From DAGs to linear path models

Ian Lundberg

Soc 212C

Learning goals for today

At the end of class, you will be able to:

1. Understand how linear path models are like DAGs, but with additional statistical assumptions
2. Determine the covariance between standardized variables using a linear path model

Despite their historical importance in quantitative social science, in many applications linear path models are not as useful as DAGs because the linear parametric assumptions may not hold.

We have been working with DAGs

$$\begin{tikzcd}
A \arrow[r] & Y \\
L \arrow[u] \arrow[ur] &
\end{tikzcd}$$

- Each node is a variable
- Each edge is a direct causal effect
- The DAG is nonparametric
  - Effect of A may be nonlinear
  - Effect of A may be heterogeneous across L

$$\mathrm{E}(Y \mid A, L) = f(A, L) \qquad \leftarrow \text{arbitrarily complex } f()$$

Today we will use linear path models

Today we will use linear path models

$$
\begin{tikzcd}
A \arrow[r, "b"] & Y \\
L \arrow[u, "a"] \arrow[ur, "c"'] &
\end{tikzcd}
$$

Today we will use linear path models

$$\begin{tikzcd}
A \arrow[r, "b"] & Y \\
L \arrow[u, "a"] \arrow[ur, "c"'] &
\end{tikzcd}$$

Adds a **parametric assumption:**
Each output is a linear, additive function of its inputs

Today we will use linear path models

$$\begin{tikzcd}
A \arrow[r, "b"] & Y \\
L \arrow[u, "a"] \arrow[ur, "c"'] &
\end{tikzcd}$$

Adds a **parametric assumption:**
Each output is a linear, additive function of its inputs

$$E(Y \mid A, L) = \alpha + bA + cL$$

Today we will use linear path models

$$
\begin{tikzcd}
A \arrow[r, "b"] & Y \\
L \arrow[u, "a"] \arrow[ur, "c"'] &
\end{tikzcd}
$$

Adds a **parametric assumption:**
Each output is a linear, additive function of its inputs

$$\mathrm{E}(Y \mid A, L) = \alpha + bA + cL$$

- Homogeneous causal effects

Today we will use linear path models

$$\begin{tikzcd} A \arrow[r, "b"] & Y \\ L \arrow[u, "a"] \arrow[ur, "c"'] & \end{tikzcd}$$

Adds a **parametric assumption:**
Each output is a linear, additive function of its inputs

$$E(Y \mid A, L) = \alpha + bA + cL$$

- Homogeneous causal effects
- Linear causal effects

Today we will use linear path models

$$\begin{tikzcd}
A \arrow[r, "b"] & Y \\
L \arrow[u, "a"] \arrow[ur, "c"'] &
\end{tikzcd}$$

Adds a **parametric assumption:**
Each output is a linear, additive function of its inputs

$$\mathrm{E}(Y \mid A, L) = \alpha + bA + cL$$

- Homogeneous causal effects
- Linear causal effects

We will also assume all variables are **standardized** to mean 0 and variance 1

CORRELATION AND CAUSATION

By SEWALL WRIGHT

Senior Animal Husbandman in Animal Genetics, Bureau of Animal Industry, United States Department of Agriculture

PART I. METHOD OF PATH COEFFICIENTS

INTRODUCTION

The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated by well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.

The present paper is an attempt to present a method of measuring the direct influence along each separate path in such a system and thus of finding the degree to which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain the method can be used to find the logical consequences of any particular hypothesis in regard to them.

CORRELATION

Relations between variables which can be measured quantitatively are usually expressed in terms of Galton's (*4*)[1] coefficient of correlation, $r_{XY}=\frac{\Sigma X'Y'}{n\sigma_X\sigma_Y}$ (the ratio of the average product of deviations of *X* and *Y* to the product of their standard deviations), or of Pearson's (*7*) correlation ratio, $\eta_{X\cdot Y}=\frac{\sigma\left({}_{Y}{}^{M}{}_{X}\right)}{\sigma_X}$ (the ratio of the standard deviation of the mean values of *X* for each value of *Y* to the total standard deviation of *X*), the standard deviation being the square root of the mean square deviation.

Use of the coefficient of correlation (*r*) assumes that there is a linear relation between the two variables—that is, that a given change in one variable always involves a certain constant change in the corresponding average value of the other. The value of the coefficient can never exceed

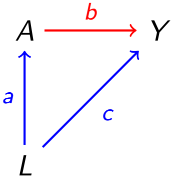
[1] Reference is made by number (italic) to "Literature cited," p. 585.

Journal of Agricultural Research, Washington, D. C. wh

Vol. XX, No. 7 Jan. 3, 1921 Key No. A-55

17777°—21——4

Wright's (1921) path rule:
Connecting causal paths to statistical associations



$$\mathrm{Cov}(A, Y) = b + ac$$

Wright's rule:

When all variables are standardized to variance 1,
the covariance between two variables
is the sum over unblocked paths
of the product of coefficients on that path

When is Wright's rule helpful? Reasoning about biases

$$\begin{tikzcd} A \arrow[r, "b"] & Y \\ U \arrow[u, "a"] \arrow[ur, "c"'] & \end{tikzcd}$$

Suppose we don't observe U.

When is Wright's rule helpful? Reasoning about biases

$$\begin{tikzcd} A \arrow[r, "b"] & Y \\ U \arrow[u, "a"] \arrow[ur, "c"'] & \end{tikzcd}$$

Suppose we don't observe U. We regress Y on A.

$$\mathrm{E}(Y \mid A) = \alpha + \beta A$$

When is Wright's rule helpful? Reasoning about biases

$$\begin{tikzcd} A \arrow[r, "b"] & Y \\ U \arrow[u, "a"] \arrow[ur, "c"'] & \end{tikzcd}$$

Suppose we don't observe U. We regress Y on A.

$$\mathsf{E}(Y \mid A) = \alpha + \beta A$$

What do we get?

When is Wright's rule helpful? Reasoning about biases

$$\begin{tikzcd}
A \arrow[r, "b"] & Y \\
U \arrow[u, "a"] \arrow[ur, "c"'] &
\end{tikzcd}$$

Suppose we don't observe U. We regress Y on A.

$$\mathsf{E}(Y \mid A) = \alpha + \beta A$$

What do we get?

$$\beta = \frac{\mathsf{Cov}(A, Y)}{\mathsf{V}(Y)} \qquad \text{from regression}$$

When is Wright's rule helpful? Reasoning about biases

$$\begin{tikzcd} A \arrow[r, "b"] & Y \\ U \arrow[u, "a"] \arrow[ur, "c"'] & \end{tikzcd}$$

Suppose we don't observe U. We regress Y on A.

$$\mathsf{E}(Y \mid A) = \alpha + \beta A$$

What do we get?

$$\begin{aligned} \beta &= \frac{\text{Cov}(A, Y)}{\text{V}(Y)} && \text{from regression} \\ &= \text{Cov}(A, Y) && \text{since standardized} \end{aligned}$$

When is Wright's rule helpful? Reasoning about biases

$$\begin{tikzcd} A \arrow[r, "b"] & Y \\ U \arrow[u, "a"] \arrow[ur, "c"'] & \end{tikzcd}$$

Suppose we don't observe U. We regress Y on A.

$$\mathrm{E}(Y \mid A) = \alpha + \beta A$$

What do we get?

$$\begin{aligned} \beta &= \frac{\mathrm{Cov}(A, Y)}{\mathrm{V}(Y)} && \text{from regression} \\ &= \mathrm{Cov}(A, Y) && \text{since standardized} \\ &= \underbrace{b}_{\text{Estimand}} + \underbrace{ac}_{\text{Bias}} && \text{Wright's rule} \end{aligned}$$

Why you should rarely use linear path models

1. DAGs formalize causal beliefs
   - good for human reasoning
   - not very amenable to machine learning
2. Nonlinear statistical patterns can be learned from data
   - hard for humans to reason about
   - easy to plug in machine learning

DAGs + ML separate (1) and (2)
Linear path models lump them together