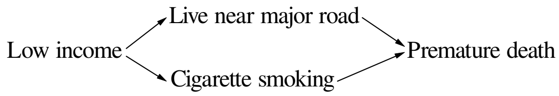
University of Waterloo W. H. Cherry

RELATIONSHIPS IN STATISTICS: Comparative Plans and Causal Structures



Based on the background provided in Statistical Highlights #57 to #63, the causal structure at the right provides a convenient context for discussing cases (1), (8), (9) and (11) (shown again overleaf on page HL64.2) from Statistical Highlight #59; the context comes from an investigation whose newspaper report is reprinted below. This investigation found that death rates were higher for people who lived near a major road – in our terminology, living near a major road (focal variate **X** with two values: living near such a road and not doing so) is associated with a higher death rate [an attribute of response variate **Y** with two values

---

EM0424: The Globe and Mail, August 6, 2004, page A11

Death rate higher near busy roads

BY STEPHEN STRAUSS

Canadian scientists have found a startling rise in death rates associated with nothing more perilous than living within 50 metres of a major highway and 100 metres of a city road that carries a slew of polluting cars and trucks.

While there have been a number of studies tying surges in deaths to city air pollution in general, what the researchers at McMaster University uncovered was a roughly 18-per-cent spike in mortality in the Hamilton area among people who lived adjacent to streets carrying 35,000 to 75,000 vehicles daily.

The rise in the pollution death rate did not come from asthma, emphysema or lung cancer but from heart attacks and other heart conditions. "Basically air pollution does not affect your lungs but your heart", is how Murray Finkelstein of McMaster's program in occupational health and environmental medicine, and a co-author of the new study, describes what his group has found.

The reason for the large heart-disease hit is still uncertain, but Dr. Finkelstein points to research in animals that suggests air pollution particles can irritate arteries and lead to their general hardening and thickening.

Although the study, which was published in the July issue of the *American Journal of Epidemiology*, focused on the roads and highways of Hamilton, the researchers see no reason why the findings shouldn't apply to city dwellers perched above traffic surging along St. Lawrence Street in Montreal, Yonge Street in Toronto, or Hastings Street in Vancouver. Not to mention anyone whose dwelling is on the skirt of the traffic behemoth known as the Trans-Canada Highway, which, as it moves 400,000 people a day in some locations, is North America's second-busiest highway.

What the researchers also did is translate the increasing death rates, which have a relatively small impact on younger people, into something closer to an insurance company's life-expectancy table. They found there is a 2.5-year increase in age-related death levels for people whose dwellings are located cheek-to-jowl with heavy traffic.

But there is also a class-related confounding factor to the data. ..... more poor people may be more likely to live near busy streets than rich people, and poor people have other behaviours – smoking in particular – that might kill them in larger numbers.

"Basically, that means your mortality pattern if you are 50 years old is the same as someone 52.5 years old who doesn't live on a busy road", said Dr. Finkelstein. What is even more sobering is the fact that the deadliness of living near major thoroughfares is not far off the life-shortening effects of such known killers as diabetes or chronic lung disease.

The McMaster scientists say their research leads to a very simple bit of advice for a health-conscious individual. "If you have a heart condition, I would advise not buying a place very close to major roadways or highways", said Michael Jerrett, a McMaster University geography professor who is another co-author on the study.

He also suggests that susceptible people who live close to the busy thoroughfares consider air purification systems in their homes as a preventative act.

There some some caveats to the new study, which replicates Dutch research published two years ago. There was no direct measure of how much higher the motor-vehicle related pollution was near major roads. This omission should be remedied next month when the McMaster group tracks road-pollution level themselves.

But there is also a class-related confounding factor to the data. Because of existing concerns over noise and pollution, Dr. Finkelstein says more poor people may be more likely to live near busy streets than rich people, and poor people have other behaviours – smoking in particular – that might kill them in larger numbers.

REFERENCE: Finkelstein, M.M., Jerrett, M. and M.R. Sears: Traffic Air Pollution and Mortality Rate Advancement Periods. *American Journal of Epidemiolgy* **160**(#2): 173-177, July 15 (2004). [UW Library E-journal]

The abstract given in the original article is:

Chronic exposure to air pollution is associated with increased mortality rates. The impact of air pollution relative to other causes of death in a population is of public health importance and has not been well established. In this study, the rate advancement periods associated with traffic pollution exposures were estimated. Study subjects underwent pulmonary function testing at a clinic in Hamilton, Ontario, Canada, between 1985 and 1999. Cox regression was used to model mortality from all natural causes during 1992-2001 in relation to lung function, body mass index, a diagnosis of chronic pulmonary disease, chronic ischemic heart disease or diabetes mellitus, household income, and residence within 50 m of a major urban road or within 100 m of a highway. Subjects living close to a major road had an increased risk of mortality (relative risk = 1.18, 95% confidence interval: 1.02, 1.38). The mortality rate advancement period associated with residence near a major road was 2.5 years (95% confidence interval: 0.2, 4.8). By comparison, the rate advancement periods attributable to chronic pulmonary disease, chronic ischemic heart disease, and diabetes were 3.4 years, 3.1 years, and 4.4 years, respectively.

(*continued overleaf*)

(alive or dead), quantified in this investigation as a 'mortality rate advancement period' (MRAP)].

(11) $$\begin{tikzcd} & \mathbf{X} \arrow[dr] & \\ \mathbf{Z}_j \arrow[ur] \arrow[dr] & & \mathbf{Y} \\ & \mathbf{Z}_i \arrow[ur] & \end{tikzcd}$$

Case (11): The causal structure at the upper right overleaf on page HL64.1 is an instance of case (11) [shown again at the right], although the investigation (described overleaf) of the effect of living near a major road was not explicitly concerned with $\mathbf{Z}_j$ (low income). However, this causal structure does not raise Questions that are not also raised by the three other cases (1), (8) and (9) discussed below, because it is a composite of them (and other) cases as follows:

- the left-hand side has $\mathbf{Z}_j$ as the *common cause* of $\mathbf{X}$ and $\mathbf{Z}_i$ which is our case case (9) (so-called **common response**),
- the right-hand side has $\mathbf{X}$ and $\mathbf{Z}_i$ as causes of (the *common response*) $\mathbf{Y}$ which is our case (8),
- the top and bottom are the causal chains of cases (4) and (6) [which is which depends on variate assignment];

there are also four instances of case (1): $\mathbf{Z}_j \rightarrow \mathbf{X}$, $\mathbf{Z}_j \rightarrow \mathbf{Z}_i$, $\mathbf{X} \rightarrow \mathbf{Y}$, $\mathbf{Z}_i \rightarrow \mathbf{Y}$.

Cases (1), (8) and (9): These cases involve five Questions that could be investigated; they are numbered 1 to 5 for convenient reference and given with other information in Table HL64.1 below – 'E' or 'O' in the Plan column denotes 'experimental' or 'observational'.

Table HL64.1

No.	Case	Question	Plan
1	(1) $\mathbf{X} \rightarrow \mathbf{Y}$	Is low income associated with premature death?	O
2	(8) = (1) $\mathbf{X} \rightarrow \mathbf{Y}$	Is living near a major road associated with premature death?	O
3	(8) = (1) $\mathbf{X} \rightarrow \mathbf{Y}$	Is cigarette smoking a cause of premature death?	O
4	(9) $\mathbf{Z} \begin{smallmatrix} \nearrow \mathbf{X} \\ \searrow \mathbf{Y} \end{smallmatrix}$	Are living near a major road and cigarette smoking associated with low income?	O
5	(8) $\begin{smallmatrix} \mathbf{X} \searrow \\ \mathbf{Z} \nearrow \end{smallmatrix} \mathbf{Y}$	To what extent are living near a major road and cigarette smoking associated with premature death?	O

Question 1: It has long been known that the answer to this Question is *Yes* – for example, nineteenth century vital statistics in the U.K. showed an association between 'social class' and death rates. The inability of the investigator(s) to assign the value of the focal variate $\mathbf{X}$ (a person's income) is why the Plan can only be observational and the Question is phrased in terms of association (rather than causation).

Question 2: This Question is answered by the investigation whose newspaper report is given overleaf on page HL64.1. As the report points out, possible confounding by a lurking variate $\mathbf{Z}$ (cigarette smoking) means that comparison error imposes a severe limitation on the Answer.

Question 3: The health consequences of cigarette smoking are now well documented as a result of tens of thousands of investigations, most of them from around 1950 and later. The Plans of investigations involving humans have been observational because investigators cannot ethically (or practically) assign elements' smoking habits; *experimental* Plans have been limited to investigations involving animals, but they are relatively few in number, in part because of the difficulty (and, hence, the cost) of getting animals to smoke. [Another factor is the limited lifespans of cheaper laboratory animals (like mice and rats) in relation to the time for some health effects of smoking to become apparent.]

The Question wording involves *causation* because of the requirement that manipulation of the focal variate (reducing the prevalence of cigarette smoking) will produce a desired change in the response variate (a reduction in smoking-induced disease, resulting in better public health and reduced healthcare costs). The decades of research and the number of investigations of the health consequences of smoking are a reminder of the difficulties of establishing causation using an observational Plan.

Question 4: This Question involves $\mathbf{Z}$ (low income) as a *common cause* of $\mathbf{X}$ (living near a major road) and $\mathbf{Y}$ (cigarette smoking), although the Question wording involves (the weaker) association rather than causation – more appropriate notation would be $\mathbf{X}$ instead of $\mathbf{Z}$ and $\mathbf{Y}_1$ and $\mathbf{Y}_2$ instead of $\mathbf{X}$ and $\mathbf{Y}$.

Question 5: This Question involves the effects of *two* focal variates on a response variate, which is case (8) but with $\mathbf{X}_1$ and $\mathbf{X}_2$ in place of $\mathbf{X}$ and $\mathbf{Z}$ – see also the discussion in Appendix 1 on pages HL64.3 and HL64.4 of *quantifying* the relationship of two (or more) focal variates and a response variate.

In summary, four causal structures are introduced in the discussion of statistical association of explanatory variates at the upper right of page HL59.1 in Statistical Highlight #59; these become twelve structures at the middle right of page HL59.1 with the inclusion of the response variate. The discussion on page HL59.1 below these structures, and in this Highlight #64, shows that only *four* of these twelve are relevant to comparative Plans, for which the *primary* concern is the structure of case (1); the overlapping structures of cases (8), (9) and (11) serve mainly to inform case (1) investigating.

- The discussion in this Statistical Highlight #64 reminds us that, to develop a comparative Plan to answer a Question with a causative aspect, sufficient *extra*-statistical knowledge is needed to:
  - frame a (clear) Question about the association being investigated;
  - give a plausable causal structure that is appropriate for this Question;
  - choose (and then develop) a feasible Plan type.

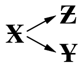
NOTES: 1. The *observational* nature of the Plans in Table HL64.1 above reflects their *context*; in contexts where the value of the focal variate(s) *could* be assigned by the investigator(s), the Plans could be experimental.

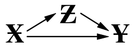
- It is also context-dependent whether the Questions involve *establishing* causation, *quantifying* (causal) relationships or *prioritizing* causes – see Appendix 1 on pages HL64.3 and HL64.4.

(*continued*)

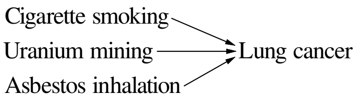
RELATIONSHIPS IN STATISTICS: Comparative Plans and Causal Structures (continued 1)

NOTES: 2. Case (10) [shown again at the right] in its lower *real* form (because Z is an explanatory variate) is *not* a viable basis for a comparative Plan, which requires either:

(10) 



- *direct* causation of Y by X [case (1)], OR:
- an *explicit* intermediate variate in *each* causal chain from X to Y, as in case (11) and illustrated at the start of this Highlight #64 in the causal structure at the upper right of page HL64.1.
  - There may be *more* than two causal chains from X to Y, as illustrated by a case of *three* possible intermediaries at the right. There may also be *interaction* (see Appendix 2 on page HL64.4) among such explanatory variates; for example, the increased risk of lung cancer among uranium miners (presumably due to radioactive dust inhalation) might be mainly among smokers but *both* non-smokers and smokers may be at an increased risk of lung cancer from asbestos inhalation.
    - + Causes of 'lung cancer' are actually more complicated than implied by this example, because there are a *number* of such cancers involving different cell types (*e.g.*, mesothelioma from asbestos inhalation).



3. The newspaper article (reprinted on page HL64.1) discussed in this Statistical Highlight #64 is concerned with higher death rates among people who live near a major road (the focal variate in an investigation with an *observational* Plan). The abstract (at the bottom of page HL64.1 below the newspaper article) of the *journal* article mentions that six possible confounders – lung function, body mass index, household income, a diagnosis of chronic pulmonary disease, chronic ischemic heart disease and diabetes – were considered in the investigation as other possible factors in premature death, and larger mortality rate advancement periods (MRAPs) than for the focal variate were found for the last three of these variates using a model called Cox regression [analogous to the response model (HL64.1) on page HL64.4 but differing in mathematical form]. Such modelling manages comparison error by trying to achieve the statistical benefit of *blocking* in an experimental Plan by *mathematically* (rather than physically) holding some (here, six) lurking variates 'fixed' as X changes. Such modelling encounters two difficulties.
   - Like blocking, it manages comparison error *only* for confounder(s) which are identified *explicitly*:
     – for use as blocking factor(s) or inclusion in the model AND: – whose values it is feasible to measure.
     [This is true also for the confounder(s) used for matching and subdividing in an observational Plan.]
   - It must be assumed that the model has the *correct* (or an *adequate*) *form* for each possible confounder in its structural component – for example, a first power, a second power, a square root, a logarithm; any Z for which this is not so will *not* be held 'fixed' by the model calculations and so can become a source of model error.
     - Holding the *same* (or *similar*) values *physically* (in an experimental Plan) for other explanatory variates as the focal variate changes manages comparison error more effectively than holding them fixed by means of the *model* (and using data from an observational Plan). Likewise, Answers which claim causation based on *physical* evidence have less limitation than those based on a *model*.

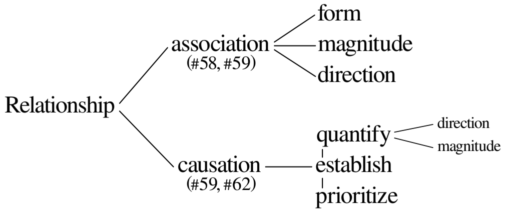
   The similarity of intent between blocking in an experimental Plan and including possible confounders in the structural component of a response model for an observational Plan continues on by managing, *under repetition*, comparison error due to unblocked, unmeasured and unknown confounders:
   - physically, by probability assigning (*e.g.*, EPA) in an experimental Plan – the greater the degree of *replicating*, the greater the reduction in comparing imprecision due to such confounders;
   - mathematically, by the residuals [and their (sub)model] in the response model for an observational Plan (but at the cost of model error becoming one of the components of overall error).

   It is *im*material in the *model* for the investigation described on page HL64.1 whether we regard the seven explanatory variates as seven focal variates or as one focal variate and six possible confounders.

The following Appendices (excerpted from other Statistical Highlights) provide easier access to some cross-referenced material.

Appendix 1: Investigating Statistical Relationships – Three Types of Causal Questions (see Statistical Highlight #63)

Relationships investigated in statistics, which *we* describe in terms of variates, are often encountered as *associations*; investigating associations includes identifying their characteristics and/or the reasons (causal or otherwise) for them (see also Statistical Highlight #60). This Appendix 1 is concerned with comparative Plans for investigating relationships where causation is to be established or *is* involved; the focus on the X-Y relationship being *causal* means that a *change* can (potentially) be induced in Y by *changing* X. These matters are summarized in the schema at the right, which reminds us that:



⁎ association is usually characterized by its *form*, *magnitude* or *direction*;
  – correlation (see Statistical Highlight #66) is one measure of magnitude ('strength') for a straight-line association; form can also be *non*-linear;

⁎ it is useful to distinguish three types of Questions with a causative aspect:
  – **Establishing** whether X *is* a cause of Y, usually with a view to manipulating X to produce

(*continued overleaf*)

- a (desired) change in **Y** – the quintessential example is whether cigarette smoking is a cause of lung cancer (and other life-threatening diseases), the topic of tens of thousands of data-based investigations over several decades starting in the 1940s. Establishing that an observed association of **X** and **Y** is causation of **Y** by **X** is answering the Question whether the relevant causal structure (shown again at the right from page 5.34) is case (1) or case (9) [= case (12)].

(1) $X \rightarrow Y$

(9) $$\begin{tikzcd} & X \\ Z \arrow[ur] \arrow[dr] & \\ & Y \end{tikzcd}$$

- **Quantifying** the relationship between **X** (or, more commonly, X_1, X_2, ..., X_q) and **Y**; this arises in the statistical area of *Design of Experiments* (DOE) – for example, the effect of temperature, humidity, light, fertilizer and insecticide levels on the growth of seedlings in a greenhouse. Quantifying a causal relationship is, in essence, investigating the case (1) causal structure – the subscripts on the case number now remind us that the Plan needs to reflect the number of focal variates involved.

$(1)_1$ $X \rightarrow Y$

$(1)_2$ $$\begin{tikzcd} X_1 \arrow[dr] & \\ & Y \\ X_2 \arrow[ur] & \end{tikzcd}$$

- **Prioritizing** causes by the size of their effect is the domain of (data-based) process improvement – trying to identify the *most important* cause (usually of excessive variation in the process output, **Y**) from among many causes X_1, X_2, ..., X_q.

$(1)_q$ $$\begin{tikzcd} X_1 \arrow[dr] \arrow[dd, dotted, no head] & \\ \arrow[r] & Y \\ X_q \arrow[ur] & \end{tikzcd}$$

Questions which involve *establishing* and *quantifying* causal relationships are typically part of the *same* investigation. For example, in the Physicians' Health Study (described in Figure 10.2 of the STAT 231 Course Materials) of the effect of aspirin on heart disease, *two* Questions, in the context of an appropriate target population, are:

- does aspirin reduce heart-attack risk?
- is the reduction in heart-attack risk due to aspirin large enough to be practically important?

The Physicians' Health Study had to answer *both* Questions; in *in*formal discussion, it is easy to consider only *one* of the Questions and overlook the other.

Similarly, when *prioritizing* causes in process improvement investigations, investigators should:

- verify that the suspected (most important) cause *is* a cause of the (variation in the) response variate(s);
- validate that the proposed Answer *does* address the Question – that the proposed 'solution' *does* solve the 'problem'.

Appendix 2: Comparative Plans – The Protocol for Setting Levels and Interaction (see Statistical Highlight #68)

The protocol for setting levels specifies the *values* to be taken by relevant explanatory variate(s); the simplest case is *two* values of *one* focal variate but there is terminology to deal with the complications of more than two values of more than one focal variate. This terminology is used mainly in the context of *experimental* Plans.

∗ A **factor** is an explanatory variate; we distinguish an explanatory variate that is:
  - a *focal* variate; – a *non*-focal variate used as a **blocking factor**;
  - a *non*-focal variate whose value is managed for other reasons – see Note 1 on page HL65.2 in Statistical Highlight #65.

  Our concern here is with factor(s) that are *focal* variate(s).

∗ Factor **levels** are the set of value(s) assigned to a factor – that is, (usually) the set of values assigned to the (or a) focal variate. Choosing the *values* for levels in the context of a particular investigation may require extra-statistical knowledge.

∗ A **treatment** is a *combination* of the levels of the factor(s) applied to a unit [in the sample (or the blocks)].

∗ A **run** is part of the Execution stage of an experimental Plan in which all the data are collected for *one* treatment.

∗ A **factorial** treatment structure involves *all* combinations of the levels of the (two or more) factors.

∗ The **(treatment) effect** of **X** on **Y** (usually) refers to the change in the *average* of **Y** for *unit* change in **X** and:
  = implies the **X**-**Y** relationship is (believed to be) *causal* – a change in **X** *causes* (brings about) a change in **Y**.

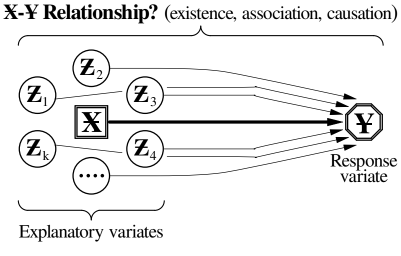
∗ **Interaction** of two factors X_1 and X_2 is said to occur when the effect of one factor on a response variate **Y** depends on the level of the other factor. Interaction means the combined effect of two factors is *not* the sum of their individual effects.

Appendix 3: Investigating Statistical Relationships: Changing and Comparing (see Statistical Highlight #57)

Relationships occur in most (perhaps all) areas of human endeavour and come in many forms. In statistics, we cast relationships in terms of *variates* – in the simplest case, between one **explanatory** variate (**X**, say, which we call the **focal** variate) and one **response** variate **Y**, over the elements of a population. However, as portrayed pictorially at the right, in statistics we can seldom ignore *other* (**non**-focal) explanatory variates (denoted Z_1, Z_2, ..., Z_k) when answering a Question about an **X**-**Y** relationship, because the Answer is predicated on Z_1, Z_2, ..., Z_k remaining *fixed* when **X** *changes* to make apparent its relationship to **Y**. This idea arises mathematically when, to analyze data for the k+2 variates of each unit in a sample of n units, we use the response model (HL64.1) in

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 z_{1j} + ... + \beta_{k+1} z_{kj} + R_j, \quad j = 1, 2, ..., n, \quad R_j \sim N(0, \sigma), \text{ pr. indep., EPS} \qquad \text{-----(HL64.1)}$$

which **Y** has a first-power (or 'straight-line') relationship to each explanatory variate; the interpretation of β_1 (the coefficient of the *focal* variate in the model) is the change in the average of **Y** for unit change in **X** while Z_1, Z_2, ..., Z_k *all remain fixed in value.* [The interpretation of *any* of the k+2 coefficents in the structural component of (HL64.1) requires a similar caveat, of course.]


X-Y Relationship? (existence, association, causation)
Z2
Z1
Z3
X
Y
Zk
Z4
...
Response variate
Explanatory variates