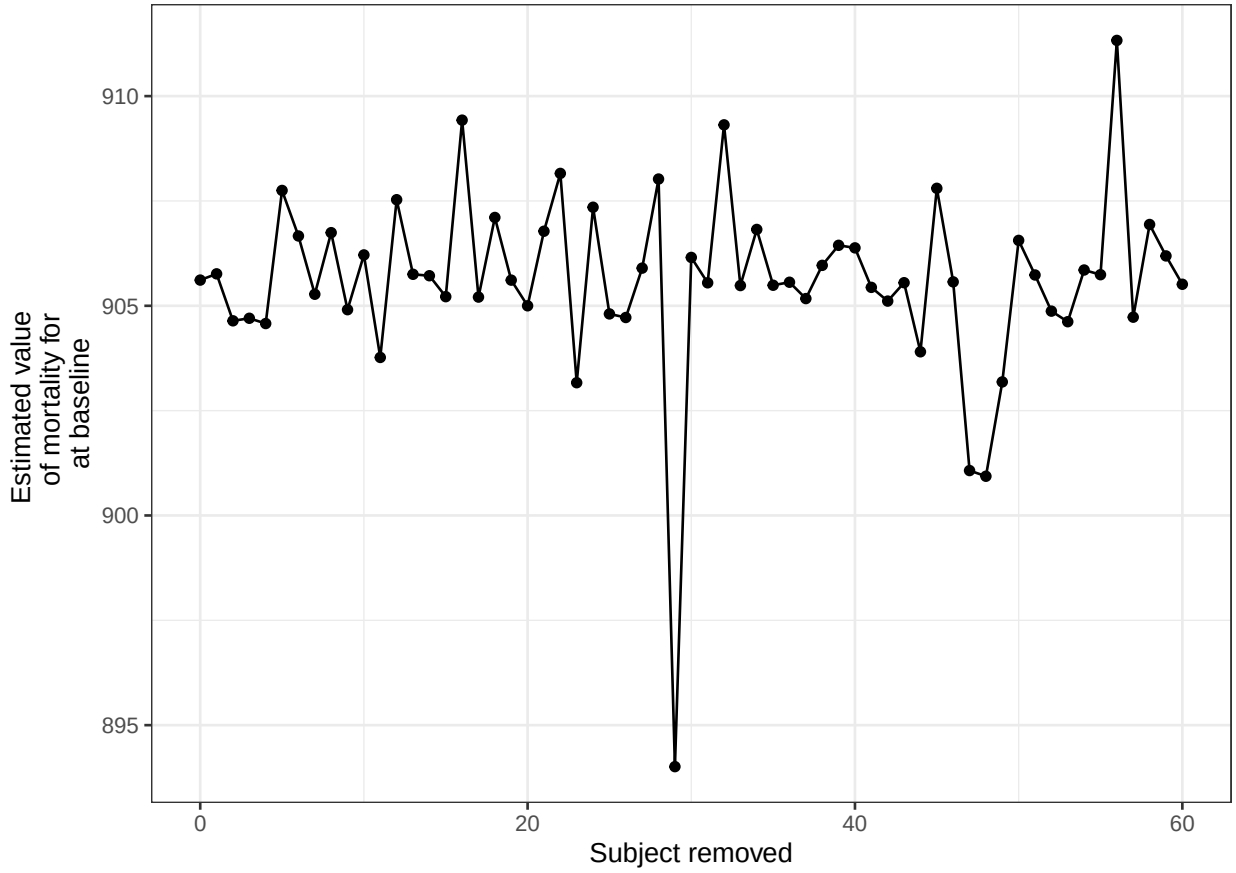


Figure 2: Leave-one-out plot for estimated mean value of mortality (per 100,000 residents) for pollution potentail of zero

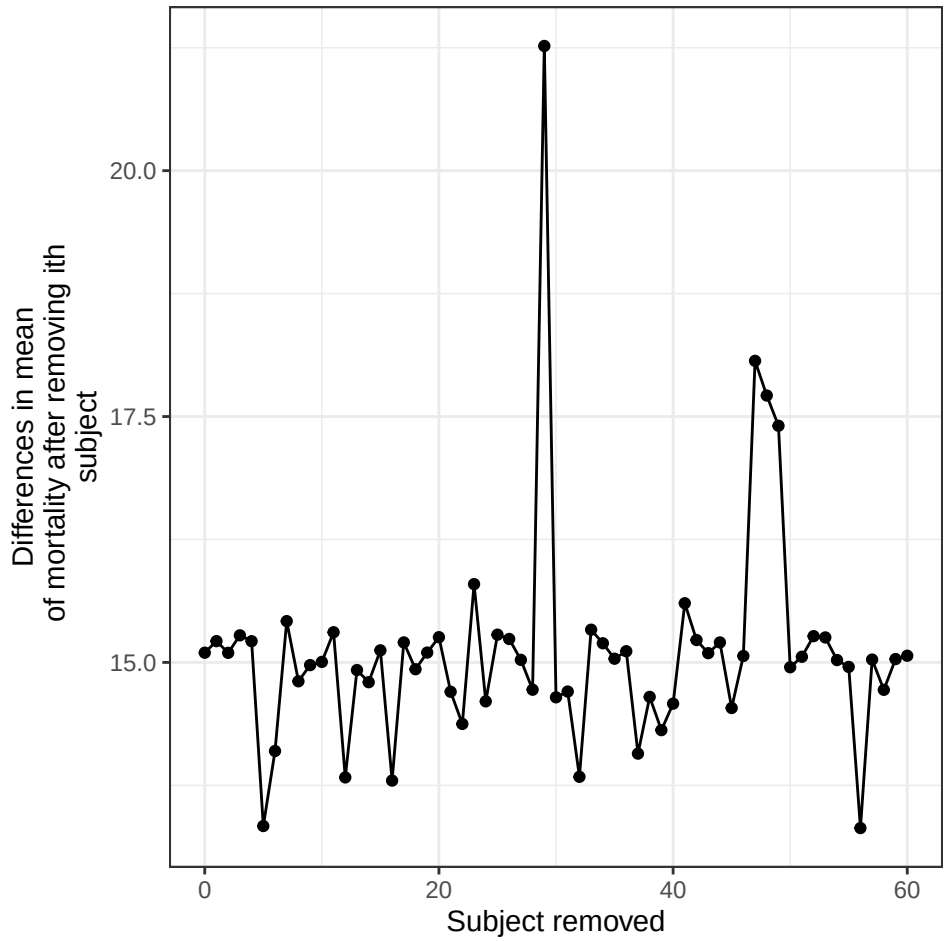


Figure 3: Leave-one-out plot for the estimated mean difference in mortality per 100,000 residents differing by one log of nitrous oxide pollution level

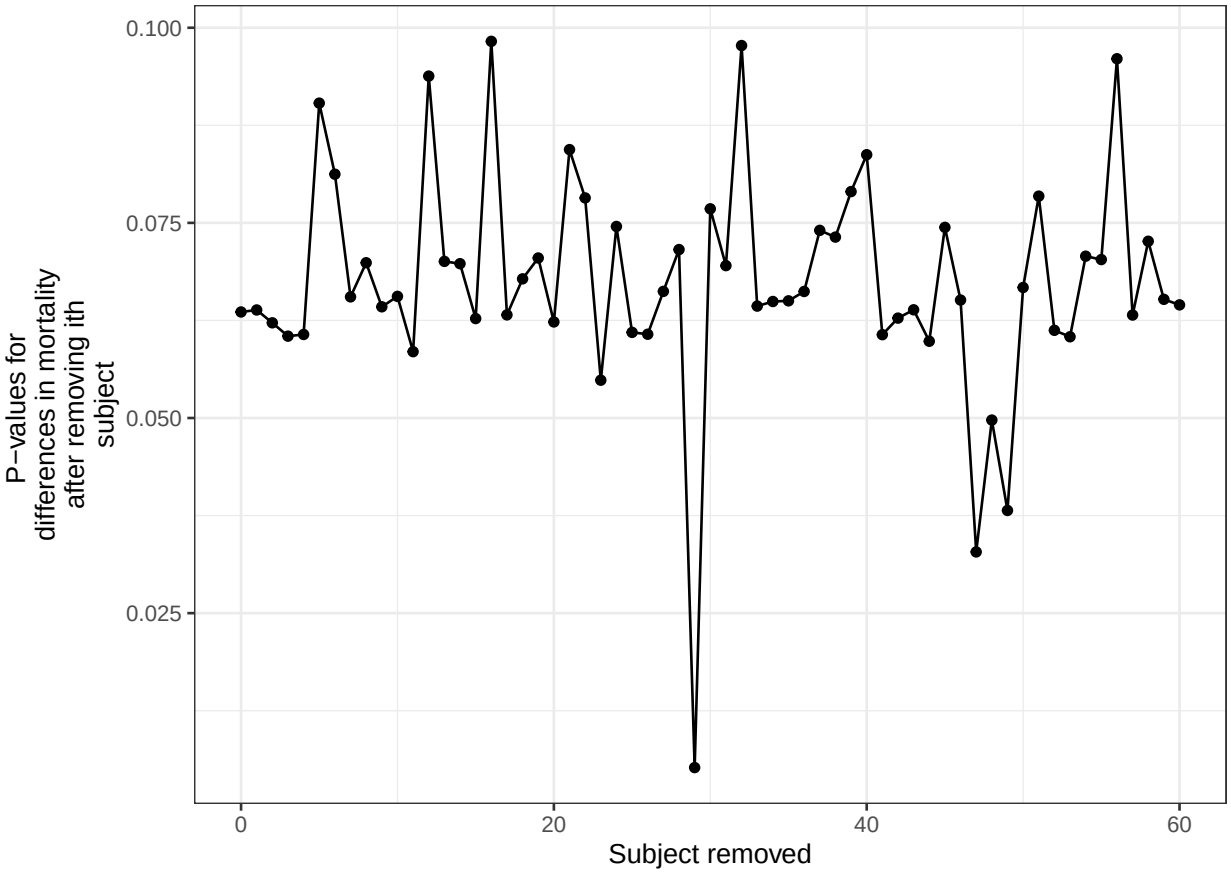


Figure 4: P-values of the predictor for leave-one-out analysis

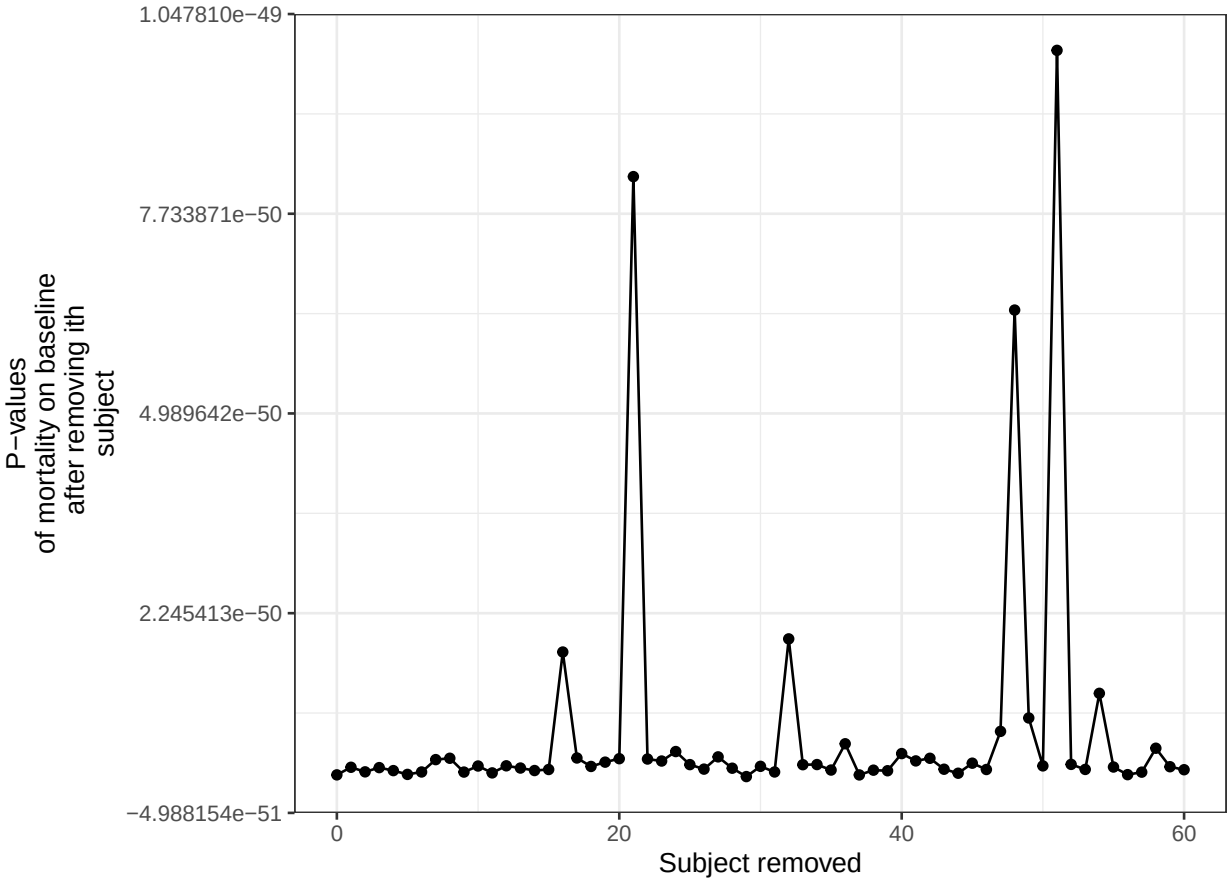
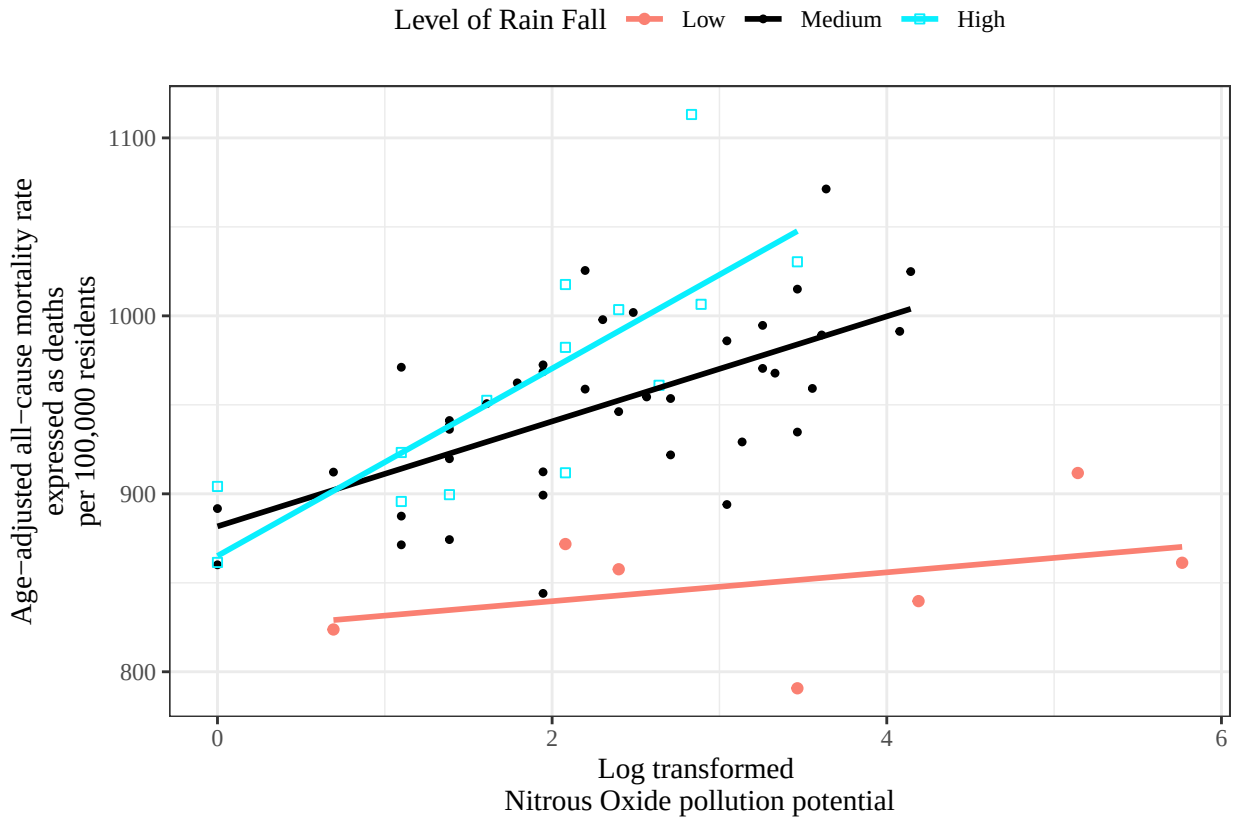


Figure 5: P-value associated with the intercept for leave-one-out analysis

Question 3 New column *Rain_cat* is created per the following rule: Low rain Fall for rain less than 30 inches (<30); medium rain fall for rain between 30 and 45 inches (including 30 but not 45); and finally, high rain fall for rain above 45 inches (including 45).
Rainfall would be a potential effect modifying variable. A critical criterion for evaluating effect modification is to examine whether the relationship between the predictor of interest, namely nitrous oxide pollution levels, and the response or outcome variable, namely age-adjusted mortality rate per 100,000 residents, varies for different values of the effect-modifying variable. As illustrated in Figure 6, the association between pollution potential and age-adjusted mortality varies as a function of changing rainfall categories. Two approaches for assessing whether this association varies involve examining differences in the estimated mean of age-adjusted mortality rates for cities that differ by 1 log unit in their pollution exposure levels across different rainfall categories, and evaluating whether the expected value of age-adjusted mortality for cities with zero log nitrous oxide pollution level (or 1 nitrous oxide pollution) varies across rainfall categories. Figure 6 confirms the validity of these two scenarios. Simply, the verdict for deciding whether a variable is effect modifier or not can start with the assessing if there are any differences in slope and intercept, which are drawn from comparing the outcome and predictor of interest.



Question 4 The fitted model is described below:

$$\hat{mortality}_i|(X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}) = \hat{\beta}_0 + \hat{\beta}_1 \times X_{i1} + \hat{\beta}_2 \times X_{i2} + \hat{\beta}_3 \times X_{i3} + \hat{\beta}_4 \times X_{i4} + \hat{\beta}_5 \times X_{i5}$$

Where $\hat{mortality}_i$ represents the fitted value of mortality rate of the i^{th} city per 100,000 residents. X_{i1} log transformed pollution potential (measured as nitrous oxide pollution level), of the i^{th} city. X_{i2} is 1 if the i^{th} city has medium annual rainfall and 0 otherwise. X_{i3} is 1 if the i^{th} city has high annual rainfall and zero otherwise. X_{i4} represents the interaction term between X_{i1} and X_{i2} ($X_{i1} \times X_{i2}$), whereas, X_{i5} represents the interaction term between X_{i1} and X_{i3} ($X_{i1} \times X_{i3}$). Keep in mind, my reference group is low annual rainfall. Predictor of interest is pollution potential (measured as nitrous oxide pollution level). The outcome is age-adjusted mortality rate for different cities.

Table 2: A summary of a linear fit of mortality rate per 100,000 residents considering varying annual rainfall levels, using nitrous oxide pollution level as a predictor of interest

	Estimate	Naive SE	Robust SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	823.44	34.32	18.16	787.03	859.84	45.35	0.00
log_NOX	8.11	9.09	5.30	-2.51	18.73	1.53	0.13
Rain_catMedium	58.26	37.59	21.99	14.17	102.35	2.65	0.01
Rain_catHigh	41.68	40.94	24.60	-7.64	91.00	1.69	0.10
log_NOX:Rain_catMedium	21.39	10.96	7.35	6.67	36.12	2.91	0.01
log_NOX:Rain_catHigh	44.55	14.03	11.46	21.57	67.53	3.89	0.00

Interpretation of parameters are given as follow: *for the definitions of parameters provided below the following thresholds are used, low rain Fall for rain less than 30 inches (<30); medium rain fall for rain between 30 and 45 inches (including 30 but not 45); and finally, high rain fall for rain above 45 inches (including 45).* In this and subsequent question my interpretation for change in predictor of interest is given as *a change in one unit of log of nitrous oxide pollution.* The reason for doing so is, as stated in the instruction of this homework, the audience for question 1 to 6 is scientific audiences with understanding of statistics.

The estimated mortality per 100,000 residents for high annual rainfall cities and with 0 log of nitrous oxide pollution (nitrous oxide pollution of 1) is 823.44 mortality rate per 100,000.

For two cities with same level of annual rainfall but differ in their log of nitrous oxide pollution level of one, we estimate cities with high nitrous oxide pollution level has mean age adjusted mortality rate (per 100,000 residents) that is 8.11 higher than cities with low nitrous oxide pollution level. Alternatively, this can be also interpreted as, for two cities with same level of annual rainfall but differ in their nitrous oxide pollution by 50%, we estimate cities with high nitrous oxide pollution level has mean mortality rate (per 100,000 residents) that is 3.29 higher than cities with low nitrous oxide pollution level. (I associated 50% increase in predictor of interest - with $\hat{\beta} * \log(1.5)$ change in outcome - mortality).

The estimated difference in mean mortality for cities with the same nitrous oxide pollution level but differ in the annual level of rainfall is 58.26 with cities with medium rainfall having higher mortality as compared to cities with lower annual rainfall. At 5% level we fail to reject the hypothesis that the difference in mean mortality for cities who are same in their nitrous oxide pollution is equal for cities with medium and low annual rainfall (95% CI robust standard errors: (14.17 to 102.35); p = 0.01).

The estimated difference in mean mortality for cities with the same nitrous oxide pollution level but differ in the annual level of rainfall is 41.68 with cities with high rainfall having higher mortality as compared to cities with low annual rainfall. At 5% level we fail to reject the hypothesis that the difference in mean mortality for cities who are same in their nitrous oxide pollution is equal for cities with high and low annual rainfall (95% CI robust standard errors: (-7.64 to 91.00); p = 0.10).

We estimate that when comparing cities who differ in their pollution level by one log of nitrous oxide pollution level, the difference in mean mortality between two cities with medium annual rainfall would be 21.39 per 100,000 residents higher than the difference in mean mortality between two cities with low annual rainfall. (p-value interpretation is included in question 5).

We estimate that when comparing cities who differ in their pollution level by one log of nitrous oxide pollution, the difference in mean mortality between two cities with high annual rainfall would be 44.55 per 100,000 residents higher than the difference in mean mortality between two cities with low annual rainfall. (p-value interpretation is included in question 5).

Question 5 We fit a multiple linear regression model for mortality rate, including pollution potential, an indicator for rainfall level, and an interaction between these two variables as predictors. We estimate that when comparing cities who differ in their pollution level by one log of nitrous oxide pollution level, the

difference in mean mortality between two cities with medium annual rainfall would be 21.39 per 100,000 residents higher than the difference in mean mortality between two cities with low annual rainfall. At the 5% level, we reject the null hypothesis that the difference in mean mortality for cities who differ in log pollution potential of 1 is equal for cities with low annual and high annual rainfall (95% CI for difference in differences based on robust standard errors: (6.67 to 36.12); p = 0.01).

We estimate that when comparing cities who differ in their pollution level by one log of nitrous oxide pollution, the difference in mean mortality between two cities with high annual rainfall would be 44.55 per 100,000 residents higher than the difference in mean mortality between two cities with low annual rainfall. At the 5% level, we reject the null hypothesis that the difference in mean mortality for cities who differ in log pollution potential of 1 is equal for cities with medium and high annual rainfalls (95% CI for difference in differences based on robust standard errors: (21.57 to 67.53); p = 0.00).

Question 6 Based on the fitted model presented, it is not feasible to draw a definitive conclusion regarding the existence of a causal relationship between the level of nitrous oxide pollution and the mortality rate. Linear models, seen earlier, serve the purpose of detecting correlation (association) and not causation. In order to estimate an Average Causal Effect (ACE) and validate the presence of causality between the two variables, it is possible to employ regression analysis, provided that the following critical causality assumptions - consistency, positivity, no interference, and no unmeasured confounding - are demonstrated or assumed to be true. However, verifying these assumptions through statistical methods or through any other technique for that matter would pose a challenge. As a result, we must depend on our collaborators and domain experts to validate our belief in the existence of causality.

Question 7 For our Los Angeles clients, our analysis found that the level of NOx in the air is positively associated with mortality, but the effect is not significant. This suggests that the relationship between NOx and mortality does not vary significantly across cities with low rainfall. However, it is crucial to note that this relationship is complex and may be influenced by other factors. Therefore, it is particularly important to monitor and reduce NOx levels to mitigate potential health risks. The decision in NOx measurement and further mitigation measures should not solely base on the information provided here.

Question 8 For our Seattle clients, our analysis found a positive and statistically significant association between NOx levels and mortality. The effect is modified by the level of rainfall, with a stronger positive association observed in areas with high rainfall. This suggests that Seattle, which receives relatively high rainfall, may need to take extra precautions to reduce NOx levels, particularly during periods of heavy rain. It is important to monitor NOx levels and take measures to reduce them in order to protect public health. However, it is important to note that this relationship is complex and may be influenced by other factors as well.

Reference

1. Richard J. Klein, M.P.H., and Charlotte A. Schoenborn, M.P.H. Age Adjustment Using the 2000 Projected U.S. Population. (2001). https://www.cdc.gov/nchs/data/statnt/statnt20.pdf

Code Appendix

```
### Setting up the packages
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
# check if packages are installed; if not, install them
packages <- c("tidyverse", "readr", "ggExtra", "plotly",
              "ggplot2","ggstatsplot","ggside","rigr","nlme","lmtest",
              "sandwich")
not_installed <- setdiff(packages, rownames(installed.packages()))
if (length(not_installed)) install.packages(not_installed)

# load packages
library("MASS")
library(sandwich)
library(readr)
library(lmtest)
library(nlme)
library(ggstatsplot)
library(ggside)
library(rigr)
library(ggExtra)
library(broom)
library(plotly)
library(ggplot2)
library(tidyverse) # don't load tidyverse package due to conflict with dplyr
### -----------------------------------------------------------
#Loading working directory of the raw data

#Please load your data/directory by changing it with your work directory
#Throughout this code module you will see a tone of places, where
#data is read and written, so please make sure to change them to your
#working directory folder format

working_directory_data <- setwd("C:/Users/latera/Desktop/Bio_ass")

smsa <- read_csv("Data/smsa.csv")
#I used natural logarithm

# I want to know why the regress doesn't work for the following categorization
smsa <- smsa %>%
  mutate(log_NOX = log(NOxPot))

smsa$Rain_cat <-
  cut(smsa$Rain, breaks = c(0,29,44,Inf),
      labels = c("Low", "Medium","High"))

smsa_fit <- lm(Mortality ~ log_NOX, data = smsa) # lm

# studentized residuals
smsa_resid <- smsa %>%
select(c(Mortality, log_NOX)) %>%
na.omit() %>%
```

```
mutate(studentized_resids = rstudent(smsa_fit))
# compute the threshold for an outlier
threshold <- qnorm((1-.9999)/2, lower.tail = FALSE)
threshold
smsa_resid %>%
filter(abs(studentized_resids) > threshold) -> out_value
out_value
leave_one_out_df <- data.frame(subj_removed = 1:nrow(smsa),
intercept = rep(NA, nrow(smsa)),
Log_NoX = rep(NA, nrow(smsa)))
leave_one_out_df_p <- data.frame(subj_removed = 1:nrow(smsa),
intercept = rep(NA, nrow(smsa)),
Log_NoX = rep(NA, nrow(smsa)))
lm_df_rigr <- regress("mean", Mortality ~ log_NOX, data = smsa)
# perform 60 regressions
for (i in 1:nrow(smsa)){
# leave out the i-th subject
# perform linear regression
# extract/stor new coefficient estimates
leave_one_out_df[i,2:3] <- (
regress("mean", Mortality ~ log_NOX, data = smsa[-i,]) %>%
coef())[,1]
}
# add row for the estimates using full data
leave_one_out_df <- rbind(c(0, (lm_df_rigr %>% coef())[,1]),
leave_one_out_df)
head(leave_one_out_df, n=4) %>% signif(3)
knitr::kable(coef(lm_df_rigr) %>% round(4), caption = "A summary of a linear fit of mortality rate per

# perform 60 regressions
for (i in 1:nrow(smsa)){
# leave out the i-th subject
# perform linear regression
# extract/stor new coefficient estimates
leave_one_out_df_p[i,2:3] <- (
regress("mean", Mortality ~ log_NOX, data = smsa[-i,]) %>%
coef())[,7]
}
# add row for the estimates using full data
leave_one_out_df_p <- rbind(c(0, (lm_df_rigr %>% coef())[,7]),
leave_one_out_df_p)
head(leave_one_out_df_p, n=4) %>% signif(3)
outliers_one <- which(smsa$log_NOX > 4.2)
outliers_two <- which(smsa$log_NOX > 4.1 & smsa$Mortality < 900)
outliers_three <- which( smsa$Mortality < 810)
outliers_four <- which( smsa$Mortality > 1100)

#Plotting scatter plot
p <- ggplot(smsa, aes(x=log_NOX, y=Mortality)) +
      geom_point(size=1.2)+
      xlab("Log transformed \nNitrous Oxide pollution potential") + ylab("Age-adjusted all-cause mortali
  #scale_color_manual(name="Smoking Status",breaks=c('nonsmoker', 'smoker'),
    #   values=c('nonsmoker'='#409df4', 'smoker'='#f54066'))+
```

```
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))+
  geom_point(data = smsa[outliers_one,], size = 8, shape = 23, color = 'red')+
  geom_point(data = smsa[outliers_two,], size = 8, shape = 23, color = 'red')+
  geom_point(data = smsa[outliers_three,], size = 8, shape = 23, color = 'red')+
  geom_point(data = smsa[outliers_four,], size = 8, shape = 23, color = 'red')

p+geom_smooth(method = "loess", se = TRUE)


ggplot(data=leave_one_out_df, aes(x=subj_removed, y=intercept  , group=1)) +
  geom_line()+
  geom_point()+
  xlab("Subject removed")+
  ylab("Estimated value \n of mortality for\nat baseline")+
  theme_bw()+
  theme(legend.background = element_rect
        (fill = "transparent"))
ggplot(data=leave_one_out_df, aes(x=subj_removed, y=Log_NoX  , group=1)) +
  geom_line()+
  geom_point()+
  xlab("Subject removed")+
  ylab("Differences in mean\n of mortality after removing ith\nsubject")+
  theme_bw()+
  theme(legend.background = element_rect
        (fill = "transparent"))

ggplot(data=leave_one_out_df_p, aes(x=subj_removed, y=Log_NoX  , group=1)) +
  geom_line()+
  geom_point()+
  xlab("Subject removed")+
  ylab("P-values for \n differences in mortality \nafter removing ith\nsubject")+
  theme_bw()+
  theme(legend.background = element_rect
        (fill = "transparent"))
ggplot(data=leave_one_out_df_p, aes(x=subj_removed, y=intercept  , group=1)) +
  geom_line()+
  geom_point()+
  xlab("Subject removed")+
  ylab("P-values\n of mortality on baseline \n after removing ith\nsubject")+
  theme_bw()+
  theme(legend.background = element_rect
        (fill = "transparent"))
mort_nox_scatter <- ggplot(smsa, aes(x=log_NOX, y=Mortality,color = Rain_cat, shape=Rain_cat)) +
```

```
  labs(
    x = "Log transformed \nNitrous Oxide pollution potential",
    y = "Age-adjusted all-cause mortality rate\n expressed as deaths \nper 100,000 residents"
  ) +
  theme_bw() +
  theme(
    text = element_text(family = "serif"),
    legend.position = "top"
  )
mort_nox_scatter +
  geom_point() +
  scale_color_manual(name = "Level of Rain Fall", values = c("salmon", "black","#09effe"))+
  scale_shape_manual(name="Level of Rain Fall", values = c(19, 20,22))+
  geom_smooth(method = "lm", se = FALSE)

lm_em <- regress("mean",
  Mortality ~ log_NOX + Rain_cat + log_NOX:Rain_cat,
  data = smsa)
knitr::kable(coef(lm_em) %>% round(2), caption = "A summary of a linear fit of mortality rate per 100,0
```