MATH: 245
Adam Loy

(Not-so) Fresh Air: A Study of the Effect of Air Pollutants on Mortality

Nate Isbell and Dean Gladish

5/8/2018

Introduction:

In a modernized world where metropolitan areas attract an increasing number of civilians, it is imperative that we recognize and understand the impact that production and industry may have on our physical well-being in the long-run. Airborne pollution in the form of nitrogen oxides, sulfur dioxides, and various hydrocarbons can impact people regardless of age and cause unnecessary rises in mortality every year. Using data gathered by social scientists from 60 different metropolitan areas, our analysis aims to discover which, if any, of these particulates is associated with the annual mortality rate; if there is an association we aim to determine its extent and generate a multiple linear regression model to address this problem.

Data:

The analyzed dataset incorporated numerous socioeconomic and climatological measurements as well as the values of HC, NO_x, and SO_2 measured in terms of relative pollution potential. These are the only three explanatory variables we will be discussing in this report. Basic summary statistics of the three variables are displayed below (see Table 1).

Table 1: Summary Statistics for Hydrocarbons, Nitrogen Oxides, and Sulfur Dioxides

	Min.	1st Q.	Median	Mean	3rd Q.	Max.	SD
Mortality (per 100,000)	790.7	898.4	943.7	940.4	983.2	1113.1	62.20
HC	1.00	7.00	14.50	37.85	30.25	648.00	91.98
NO_x	1.00	4.00	9.00	22.65	23.75	319.00	46.33
SO_2	1.00	11.00	30.00	53.77	69.00	278.00	63.39

The following scatterplots represent the basic relationship between the explanatory variables of HC, NO_X, and SO_2 and the response variable of Mortality.

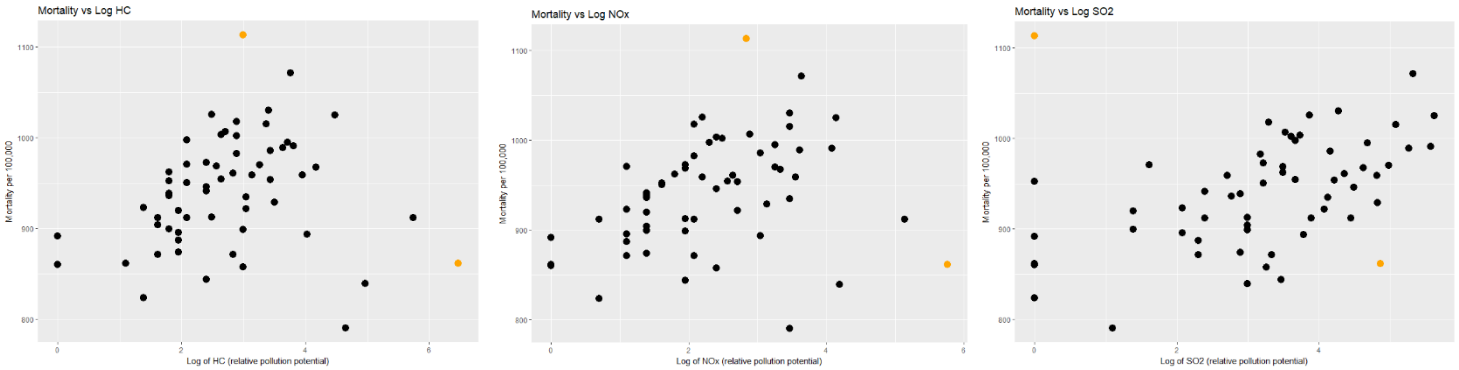


Figure 1: Three scatterplots of Mortality on log-transformed variables with potential outliers highlighted

Through trial and error, we found that log-transformed variables satisfied the conditions for multiple linear regression better than the original plots.

Results:

From the resulting logarithmic transformation of both variables, we obtain the following regression model:

$$\mu[y|\log(x1),\log(x2),\log(x3)] = 871 - 15.7\,\frac{\log(x1)+\log(x2)}{2} + 33.2\log(x3)$$

Where y represents the mortality rate and x1, x2, and x3 represent the particles HC, NO_x, and SO_2 respectively. We used a composite of both log(HC) and log(NOX) to resolve the issue of multicollinearity based off of EDA.

The model given above can be interpreted as follows. A doubling of SO_2 is associated with a 33.2*log(2) increase in Y. A doubling of HC or NO_x is associated with a 15.7*(log(2)/2) decrease in Y. This shows a positive relationship between SO_2 and Mortality rate as well as a negative relationship between our composite variable and Mortality rate. We obtained our values from a t-distribution with 55 degrees of freedom.

Table 2: Model coefficients for multiple regression model

	Estimated Values	Standard Error	t-value	p-value
Intercept	871	15.7	55.3	<2e-16
LogHC_LogNO_x_Average	-15.7	8.44	-1.86	0.0677
Log(SO_2)	33.2	6.32	5.25	2.52e-06

Evaluating these variables and their respective p-values at an α = .05 level reveals that one variable, Log(SO_2), is much more significant than the rest of the variables. This indicates that the issue of multicollinearity did play a part in the initial model.

A 95% confidence interval for the average of log HC and log NO_x is 1.18 to -32.58, and a 95% confidence interval for log SO_2 is 20.6 to 45.8.

Discussion:

Based on the dataset and subsequent analysis, we have overwhelming evidence that SO_2 is definitely positively associated with the mortality rate per 100,000. It is important to note from our Results section that the 95% CI for our composite (mean of log HC and log NO_x) includes zero, which makes it nearly impossible to infer that its impact on mortality is significant. This indicates that the effect of HC and NO_x might not be as large as the initial data suggests.

However, we cannot infer that the concentration of SO_2 has a causal relationship with the mortality rate because the data was gathered from an observation study. The researchers would need to introduce elements of experimentation in order for us to be able to draw a causal conclusion. We also have concerns about the limitations of the study – we omitted two values for Los Angeles, CA and New Orleans, LA because our analysis revealed high leverage for these values. The high studentized residual value for New Orleans also may be a concern. Our

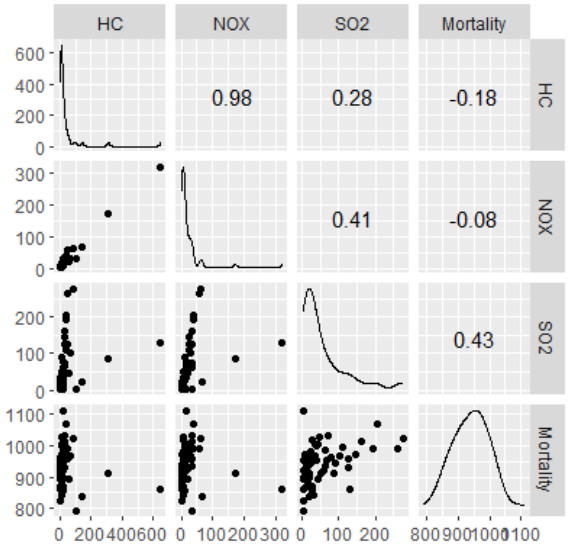
R^2 of 0.383 might indicate that our model is not an accurate fit for the data. After our exploratory analysis we found that the particle SO_2 has a much more substantial effect on mortality (between 20.6 and 45.8) than the other particles whose effect on mortality was less significant and less predictable (between -32.58 and 1.18).

Code Supplement for Case Study 2

Dean Gladish and Nate Isbell

May 8, 2018

```
library(Sleuth3)
library(GGally)
library(broom)
library(car)
library(ggformula)


data(ex1217, package = "Sleuth3")

#  Constructs a scatterplot matrix
ggscatmat(ex1217, columns = c(15:17, 2))
```

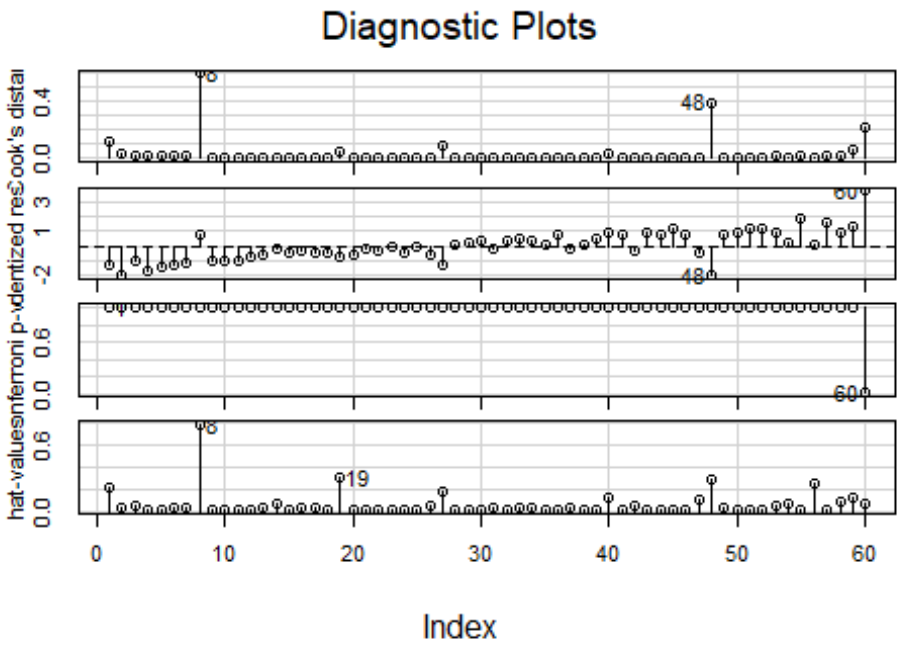



```
#  A basic SLR model.
model1 <- lm(Mortality ~ HC + NOX + SO2, data = ex1217)

#  This puts our information into a more readable format.
tidy(model1)
```

```
##          term    estimate std.error  statistic      p.value
## 1 (Intercept) 924.1631054 8.9720848 103.004276 1.557282e-65
## 2          HC  -1.6132989 0.6068211  -2.658607 1.021017e-02
## 3         NOX   2.9345822 1.2666352   2.316833 2.419511e-02
## 4         SO2   0.2006535 0.1727938   1.161231 2.504740e-01
```

```
#  This portion allows us to see Cook's distance (which governs influence),
#  Studentized residuals (which indicate deviation from the model itself),
```

```
#  the Bonferroni p-value, and the hat-values.
influenceIndexPlot(model1)
```

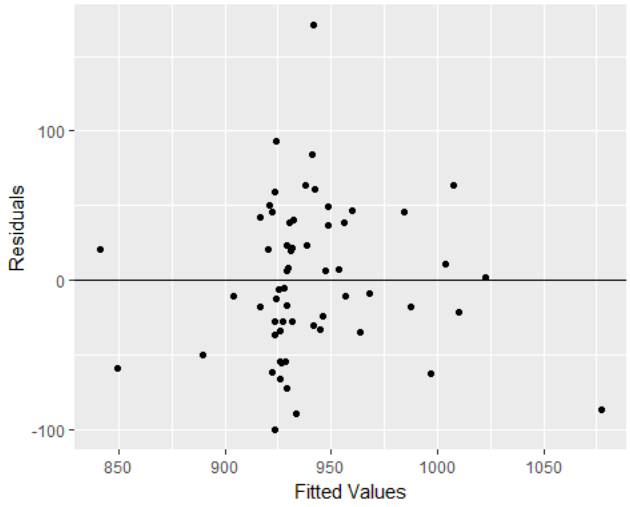


We determined that the cut-off point for leverage here is 2*((3+1)/60) ~ 0.133333. As shown on the plots, case 8 (LA) is influential as it is close to 1 in terms of Cook's Distance. This case (8) also has high leverage as indicated on the hat-values plot. This means that we should exclude case 8 as well as any other outliers that we come across.

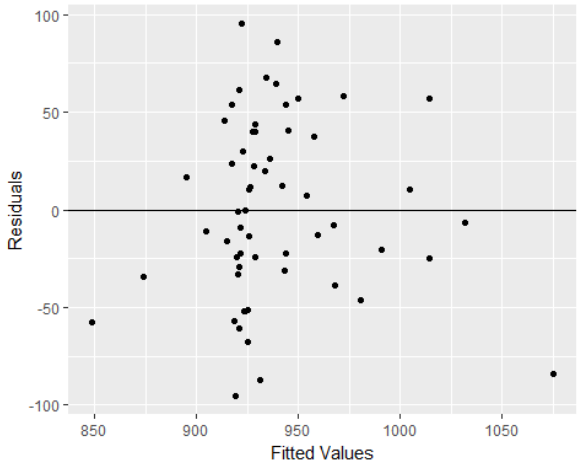
City 60 does not have a high leverage value but it deviates from the model as indicated by its high residual value. Thus we decided to remove cases 8 and 60 (New Orleans).

The following code is where we fit a new model that excludes these two cities.

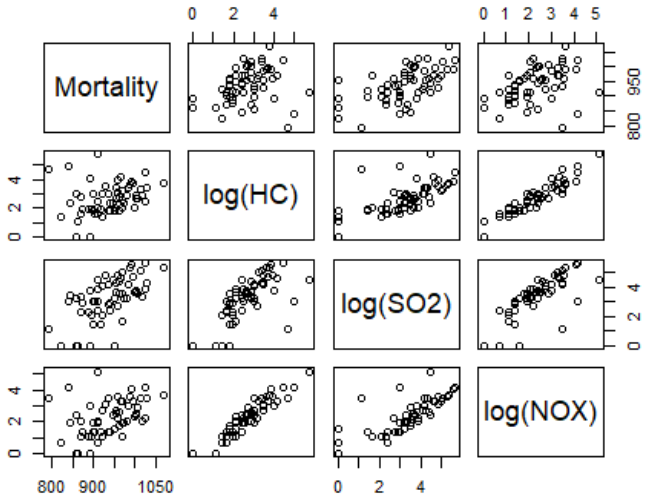
```
#  Fit a new model without case 8.
model2 <- lm(Mortality ~ HC + NOX + SO2, data = ex1217, subset = -c(8, 60))

#  The following code gives us information about the model used to
#  generate residual plots below.
model1_aug <- augment(model1)
model2_aug <- augment(model2)

# Residual plot for first model
gf_point(.resid ~ .fitted, data = model1_aug) %>%
  gf_hline(yintercept = 0, color = "blue", linetype = 2) %>%
  gf_labs(x = "Fitted Values", y = "Residuals")
```




```
# Residual plot after removals.
gf_point(.resid ~ .fitted, data = model2_aug) %>%
  gf_hline(yintercept = 0, color = "blue", linetype = 2) %>%
  gf_labs(x = "Fitted Values", y = "Residuals")
```




```
# To assess potential multicollinearity we created a correlation matrix.
pairs(Mortality ~ log(HC) + log(SO2) + log(NOX), data = ex1217, subset = -
c(8, 60))
```



The residual plots shown above indicate that after removing the two outliers we are left with less variance and a more smooth pattern among the residuals. The correlation matrix that we created essentially indicates that there is a problem with multicollinearity; specifically, log(HC) and log(NOX) appear to be highly correlated, as do log(SO2) and log(NOX). The result of this correlation between explanatory variables needs to be addressed in order for our model in order not to result in seemingly significant results due to inflated p-values.

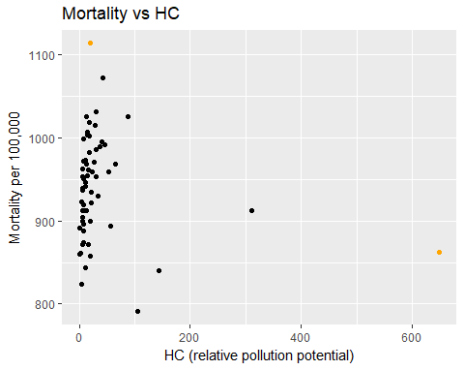
As a result, we elected to add a composite variable. This variable is the average of the logs of two columns, HC and NOX, to the dataset ex1217.

```
library(dplyr)

# This code updates the dataset with a column consisting of the composite
# variable evaluated at each value of HC and NOX.
ex1217 <- mutate(ex1217, logHC_logNOX_Average = (log(HC) + log(NOX))/2)
```

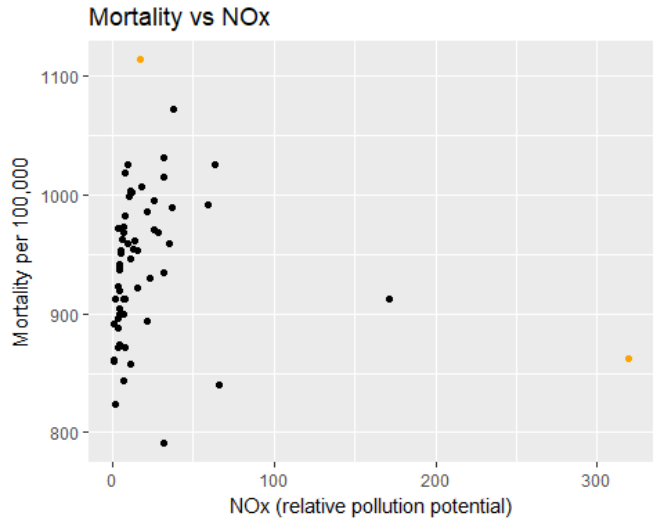
Ultimately we determined that the two variables HC and NOX did not matter very much and that the SO2 value was generally the sole variable of our interest that was associated with the mortality rate. We also looked at scatterplots of the original variables versus the transformed variables, and included our plots of the log-transformed variables in our final report to indicate that the transformation effectively solved the non-linearity (which resembles a plot following the sqrt function) problem of our data.

```
# Scatterplots of original (not transformed) variables with
# outliers highlighted in orange.
gf_point(Mortality ~ (HC), data = ex1217) %>%
  gf_point(Mortality ~ (HC), data = filter(ex1217, CITY == "Los Angeles,
CA"), color = "orange") %>%
  gf_point(Mortality ~ (HC), data = filter(ex1217, CITY == "New Orleans,
LA"), color = "orange") %>%
```

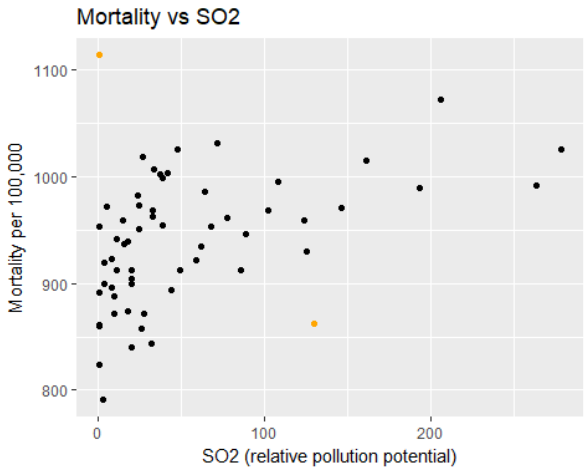
```
  gf_labs(x = "HC (relative pollution potential)", y = "Mortality per
100,000", title = "Mortality vs HC")
```



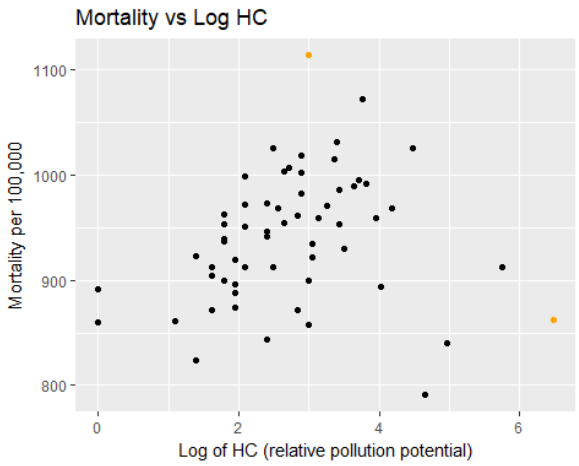

```
gf_point(Mortality ~ (NOX), data = ex1217) %>%
  gf_point(Mortality ~ (NOX), data = filter(ex1217, CITY == "Los Angeles,
CA"), color = "orange") %>%
  gf_point(Mortality ~ (NOX), data = filter(ex1217, CITY == "New Orleans,
LA"), color = "orange") %>%
  gf_labs(x = "NOx (relative pollution potential)", y = "Mortality per
100,000", title = "Mortality vs NOx")
```



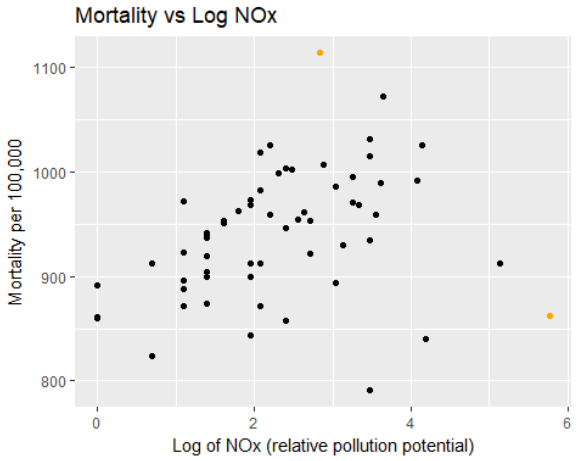

```
gf_point(Mortality ~ (SO2), data = ex1217) %>%
  gf_point(Mortality ~ (SO2), data = filter(ex1217, CITY == "Los Angeles,
CA"), color = "orange") %>%
  gf_point(Mortality ~ (SO2), data = filter(ex1217, CITY == "New Orleans,
LA"), color = "orange") %>%
  gf_labs(x = "SO2 (relative pollution potential)", y = "Mortality per
100,000", title = "Mortality vs SO2")
```



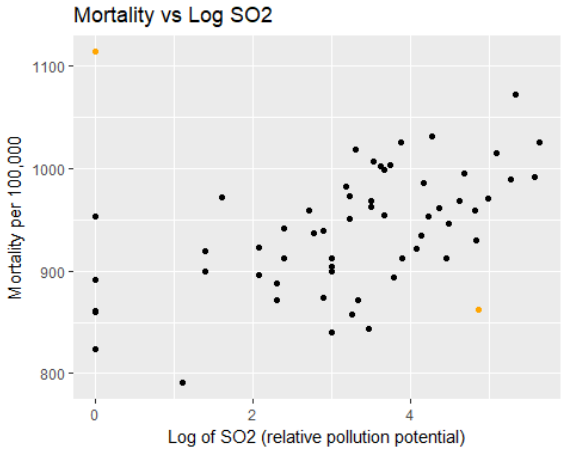

```
# Scatterplots of log-transformed variables with outlier highlighted.
gf_point(Mortality ~ log(HC), data = ex1217) %>%
  gf_point(Mortality ~ log(HC), data = filter(ex1217, CITY == "Los Angeles,
CA"), color = "orange") %>%
  gf_point(Mortality ~ log(HC), data = filter(ex1217, CITY == "New Orleans,
LA"), color = "orange") %>%
  gf_labs(x = "Log of HC (relative pollution potential)", y = "Mortality per
100,000", title = "Mortality vs Log HC")
```




```
gf_point(Mortality ~ log(NOX), data = ex1217) %>%
  gf_point(Mortality ~ log(NOX), data = filter(ex1217, CITY == "Los Angeles,
CA"), color = "orange") %>%
  gf_point(Mortality ~ log(NOX), data = filter(ex1217, CITY == "New Orleans,
LA"), color = "orange") %>%
  gf_labs(x = "Log of NOx (relative pollution potential)", y = "Mortality per
100,000", title = "Mortality vs Log NOx")
```



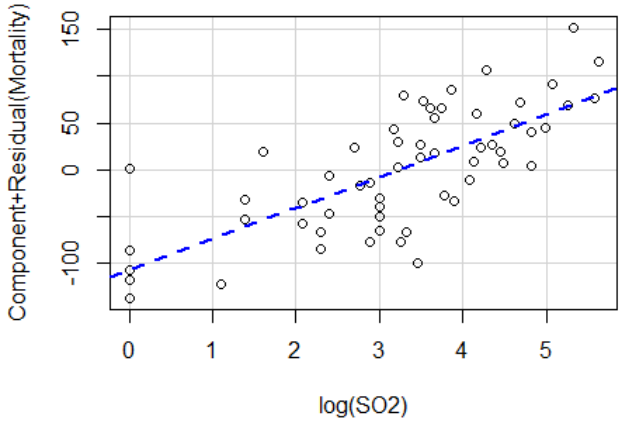

```
gf_point(Mortality ~ log(SO2), data = ex1217) %>%
  gf_point(Mortality ~ log(SO2), data = filter(ex1217, CITY == "Los Angeles,
CA"), color = "orange") %>%
  gf_point(Mortality ~ log(SO2), data = filter(ex1217, CITY == "New Orleans,
LA"), color = "orange") %>%
  gf_labs(x = "Log of SO2 (relative pollution potential)", y = "Mortality per
100,000", title = "Mortality vs Log SO2")
```



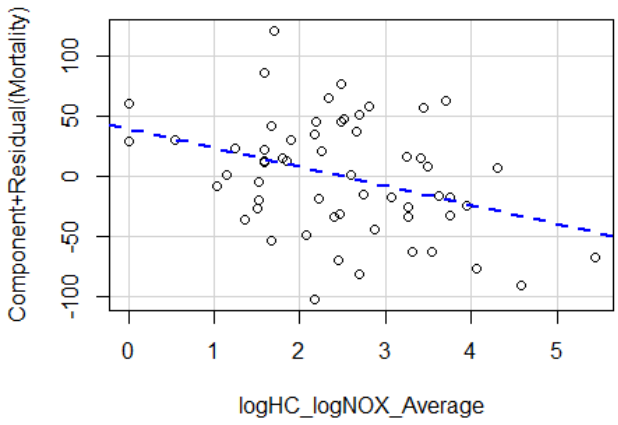
After deciding on our composite variable we fit our final model.

```
# Fit a third model based on our composite:
model3 <- lm(Mortality ~ logHC_logNOX_Average + log(SO2), data = ex1217,
subset = -c(8, 60))

# We computed values for a 95% confidence
# interval for the coefficient
# of our composite variable.
-15.7+2*8.44

## [1] 1.18

-15.7-2*8.44
```

```
## [1] -32.58

# and for a 95% CI of our log SO2 variable.
33.2+2*6.32

## [1] 45.84

33.2-2*6.32

## [1] 20.56

# The Adjusted R-Squared value indicated
# below merits some justification.
summary(model3)

##
## Call:
## lm(formula = Mortality ~ logHC_logNOX_Average + log(SO2), data = ex1217,
##     subset = -c(8, 60))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107.41  -31.58   -1.02   29.25  108.57
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            870.708     15.748  55.289  < 2e-16 ***
## logHC_logNOX_Average  -15.730      8.439  -1.864   0.0677 .
## log(SO2)               33.156      6.315   5.251 2.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.59 on 55 degrees of freedom
## Multiple R-squared:  0.4049, Adjusted R-squared:  0.3833
## F-statistic: 18.71 on 2 and 55 DF,  p-value: 6.316e-07

# These partial residual plots allow us to analyze
# potential non-linearity concerns.
crPlot(model3, variable = "log(SO2)", smooth = FALSE)
```




```
crPlot(model3, variable = "logHC_logNOX_Average", smooth = FALSE)
```



Given that our partial residual plots do not indicate any obvious patterns we can say that the model does fit the log-transformed data despite not accounting for all of the variance as shown by our low adjusted R^2 value.