CHAPTER 1
AIR POLLUTION AND PUBLIC HEALTH:
A CASE STUDY FOR
REGRESSION ANALYSIS
Fig. 1.1: Illustration of simple linear regression
Fundamental Equations

Simple linear regression:

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, ..., n.$$  

Multiple regression:

$$y_i = \sum_{j=1}^{p} x_{ij} \beta_j + \epsilon_i, \quad i = 1, ..., n.$$  

The $\epsilon_i$ are random errors (typically mean 0, common variance $\sigma^2$)
QUESTIONS

• Which variable should be $y$ and which should be $x$?
• Should either or both variables be transformed (for example, by taking logarithms) prior to analysis?
• Is the relationship between $y$ and $x$ linear, or should some nonlinear function be considered?
• Are there omitted variables whose inclusion might substantially change our conclusions about the form of the relationship?
• Are there outliers or influential values among either the $x$ or $y$ variables which may be distorting the interpretation of the data?
• Are the random errors truly independent, of equal variance, and normally distributed?
• Finally and perhaps most problematically of all — having found a relationship between the two variables and tested its statistical significance, can this be taken as evidence of a causal effect?
Background on Air Pollution and Health

It is widely recognized that high air pollution has a significant public health impact. However there are many issues that are still debated in the context of present-day air pollution regulations:

- Which precise pollutants are responsible, e.g. ozone, sulfur dioxide, fine particulate matter, coarse particulate matter, etc.?

- Is there a threshold effect, i.e. a safe level below which air pollution essentially has no adverse consequences?

- Which subsets of the population are most affected, and in what ways (what kinds of deaths or other adverse reactions such as asthma)?

- Exactly how should one quantify the overall effect?
Early Studies — London in the 1950s

• Very high air pollution levels (20–30 times current EPA standards?) and sharp rises in deaths (e.g. the December 1952 smog in London is generally reckoned to have caused 4,000 “excess deaths” in a 4–5 day period)

• Graphs of daily deaths versus air pollution levels give some nice simple examples of linear regression relationships

• Research on “what really happened” continues up to this day
The Great Smog of London Was Anything but Great

By ERIC NAGOURNEY

In early December 1952, a great mass of cold air moved off the English Channel, draped itself over London like an icy comforter and then simply stayed put.

Trying to keep warm, Londoners piled extra coal into their fireplaces, sending plumes of black, sooty smoke into the air that mixed with clouds of exhaust from factories and coal-burning power plants. But instead of rising into the atmosphere and dispersing, the smoke stayed close to the ground, trapped by the cold air above.

Over the next five days, a city already famous for its smog experienced the worst air pollution it had ever seen. A thick haze hovered over the streets, penetrating homes and offices. Public transportation nearly ground to a halt, and at night the visibility was so poor that some parts of London became un navigable. Indoor concerts were canceled because the audiences could not see the stage.

---

Toxicologic and Epidemiologic Clues from the Characterization of the 1952 London Smog Fine Particulate Matter in Archival Autopsy Lung Tissues

Andrew Hunt,† Jarrold L. Abraham,‡ Brett Judson,† and Colin L. Berry‡

†Department of Pathology, State University of New York Upstate Medical University, Syracuse, New York, USA; ‡Department of Morbid Anatomy and Histo pathology, Queen Mary and Westfield College, Royal London Hospital, London, United Kingdom

Environmental Health Perspectives • VOLUME 111 | NUMBER 9 | July 2003
Plots of data show:

- Strong visible association between smoke levels (circles) and deaths (black dots) — suggestion of a one-day lag (Fig. 1.2)
- A scatterplot shows a more direct relationship with high correlation (Fig. 1.3)
- However, more detailed plots from a different smog event show a number of possible complicating factors, e.g. temperatures were also very low on the days when deaths were high. We can also see some distinction among different age groups (Fig. 1.4)
- Individual scatterplots give more information, but just looking at correlations could be misleading for determining the true causal relationship (Fig. 1.5)
- Time series over several London winters show gradually decreasing deaths as smoke levels sharply decreased (Fig. 1.6)
We can also look at scatterplots of the annual aggregated deaths (Fig. 1.7). Note the effect of the outlier due to the very cold 1963 winter.
Fig. 1.2: Smoke levels and deaths during 1952 smog
Fig. 1.3: Deaths (lagged 1 day) vs. Smoke

Corr = .91
Fig. 1.4: Time Series for Dec. 1957 Smog

**Smoke**

- Y-axis: Smoke
- X-axis: Day

**SO2**

- Y-axis: SO2
- X-axis: Day

**Max and Min Temp**

- Y-axis: Temperature
- X-axis: Day

**Deaths by Age Group**

- Y-axis: Deaths
- X-axis: Day
Fig. 1.5: Scatterplots of Lagged Deaths vs. Predictors

- Smoke: Corr = 0.55
- SO2: Corr = 0.54
- Max Temperature: Corr = -0.25
- Min Temperature: Corr = -0.46
Fig. 1.6: Time Series for 14 London Winters
Fig. 1.7: Scatterplots of Annual Deaths vs. Predictors (Dotted lines omit outlier; dashed lines include)
Modern Studies of Air Pollution and Health

Modern studies apply many of the same methods of analysis to much larger data sets.

For example, Kelsall et al.* analyzed a 14-year series of daily mortality and air pollution in Philadelphia. They decomposed the observed daily time series of deaths into three components, (a) seasonal and long-term trend, (b) meteorology, (c) air pollution. They concluded that even though much of the variability can be attributed to (a) and (b), there is a strong enough residual effect under (c) to conclude that air pollution has a very significant effect, even at levels of air pollution much lower than the infamous London smogs.

The present study is based on a re-analysis of the same data and is intended to illuminate many of the issues, including some that remain controversial.

**Figure 1.8.** Time series plot of weekly deaths in Philadelphia, with smoothed lowess curve. The vertical dotted lines are placed to indicate the ends of years.

**Figure 1.9.** Scatterplots of daily deaths against four covariates, with fitted subsample averages (the text will describe the details of how these were calculated). TSP is Total Suspended Particulates (this measure has now been replaced in most studies by more specific measures of particles in different size ranges). SO2 is sulfur dioxide.
Outline of analysis (more details in Chapter 5):

- The $y$ variable was taken to be the square root of daily death count. The square root transformation is motivated in part by the fact that if the death counts have a Poisson distribution, which seems a reasonable intuitive assumption, then taking square roots stabilizes the variances. See Section 5.3 for a more detailed discussion of transformations.

- Seasonal trend was handled by representing the smooth curve in Fig. 1.8 as a linear combination of 180 fixed basis functions. Since we believe that the acute effects of interest occur within at most one week of a high-pollution event, these seasonal covariates should not confound the pollution effect.
• Meteorological effects were handled by introducing linear and quadratic terms for both temperature and dewpoint, with additional terms for the range above 75°F and 60°F respectively. Moreover, to allow for delayed effects of up to four days, the current day’s values were supplemented by those lagged from one to four days. A variable selection was performed to remove insignificant terms which could still confound the pollution effects we are trying to measure. Section 5.2 in Chapter 5 has more detail on variable selection.

• Five pollutants were introduced both singly and in combination. Since we do not know which lags of pollutant are most relevant, we used all possible lags between 0 and 4 days, as well as averages of 2, 3, 4 and 5 consecutive days within this period. Among
the 15 possible “exposure measures” thus generated, the one with the most significant effect was selected and used in the subsequent analysis. Possible “data snooping” criticisms of this.
Conclusions

• The most significant exposure measure for TSP is current day’s value. The $t$ value (estimate divided by standard error) is 3.1, which is significant at the (two-sided) level of .002.

• The most significant exposure measure for SO$_2$ is current day’s value, with $t = 3.3$.

• The most significant exposure measure for NO$_2$ is the 4-day lagged value, with $t = 2.1$.

• The most significant exposure measure for CO is the average of the 3-day and 4-day lags, with $t = 2.8$.

• The most significant exposure measure for O$_3$ is the average of the current day’s value with those for lags 1 and 2. This leads to $t = 2.9$.

All of these are statistically significant at the 5% level, but note the “selection effect” of comparing different measures of exposure and only picking out the most significant.
When different combinations of the variables are included, the results change. For instance, with TSP and SO$_2$ together, the $t$ statistics are 1.5 (TSP) and 1.4 (SO$_2$). Other analyses are:

TSP, O$_3$ together: $t=2.8$ (TSP), 2.9 (O$_3$).

TSP, SO$_2$, O$_3$ together: $t=1.1$, 1.6, 2.9.

TSP, SO$_2$, NO$_2$, CO, O$_3$ all included: $t=1.2$, 1.7, 0.4, 2.1, 2.7.
We can also look for *nonlinear* relationships between airpollution and mortality, which are fitted by more complicated versions of regression analysis. For example, Fig. 1.10 shows some possible piecewise linear relationships between TSP and mortality, suggesting higher slopes at higher TSP levels (relevant to determination of standards).

The conclusions about multiple pollutants and nonlinear relationships suggest that the truth about how air pollution affects mortality may be more complicated than a simple linear relationship based on a single pollutant would suggest.
Other types of study lead to different types of conclusion. For example, Fig. 1.11 is based on a famous prospective study* in which adjusted mortality rate was plotted against median level of fine particulate matter in each of 51 cities. The difficulties here include that measurement of particulate matter is very imprecise if we are only using a single value to represent a long period of study, and there are many other possible reasons why the death rate in these cities might differ (ecological bias problem).

Figure 2. Age-, sex-, and race-adjusted population-based mortality rates for 1980 plotted against mean fine particulate air pollution levels for 1979 to 1983. Data from metropolitan areas that correspond approximately to areas used in prospective cohort analysis.
Summary of Chapter

Air pollution is a major public health issue

The relationship between air pollution (particulate matter, sulfur dioxide, etc.) can be examined through regression relationships

Simple analyses — individual-day analyses of high pollution episodes, analyses based on annual summary statistics from long time series — are interesting exercises but don’t really help understand the full phenomenon

Therefore, in recent years attention has move to the daily analyses of long time series which are much more informative, but also pose many complicated problems of interpretation

There are other types of data sets (e.g. prospective studies) that add to the information (and the confusion)